CHARACTERIZATION OF RNA VIRUSES FROM THE COASTAL WATERS OF BRITISH COLUMBIA

by

ALEXANDER IAN CULLEY

B.Sc., University of Oregon, 1993

M.Sc., Moss Landing Marine Laboratories, 2000

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE STUDIES
(Botany)

THE UNIVERSITY OF BRITISH COLUMBIA

January 2007

Abstract

RNA viruses are major pathogens of animals and plants and include viruses that are of enormous economic and public-health concern. In the ocean, RNA viruses infect organisms from bacteria to whales, but RNA virus communities in the sea remain essentially unknown. Although what we know of marine RNA viruses is restricted to a limited number of isolates, emerging data suggest that RNA viruses might be more abundant, and are likely more ecologically important, than has been suggested. Therefore the hypothesis of this dissertation is that RNA viruses comprise a detectable and diverse fraction of the marine virus community. Towards testing this premise, the research objectives were to sequence a marine RNA virus isolate, *Heterosigma akashiwo* RNA virus (HaRNAV), evaluate the diversity of picorna-like viruses, a superfamily of positive-sense single-stranded (ss)RNA viruses, and construct whole-genome shotgun libraries to characterize two complete RNA virus assemblages. The results of all three studies underline the novelty of the marine RNA virus community. For example, HaRNAV is related to picorna-like viruses, but does not belong within any currently defined virus family and has therefore been classified in the *Marnaviridae*, a newly established virus family. Furthermore, on the basis of analysis of RNA-dependent RNA polymerase sequences amplified from marine virus communities from the Strait of Georgia, a diverse array of picorna-like viruses exists in the ocean. All of the sequences amplified were divergent from known picorna-like viruses, and fell within four monophyletic groups. Finally, analysis of reverse transcribed whole-genome shotgun libraries revealed a diverse assemblage of RNA viruses, including a broad group of marine picorna-like viruses and distant relatives of viruses infecting arthropods and higher plants. Moreover, the genomes of several hitherto undiscovered viruses were completely assembled. These data are among the first characterizations of the *in situ* marine RNA virus community and represent a preliminary step in the elucidation of their role in the marine environment. The discovery of novel groups of viruses that are significantly divergent from established taxa should be of interest to virologists, oceanographers, and microbial ecologists.

Table of Contents

List of Tables

List of Figures

List of Symbols and Abbreviations

A	adenine
aa	amino acid
D	aspartic acid
bp	base-pair
BLAST	basic local alignment search tool
CO_2	carbon dioxide
contig	contiguous segment of overlapping sequence fragments
C	cytosine
°C	degrees Celcius
DGGE	denaturing gradient gel electrophoresis
DNA	deoxyribonucleic acid
dNTP	deoxyribonucleoside triphosphate
dia	diameter
DTT	dithiothreitol
dsDNA	double-stranded DNA
dsRNA	double-stranded RNA
EDTA	ethylenediaminetetraacetic acid
E value	expected value
FRP	Fraser river plume
G	guanine/glysine
h	hour

JP	Jericho pier
IRES	internal ribosome entry site
kbp	kilobasepairs
kDa	kilodalton
kPa	kilopascal
LASL	linker amplified shotgun library
l	litre
$MgCl_2$	magnesium chloride
mg	milligram
min	minute
ml	millilitre
mM	millimolar
mm	millimeter
M	molar
ng	nanogram
nM	nanomolar
nm	nanometer
NCBI	national center for biotechnology information
nt(s)	nucleotide(s)
oligo	oligonucleotide
ORF	open reading frame
ppt	parts per thousand
PHACCS	phage communities from contig spectrum
PAUP	phylogenetic analysis using parsimony

pmol	picomole
poly(A)	polyadenylate
PCR	polymerase chain reaction
PVDF	polyvinylidene fluoride
PSI-BLAST	position-specific iterated BLAST
PFGE	pulsed field gel electrophoresis
RACE	rapid amplification of complementary ends
RT	reverse transcription
RT-PCR	reverse transcription- polymerase chain reaction
RNA	ribonucleic acid
RdRp	RNA-dependent RNA polymerase
s	second
ssDNA	single-stranded DNA
ssRNA	single-stranded RNA
SDS-PAGE	sodium dodecyl sulphate polyacrilimide gel elctrophoresis
km^2	square kilometers
SOG	Strait of Georgia
T	thymine
TEM	transmission electron microscope
TAE	tris acetate EDTA
TBE	tris borate EDTA
unid.	unidentified
U	units/uracil
UTR	untranslated region

v	version
v/v	volume per volume
w/v	weight per volume
WGS	whole-genome shotgun
~	approximately
µg	microgram
µl	microlitre
µm	micrometer
µM	micromolar
-	negative-sense
+	positive-sense
’	prime
× g	times gravity

Acknowledgements

I thank Amanda Toperoff for her unwavering support, friendship, creative input and eternal optimism and my family for their encouragement and patience. Thanks to my advisor Curtis Suttle for his guidance, ingenuity and proficient editing and to Andrew Lang for demonstrating the delicate balance of persistence, precision, faith and resiliency required in molecular biology. I offer my gratitude to the members of the Suttle laboratory past and present for lively scientific discourse, assistance in sample collection and buoyant camaraderie, in particular, Amy Chan, Caroline Chénard, Jessie Clasen, André Comeau, Matt Fischer, Emma Hambly, Janice Lawrence, Pascal Loret, Jérôme Payet, Nina Nemcek, Cindy Short, Steven Short and Vera Tai. I am grateful to Debbie Adam, Mary Berbee, Patrick Keeling, Keizo Nagasaki, D'Ann Rochon, Hélène Sanfaçon and Lee Taylor for their assistance during the publication process, and my committee members, François Jean, Patrick Keeling, Bill Mohn and Curtis Suttle for their valuable input. My work could not have been completed without the financial support of the the Department of Botany, the Department of Earth and Ocean Sciences, the University of British Colombia and the Natural Science and Engineering Research Council of Canada.

Dedication

to blue

Co-Authorship Statement

In Chapter II, Andrew Lang designed and directed the sequencing effort. I participated in all aspects of the sequencing of the HaRNAV genome with the exception of N-terminal protein sequencing of structural proteins. My primary contribution to data analyses was to the phylogenetic analysis of conserved putative proteins. My contributions to the manuscript included a discussion of the results of the phylogenetics and in manuscript preparation and revision.

In Chapters III, IV and V, I designed and performed the research, analyzed the data and was the lead author of the manuscripts. Andrew Lang made significant contributions to methods development and participated in the preparation and revision of the manuscripts.

As the research supervisor, Curtis Suttle was involved in the conceptualization and design of the research and in manuscript preparation and revision, but was not involved in the execution of the research.

Chapter I. Introduction to marine RNA viruses

1.1 Background

1.1.1 Importance of marine viruses

Viruses are the most numerous biological entities in the ocean, typically present at concentrations of tens of billions of free virions per liter of seawater (Wommack & Colwell 2000). The virus community or virioplankton is comprised of a morphologically (Weinbauer 2004) and genetically (Edwards & Rohwer 2005) diverse array of pathogens that infect organisms from every level of the marine food web ranging from cetaceans (Bracht et al. 2006) to microbes (Weinbauer 2004). Viruses play a significant ecological role in the marine environment, including controlling the population structure of planktonic communities (Wommack et al. 1999), as direct agents of mortality (Fuhrman & Noble 1995) and as mediators of horizontal gene transfer (Jiang & Paul 1998), ultimately resulting in viruses influencing the way nutrients cycle in the ocean. Moreover, there are data indicating that the ocean can act as a reservoir for the transmission of viruses that cause disease in humans and terrestrial plants and animals (Munn 2006).

As well as being abundant and ubiquitous, viruses are agents of mortality in the ocean. On a community level, changing the concentration of viruses in seawater can affect prokaryote abundance (Proctor & Fuhrman 1990), and phytoplankton biomass (Suttle et al. 1990). Fuhrman and Noble (1995) demonstrated that viral lysis and zooplankton grazing could contribute equally to mortality in marine prokaryote communities and Fischer et al. (2006) estimated that viral-induced lysis accounted for on average an order of magnitude more prokaryote mortality than grazing in marine sediments. Nevertheless, investigations of viruses in the water column (Pedrós-Alió et al. 2000) and sediment (Filippini et al. 2006) found the relative contribution of viruses to be insignificant, demonstrating that the impact of viruses can be variable. Virus-induced mortality of prokaryotes is on average 25% in oceanic and 58% in coastal waters (Weinbauer 2004), however, estimates range from 0 to greater than 100% and clearly suffer from poorly constrained assumptions (Suttle 2005). Viruses have also been implicated in the termination of plankton blooms. For example, blooms of a strain of *Vibrio natriegens* were terminated by the addition of a natural virus community (Hennes et al. 1995). The marine coccolithophorid *Emiliania huxleyi* is capable of forming blooms in temperate waters that cover upwards of 10,000 km^2 (Dunnigan et al. 2006). *E. huxleyi* are armored in calcium carbonate

scales and thus the demise of these immense blooms results in a significant flux of carbon from the surface to deeper waters (Dunigan et al. 2006). In several cases, viruses have been identified as the primary agent of *E. huxleyi* bloom termination (Bratbak et al. 1993, Wilson et al. 2002). Viruses have been recognized in greater than 50 species of algae, however the influence of viral infection on these organisms remains obscure (Van Etten et al. 2002). Viruses are responsible for the demise of multicellular organisms as well. Ostreid herpesvirus 1 (OsHV1) infects several species of bivalves and appears to be responsible for sudden die offs in populations of cultured abalone and juvenile pacific oysters (Friedman et al. 2005). The lethal disease white-spot syndrome is caused by WSSV (white-spot syndrome virus). Epidemics of WSSV have decimated shrimp populations in Asia and prompted a worldwide effort to contain and eradicate the virus (Flegel 2006).

It has been hypothesized that viruses sustain diversity in host populations by culling the most abundant populations of successful competitors and therefore making available niche space for less competitive species to occupy (Fuhrman & Suttle 1993). Suttle (1992) observed shifts in the composition of a primary producer community with the addition of a natural virus community. The close coupling of changes in the population structure of virus, phytoplankton and bacterial communities suggests a relationship between viral activity and host diversity (Larsen et al. 2001). An indirect affect of virus-induced mortality on host diversity was observed by Van Hannen et al. (1999), who remarked that the termination of a bloom of cyanobacteria due to viral lysis corresponded with a dramatic shift in the heterotrophic bacterial community composition, possibly due to the pulse of available organic material liberated by viral activity.

It is likely that virus-mediated gene transfer, via transduction and/or transformation, is an important mechanism of host evolution in the marine environment. Weinbauer et al. (2003) estimated that on average 35% of marine prokaryotes from the Mediterranean Sea harboured a viable, inducible prophage. Additional averaged estimates of percent lysogeny in the marine prokaryotic community range from 3 to 30% (Weinbauer 2004). Jiang and Paul (1998) calculated that 1.3×10^{14} transduction events per year occurred between phages and host communities in the Tampa Bay estuary. Moreover there are data that transduction can occur between marine bacteria of different genera (Chiura 1997). Analyses of a rapidly increasing number of microbial genome sequences suggest that gene exchange between virus and host is a common occurrence. For example, cyanophages contain homologs of essential components of

the photosynthetic apparatus of *Synechococcus* (Mann et al. 2003) and *Prochlorococcus* (Sullivan et al. 2006). Based on sequence analysis, Zeidner et al. (2005) concluded that exchange of these genes between virus and host has occurred on numerous occasions. In an analysis of PBCV (*Paramecium bursaria* chlorella virus), the type species for the *Phycodnaviridae*, a family of large, double-stranded (ds)DNA viruses common in the marine environment, Iyer et al (2006) identified 46 and 40 genes of eukaryotic and prokaryotic origin respectively. There is no data evaluating the importance of transfomation in the marine environment. Neverthless, Jiang and Paul (1995) calculated that viral lysis was responsible for up to 37% of dissolved DNA in the water column. Whether this pool is available for uptake and integration by microbes is unknown.

An essential step in understanding the global cycling of nutrients such as carbon is the elucidation of nutrient cycles in the ocean. Microbes are uniquely equipped to convert nutrients from one form to another and therefore play a vital role in marine biogeochemical cycles (DeLong & Karl 2005). Because viruses infect marine microbes, it is assumed that viruses play a role in the cycling of nutrients in the ocean. Wilhelm and Suttle (1999) postulated that from 6 to 26% of carbon resulting from photosynthesis is diverted back into a dissolved form due to viral lysis and that the overall effect of viral activity is to augment the rate of movement of nutrients from particulate organic matter to dissolved organic matter, diverting nutrients from higher trophic levels back into the microbial fraction (Suttle 2005). Gobler et al. (1997) observed a detectable increase in bioavailable nutrients during the termination of a phytoplankton bloom due to viral lysis. Moreover, products of viral lysis were shown to be responsible for increases in bacterial production (Middelbøe 2000). In oligotrophic environments where infusions of allochthonous nutrients are infrequent, viral lysis may represent a significant pathway for phosphate recycling (Middelbøe et al. 1996). However, Stoderegger & Herndl (1998) found that the fragments of bacterial cells that are a result of viral lysis are largely unavailable for incorporation by the microbial community. The virioplankton itself represents the second largest pool of carbon in the ocean, orders of magnitude greater than marine protists (Weinbauer 2004). Direct consumption of viral particles has been documented (González & Suttle 1993), however Wilhelm & Suttle (1999) estimated that only 1% of the viral community is removed due to grazing. The products of viral lysis form large colloids (Shibata et al. 1997) that may increase the rate at which organic matter is advected out of the photic zone, effectively reducing the amount of nutrients available to primary producers.

The study of marine viruses is justified based on the indisputable fact that viruses infect marine organisms and must therefore have some influence on the marine ecosystem. As discussed above, viral infection can have direct and indirect effects on marine organisms that result in changes in microbial food webs, population structure and biogeochemical cycling. The continued characterization of virus communities and isolates on a molecular level may lead to new insights into virology, including virus evolution. Before our understanding of marine viruses on a community level can improve, however, we must better characterize the individual viruses that comprise the virioplankton. For example we lack data on fundamental areas of research such as infection, reproduction and extra and intracellular persistence.

1.1.2 Most marine viruses are bacteriophages

Viruses are obligate parasites and therefore the composition of the virus community generally reflects the composition of the host community (this is of course a generalization as among other factors, community composition is affected by virus burst size and decay rate). In the marine environment, prokaryotes are on average an order of magnitude more abundant than eukaryotes and thus the majority of marine viruses are believed to be phages (Cochlan et al. 1993). Several independent lines of evidence support this contention. In a wide range of marine environments, the greatest spatial and temporal covariance of viral abundance is with prokaryote abundance and prokaryote activity suggesting that prokaryotes are the hosts of most marine viruses (Wommack & Colwell 2000). It should be noted that these data do not discriminate between bacteria and archaea. Whole genome shotgun libraries of marine coastal DNA virus communities demonstrated that of the environmental sequences with homologues in the NCBI database up to 90% were most similar to bacteriophages (Edwards & Rohwer 2005). Nevertheless, a majority of sequences were unidentifiable and thus the composition of the host community remains uncertain. The majority of marine DNA virus genomes are between 30 and 60 kbp in size, a size range characteristic of bacteriophages (Steward et al. 2000). To this point, the genome size range of marine RNA viruses remains unknown.

1.1.3 A majority of marine viruses have DNA genomes

Most marine viruses are believed to have DNA genomes. Of the greater than 5000 bacteriophages isolated, 96% are tailed and have dsDNA genomes (Ackerman 2000). Moreover, a majority of marine phage isolates belong to the *Myoviridae*, a family of viruses with contractile

tails and dsDNA genomes. Although none infect marine hosts, all archaeal phage isolates to date have linear or circular dsDNA genomes (Prangishvill et al. 2006). Estimates of the proportion of tailed phages (and thus viruses assumed to have dsDNA genomes) from in situ viral communities based on transmission electron microscopy (TEM) range from less than 50% (Wommack et al. 1992) to approximately 90% (Demuth et al. 1993), although it is likely that these results underestimate the proportion of tailed phages due to improper sample preparation and staining (Weinbauer 2004).

1.1.4 Marine RNA bacteriophages

Of the hundreds of marine bacteriophages that have been characterized (Børsheim 1992), only two have RNA genomes (Lewin 1963, Hidaka 1971). Lewin (1963) described tailed, rod-like particles with RNA genomes approximately 200 nm in length that lysed isolates of the marine flexibacterium *Saprospira grandis*. Subsequently, Hidaka (1971) isolated and characterized (Hidaka & Ichida 1976) a single-stranded (ss)RNA virus named 06N-58P from coastal Japanese waters that infected a marine strain of *Pseudomonas*. 06N-58P has a narrow host range, capable of infecting only one of several *Pseudomonas* strains challenged (Hidaka & Ichida 1976). The viral particle has an envelope, is icosahedral in shape and approximately 60 nm in diameter. Viruses with DNA genomes also dominate classified groups of viruses that infect prokaryotes. Of the 38 established genera of bacteriophages, 3 of them include viruses with RNA genomes. These genera fall into two families, the *Cystoviridae* and the *Leviviridae*, neither of which includes viruses with marine hosts (http://www.ncbi.nlm.nih.gov/ICTVdb/Ictv/index.htm). Viruses in the *Cystoviridae* have a segmented dsRNA genome and infect several phytopathogenic (pathogenic to plants) species of *Pseudomonas* (Mertens 2004). Virions of the *Leviviridae* are small (~25 nm in diameter) icosahedrans that encapsulate a positive ssRNA genome (Bollback & Huelsenbeck 2004). The hosts of viruses in the *Leviviridae* appear to be restricted to gram-negative bacteria associated with sewage (Bollback & Huelsenbeck 2004). A comparison of epifluorescence counts of total viruses using a universal nucleic acid stain (YOPRO-1) and one specific to dsDNA (DAPI) suggests that ~90% of the community have dsDNA genomes (Weinbauer et al. 1997). However, YOPRO-1 appears to stain RNA viruses weakly (author, unpublished data), which would result in an underestimate of the contribution of RNA viruses to total virus abundance.

These data have led to the assumption that RNA phages comprise an insignificant fraction of the marine bacteriophage community (Steward 1992, Weinbauer and Suttle 1997), although this supposition has not been directly tested, and the data leaves room for uncertainty. For example, it is now well established that the microbes in culture do not reflect the immense diversity of the natural marine prokaryote community (Rappé & Giovannoni 2003). Because the isolation of a virus is dependent on the availability of a host, estimates of viral diversity are limited by host cultivability; thus, the characteristics of marine viral isolates may not accurately reflect the characteristics of the natural community. Additionally, 06N-58P, the marine RNA phage characterized by Hidaka and Ichida (1976), is very unstable, being susceptible to modest changes in temperature, pH and salinity. If these are general characteristics of marine RNA phages, they may not survive standard marine virus collection techniques.

1.1.5 Virus taxonomy

The nature of viruses and viral evolution preclude a classification scheme modeled on classical Linnaean taxonomy (Condit 2001). Specifically, viruses likely have multiple origins and, therefore, a single common ancestor does not exist. Recombination and reassortment in viruses occurs frequently, resulting in viruses with polyphyletic genomes. Temperate viruses are subject to radically different evolutionary pressures depending on whether they are in an integrated or lytic phase of reproduction (Ball 2004). In order to accommodate these characteristics (among others), virus taxonomy is based on a nonsystematic, polythetic, hierarchical system in which viruses are classified by comparing a collection of equivalent properties where the set of properties can change in different taxonomic branches (Condit 2001). Viruses are generally classified by virion morphology (e.g. capsid symmetry and size), virion physical properties (e.g. genome structure and antigenic properties) and biological properties (e.g. replications strategy and pathogenicity) and most recently, with genetic analyses (Condit 2001). In some cases, phylogenies of groups of viruses based on single genes are congruent with established taxa. For example, the RNA-dependent RNA polymerase (RdRp) is one of the few proteins conserved among almost all RNA viruses with the exception of retroviruses (Koonin 1991). Phylogenies based on the RdRp are routinely used to group RNA viruses into species, genera and families, although groupings higher than the family level are unreliable (Zanotto et al. 1996).

1.1.6 Relevant molecular methods in marine virology

Marine virologists have used specific primers and the polymerase chain reaction (PCR) to target evolutionarily informative genes in order to explore the richness of a variety of virus groups in the ocean. In this approach, a fragment of the target molecule is amplified from a community of extracted viral nucleic acids by PCR. The diversity of amplicons in this reaction can be assessed by several methods including cloning of amplicons and sequencing (Cottrell & Suttle 1995), cloning followed by a comparison of insert restriction enzyme digestion patterns (Chen et al. 1996), separation of products on a denaturing gel and comparison of community fingerprints (Short & Suttle 2002) and endonuclease digestion of fluorescently end-labelled PCR products followed by the generation of community profiles on an automated sequence analyzer (Wang & Chen 2004).

A PCR-based approach was first used to investigate the diversity of viruses in the family *Phycodnaviridae*, a group of large dsDNA viruses that infect algae (Chen et al. 1996). Targeting the viral DNA polymerase, this research revealed a vast amount of genetic variation that was not represented in cultures and showed that very similar sequences were distributed on a global scale (Short & Suttle 2002). A subsequent temporal study in British Columbia showed that the algal virus community is remarkably stable, even while the host community is undergoing dramatic shifts in composition (Short & Suttle 2003). Schroeder et al. (2003) used primers targeting a capsid gene and PCR to track the dynamics of *Emiliania huxleyi* viruses (EhV) during the termination of an *Emiliania huxleyi* bloom. Fuller et al (1998) developed a PCR-based approach targeting gp20, a capsid gene conserved in a subset of myoviruses. Investigations based on this method revealed the incredible diversity present in myovirus communities (Zhong et al. 2002). Yet, despite their incredible diversity, nearly identical sequences were recovered from virus communities ranging from Arctic waters to freshwater catfish ponds (Short & Suttle 2005). Similarly, genetically indistinguishable podovirus sequences have been found in a wide range of environments (Breitbart et al. 2004a).

The chimeric nature of viral genomes and the multiple ancestries of their constituent genes not only complicate the classification of viruses (Lawrence et al. 2002), but also limit the utility of a single gene approach to investigate environmental virus diversity. It is therefore essential that the targeted molecular marker reflects a meaningful biological relationship.

Furthermore a single gene approach to characterizing diversity may not be a viable option with some groups of viruses (Hendrix et al. 1999). However, viral genes that constitute a "core genome" that are resistant to lateral transfer may represent attractive targets (Jain et al. 1999).

As well as targeting molecular markers, marine virologists have used PCR to construct whole-genome shotgun (WGS) libraries of natural virus communities. Breitbart et al. (2002) used a linker amplified shotgun library (LASL) approach to construct a metagenomic library of two coastal DNA phage communities. In the LASL method, 200 l of seawater were pre-filtered and concentrated using tangential flow filtration. The phage fraction was purified from a cesium chloride gradient and the viral DNA extracted and sheared. After end-repairing the sequence fragments, linkers were added and PCR conducted with primers targeting sites specific to these linkers. Amplicons were subsequently cloned and sequenced (Brietbart et al. 2002). The LASL approach effectively overcame obstacles particular to environmental phage communities, including low concentration of nucleic acids per viral genome (sub-femtogram), modified viral DNA and genes lethal to transformed cells during cloning (Edwards & Rohwer 2005). Subsequently, the LASL method was used to examine phage communities from marine sediment (Breitbart et al. 2004b) and human (Breitbart et al. 2003) and equine (Cann et al. 2005) feces. Analysis of these libraries demonstrated that most of the sequence fragments are novel. The results of a model based on the observed overlap of sequence fragments, estimated that the number of different viral genotypes ranged from approximately 1000 in the fecal communities (Breitbart et al. 2003, Cann et al. 2005) to one million in the marine sediment (Breitbart et al. 2004b), suggesting that phage communities are some of the most diverse on the planet (Edwards & Rohwer 2005).

1.1.7 Introduction to RNA virology

The genetic material of RNA viruses can be single-stranded or double-stranded molecules of ribonucleic acid. Single-stranded RNA viruses are further classified by the polarity of their genomes. During replication, the genome of a positive-sense RNA virus is translated directly while the genome of a negative-sense RNA virus is first converted to positive-sense RNA by a RNA polymerase (Prescott et al. 1993). RNA virus genomes can occur in segments where each segment encodes one protein, or in a single molecule that is transcribed into a polyprotein from the entire genome (Roizman & Palese 1996). Although some RNA virions

(complete viral particles) have other constituents such as an envelope, all RNA viruses have a nucleocapsid core composed of RNA surrounded by a protein shell called a capsid. Capsid morphology is generally icosahedral or helical, however there are many exceptions (Prescott et al. 1993). Of the ~3600 virus species characterized, ssRNA viruses are the most diverse, followed by dsDNA, dsRNA and ssDNA respectively. However, these data are greatly influenced by the focus of virology on pathogens of humans and economically important organisms (Villareal 2005).

1.1.8 Marine RNA viruses

RNA viruses of every classification that infect a diversity of host organisms have been isolated from the ocean. For example, marine birnavirus (MABV) has been isolated in a variety of bivalves from Japanese waters (Suzuki & Nojima 1999) and has been responsible for significant losses in populations of cultivated pearl oysters (Kitamura et al. 2001). Marine RNA viruses have also been associated with several species of crustaceans. For example, positive-sense, single-stranded viruses have been found in species of penaeid shrimp (Mari et al. 2002, Sritunyalucksana et al. 2006) including Taura shrimp virus (TSV). Infection by this virus can be fatal to the host organism and the virus has several variants that are wide spread in North American waters (Erickson et al. 2005). Moreover, dsRNA viruses have been observed in crustaceans, including several species of crab (Pappalardo et al. 1986, Zhang et al. 2004). The negative-sense, ssRNA rhabdoviruses and paramyxoviruses are major pathogens of fish (Hoffmann et al. 2005). Viral haemorrhagic septicaemia virus (VHSV) and infectious haematopoietic necrosis virus (IHNV) are negative-sense ssRNA viruses that cause disease in trout and salmon. Double-stranded RNA reoviruses infect fish including Chinook, Chum and Coho Salmon, Striped Bass, and Turbot (Mertens 2004) and retroviral sequences have been amplified from sharks and Pufferfish (Herniou et al. 1998). The positive-sense ssRNA betadnaviruses can cause encephalopathy (alterations in brain function) in multiple species of wild and farmed fish (Gomez et al. 2004). RNA virus infection is also common in marine mammals. Phocine distemper virus (PDV) is a negative-sense ssRNA morbillivirus that has decimated European seal populations over the past two decades (Barrett et al. 2003). Cetacean morbillivirus (CMV) and its variants have been isolated from whales, dolphins and porpoises (Rima et al. 2005) and CMV pathology is frequently present in stranded animals (Taubenberger et al. 2000). RNA viruses such as caliciviruses (Smith 2000) and influenza viruses (Van Bressem

et al. 1999) are found in many marine mammal populations, although infections tend not to be lethal. Although there are numerous examples of RNA viruses from the ocean, most known ones infect marine animals, which make up a small fraction of the living biomass in the ocean, and which are unlikely to be the major hosts for the natural RNA virioplankton.

1.1.9 RNA viruses that infect marine protists

Assuming that marine RNA phages are rare (see section 1.1.4), the most likely major hosts of RNA viruses in the ocean are the diverse, abundant and ecologically and economically important marine protists. The first RNA virus reported to lyse a eukaryotic phytoplankter is HaRNAV (*Heterosigma akashiwo* RNA virus), a ssRNA virus that infects the unicellular photosynthetic marine flagellate *Heterosigma akashiwo* (Tai et al. 2003). The population dynamics of *H. akashiwo* are of special economic interest as blooms of this alga are responsible for extensive fish kills worldwide, affecting the aquaculture industry in particular (Smayda 1998). HaRNAV was isolated from the southern Strait of Georgia, British Columbia. It has a non-enveloped, icosahedral capsid that is 25 nm in diameter, composed of five major proteins and contains a genome of ~9 kbp (Tai et al. 2003). Characteristics of HaRNAV infection include the presence of viral crystalline arrays, the swelling of the endoplasmic reticulum and vacuolation of the cytoplasm, pathology that is similar to established groups of positive-sense ssRNA viruses (Tai et al. 2003). HaRNAV's host range is restricted to *H. akashiwo* strains isolated from the North East Pacific (Tai et al. 2003), complementing other research that shows that most marine viruses are strain-specific (Tomaru et al. 2004). The characterization of the genome sequence of HaRNAV can be found in Chapter II.

RNA viruses are known to infect and lyse ecologically important marine organisms including a diatom (Nagasaki et al. 2004a) and a dinoflagellate (Tomaru et al. 2004). The first virus isolate shown to infect a diatom is the positive-sense ssRNA virus RsRNAV, which infects *Rhizosolenia setigera* (Nagasaki et al. 2004a), a diatom common in temperate coastal waters and a reoccurring member of the Fall and Spring blooms (Graham & Wilcox 2000). Diatoms are globally distributed and incredibly diverse, encompassing approximately 12,000 recognized species (Graham & Wilcox 2000). They are an integral part of biogeochemical cycling in the ocean and typically dominate the phytoplankton in nutrient rich waters (Graham & Wilcox 2000). RsRNAV has an icosohedral, non-enveloped capsid (32 nm dia). The positive-sense 8.9

kbp genome has a polyadenylate [poly(A)] tail and is polycistronic, encoding three major structural proteins (Nagasaki et al. 2004a). The latent period of RsRNAV infection is two days culminating in lysis, at which point 1000 to 3000 virions are released (Nagasaki et al. 2004a). The impact of RsRNAV on *R. setigera* populations is unknown; however, the existence of RsRNAV indicates that RNA viruses are agents of mortality of marine diatoms and hence likely affect diatom population dynamics. *Heterocapsa circularisquama* RNA virus (HcRNAV) is a non-enveloped, positive-sense ssRNA virus approximately 30 nm in diameter that infects *Heterocapsa circularisquama* (Tomaru et al. 2004), a harmful bloom-forming dinoflagellate responsible for the mass mortality of shellfish in Japanese waters (Matsuyama et al. 1999). Dinoflagellates are a diverse group of protists that include autotrophs, mixotrophs, osmotrophs (i.e. they are capable of the direct uptake of dissolved organic compounds) and parasites. They are found in most aquatic habitats and are vital primary producers primarily in coastal waters (Graham & Wilcox 2000). The HcRNAV genome is 4.4 kbp, lacks a poly (A) tail and contains two open reading frames (ORFs). ORF 1 is proximal to the 5' end and has identifiable protease and RdRp motifs, and ORF 2 encodes a single structural protein estimated to be 38 kDa in size (Nagasaki et al. 2005). Phylogenetic analysis based on translated RdRp alignments suggests HcRNAV is related to viruses from the *Luteoviridae*, *Barnaviridae* and *Tetraviridae*, but falls outside these families (Nagasaki et al. 2005). In an additional study, Nagasaki et al. (2004b) demonstrated that during the peak of a bloom in Ago Bay, Japan, a remarkable 88% of *H. circularisquama* cells contained virus-like particles resembling HcRNAV; which suggests that HcRNAV plays a significant role in *H. circularisquama* bloom termination.

As well as infecting ecologically important eukaryotic phytoplankton, RNA viruses infect heterotrophic protists as well. SssRNAV (Schizocytrium single-stranded RNA virus) is a ssRNA virus that infects the thraustochytrid *Schizocytrium* sp. (Takao et al. 2005) Thraustochytrids are osmotrophic marine fungoid protists found in a wide-range of aquatic habitats, where they serve as important decomposers (Kimura et al. 1999). The genome of SssRNAV is positive-sense, 9018 bp in length and has a poly(A) tail. The viral genome codes for two putative polyproteins, a non-structural polyprotein proximal to the 5' end that includes helicase, protease and RdRp domains and, a structural polyprotein that encodes three major and two minor structural proteins (Takao et al. 2006). Although SssRNAV has a similar genome organization to viruses in the family *Dicistroviridae*, subgenomic RNAs are present during SssRNAV replication but not

during dicistrovirus replication. Moreover, phylogenetic analysis strongly supports the placement of SssRNAV outside established families of RNA viruses (Takao et al. 2006).

The isolation of *Micromonas pusilla* reovirus (MpRV) demonstrated that dsRNA viruses are capable of infecting a marine primary producer as well. *Micromonas pusilla* is a flagellated marine phytoplankter identified as the most abundant picoeukaryote (< 2 μm) in oceanic and coastal regions (Not et al. 2004). The MpRV virion is 75 nm in diameter, contains five major proteins (Brussard et al. 2004) and does not possess the projections (i.e. 'turrets') characteristic of some viral genera in the family *Reoviridae* (Attoui et al. 2006). The genome is composed of eleven segments of dsRNA, that total ~ 25.5 kbp. Putative cell attachment, capsid and non-structural proteins, including a polymerase were identified based on significant similarity to sequences in the NCBI database (Attoui et al. 2006). Like MpRV, some rotavirus and aquareovirus genomes have eleven segments, however phylogenies generated from alignments of the MpRV polymerase with representative viruses of the eleven genera of the family *Reoviridae*, as well as several unique genomic features including an unusually long segment 1 and novel 5' and 3' terminal sequences, suggest that MpRV belongs in a new genus (Attoui et al. 2006). Table 1.1 provides a synopsis of the newly discovered RNA viruses that infect marine protists.

1.2 Thesis theme

The information presented above suggests that RNA viruses are more abundant, diverse and ecologically important in the sea than has generally been assumed; however, this hypothesis has never been tested. The overall theme of this dissertation is therefore the characterization of the natural RNA virioplankton in the marine environment. Toward this end, the second chapter is an analysis of the complete genomic sequence of HaRNAV, the first positive-sense ssRNA virus isolated that infects a marine protist. The presence of a persistent and widespread marine RNA virus prompted an investigation of the marine RNA virus community, a component of the marine environment almost completely uncharacterized. Thus, the third chapter discusses the results of research assessing the diversity of picorna-like viruses, a group of positive-sense ssRNA viruses with similar genome features and sharing conserved regions in the RdRp, from marine virus communities. In chapter four, randomly reverse-transcribed whole-genome shotgun sequencing is used to characterize the diversity of two complete marine RNA virus assemblages. These virus

communities were heavily dominated by different genotypes with small genome sizes, allowing the complete assembly of the genomes from three previously unknown viruses. The complete genome sequences of these viruses are analyzed in chapter five. The concluding chapter summarizes the findings of the dissertation, examines methodological bias and discusses the general importance and implications of this research.

1.3 Table

Table 1.1 RNA viruses that infect marine protists

Virus	Acronym	Host	Genome	Size (kbp)	Reference
Heterocapsa circularisquama RNA	HcRNAV	*H.circularisquama*	+ ss	4.4	Tomaru et al. 2004
Heterosigma akashiwo RNA	HaRNAV	*H. akashiwo*	+ ss	8.6	Tai et al. 2003
Micromonas pusilla reo-	MpRV	*M. pusilla*	ds	25.5	Brussaard et al. 2004
Rhizosolenia setigera RNA	RsRNAV	*R. setigera*	+ ss	8.9	Nagasaki et al. 2004a
Schizochytrium single-stranded RNA	SssRNAV	*Schizochytrium* sp.	+ ss	9.0	Takao et al. 2005

1.4 References

Ackermann, H. W. 2001. Frequency of morphological phage descriptions in the year. 2000. Archives of Virology **146**: 843-857.

Attoui, H., F. M. Jaafar, M. Belhouchet, P. De Micco, X. De Lamballerie, and C. P. D. Brussaard. 2006. Micromonas pusilla reovirus: a new member of the family *Reoviridae* assigned to a novel proposed genus (*Mimoreovirus*). Journal of General Virology **87:** 1375-1383.

Ball, L. A. 2004. Introduction to universal virus taxonomy, p. 3-9. *In* C. M. Fauquet, M. A. Mayo, J. Maniloff, U. Desselberger and L. A. Ball [eds.], Virus Taxnomy: VIIIth Report of the International Committee on Taxonomy of Viruses. Academic Press.

Barrett, T., P. Sahoo, and P. D. Jepson. 2003. Seal distemper outbreak 2002. Microbiology Today **30:** 162-164.

Børsheim, K. Y. 1993. Native marine bacteriophages. FEMS Microbiology Letters **102:** 141-159.

Bracht, A. J., R. L. Brudek, R. Y. Ewing, C. A. Manire, K. A. Burek, C. Rosa, K. B. Beckmen, J. E. Maruniak, and C. H. Romero. 2006. Genetic identification of novel poxviruses of cetaceans and pinnipeds. Archives of Virology **151:** 423-438.

Bratbak, G., J. K. Egge, and M. Heldal. 1993. Viral mortality of the marine alga *Emiliania huxleyi* (Haptophyceae) and termination of algal blooms. Marine Ecology-Progress Series **93:** 39-48.

Breitbart, M., J. H. Miyake, and F. Rohwer. 2004a. Global distribution of nearly identical phage-encoded DNA sequences. FEMS Microbiology Letters **236:** 249-256.

Breitbart, M., B. Felts, S. Kelley, J. M. Mahaffy, J. Nulton, P. Salamon, and F. Rohwer. 2004b. Diversity and population structure of a near-shore marine-sediment viral community. Proceedings of the Royal Society of London Series B-Biological Sciences **271:** 565-574.

Breitbart, M., I. Hewson, B. Felts, J. M. Mahaffy, J. Nulton, P. Salamon, and F. Rohwer. 2003. Metagenomic analyses of an uncultured viral community from human feces. Journal of Bacteriology **185:** 6220-6223.

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

Brussaard, C. P. D., A. A. M. Noordeloos, R. A. Sandaa, M. Heldal, and G. Bratbak. 2004. Discovery of a dsRNA virus infecting the marine photosynthetic protist *Micromonas pusilla*. Virology **319:** 280-291.

Cann, A. J., S. E. Fandrich, and S. Heaphy. 2005. Analysis of the virus population present in equine feces indicates the presence of hundreds of uncharacterized virus genomes. Virus Genes **30:** 151-156.

Chen, F., C. A. Suttle, and S. M. Short. 1996. Genetic diversity in marine algal virus communities as revealed by sequence analysis of DNA polymerase genes. Applied and Environmental Microbiology **62:** 2869-2874.

Chiura, H. X. 1997. Generalized gene transfer by virus-like particles from marine bacteria. Aquatic Microbial Ecology **13:** 75-83.

Cochlan, W. P., J. Wikner, G. F. Steward, D. C. Smith, and F. Azam. 1993. Spatial-distribution of viruses, bacteria and chlorophyll-a in neritic, oceanic and estuarine environments. Marine Ecology-Progress Series **92:** 77-87.

Condit, R. C. 2001. Principles of virology, p. 19-51. *In* D. M. Knipe, P. M. Howley, D. E. Griffin, R. A. Lamb, M. A. Martin, B. Roizman and S. E. Straus [eds.], Field's Virology. Lippincott Williams & Wilkins.

Cottrell, M. T., and C. A. Suttle. 1995. Genetic diversity of algal viruses which lyse the photosynthetic picoflagellate *Micromonas pusilla* (Prasinophyceae). Applied and Environmental Microbiology **61:** 3088-3091.

DeLong, E. F., and D. M. Karl. 2005. Genomic perspectives in microbial oceanography. Nature **437:** 336-342.

Demuth, J., H. Neve, and K.P. Witzel. 1993. Direct electron microscopy study on the morphological diversity of bacteriophage populations in Lake PluBsee. Applied and Environmental Microbiology **59:** 3378-3384.

Dunigan, D. D., L. A. Fitzgerald, and J. L. Van Etten. 2006. Phycodnaviruses: A peek at genetic diversity. Virus Research **117:** 119-132.

Edwards, R. A., and F. Rohwer. 2005. Viral metagenomics. Nature Reviews Microbiology **3:** 504-510.

Erickson, H. S., B. T. Poulos, K. F. J. Tang, D. Bradley-Dunlop, and D. V. Lightner. 2005. Taura syndrome virus from Belize represents a unique variant. Diseases of Aquatic Organisms **64:** 91-98.

Filippini, M., N. Buesing, Y. Bettarel, T. Sime-Ngando, and M. O. Gessner. 2006. Infection paradox: High abundance but low impact of freshwater benthic viruses. Applied and Environmental Microbiology **72:** 4893-4898.

Fischer, U. R., C. Wieltschnig, A. K. T. Kirschner, and B. Velimirov. 2006. Contribution of virus-induced lysis and protozoan grazing to benthic bacterial mortality estimated simultaneously in microcosms. Environmental Microbiology **8:** 1394-1407.

Flegel, T. W. 2006. Detection of major penaeid shrimp viruses in Asia, a historical perspective with emphasis on Thailand. Aquaculture **258:** 1-33.

Friedman, C. S., R. M. Estes, N. A. Stokes, C. A. Burge, J. S. Hargove, B. J. Barber, R. A. Elston, E. M. Burreson, and K. S. Reece. 2005. Herpes virus in juvenile Pacific oysters *Crassostrea gigas* from Tomales Bay, California, coincides with summer mortality episodes. Diseases of Aquatic Organisms **63:** 33-41.

Fuhrman, J. A., and C. A. Suttle. 1993. Viruses in marine planktonic systems. Oceanography **6:** 51-63.

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

Fuller, N. J., W. H. Wilson, I. R. Joint, and N. H. Mann. 1998. Occurrence of a sequence in marine cyanophages similar to that of T4 g20 and its application to PCR-based detection and quantification techniques. Applied and Environmental Microbiology **64:** 2051-2060.

Gobler, C. J., D. A. Hutchins, N. S. Fisher, E. M. Cosper, and S. Sanudo-Wilhelmy. 1997. Release and bioavailability of C, N, P, Se, and Fe following viral lysis of a marine Chrysophyte. Limnology and Oceanography **42:** 1492-1504.

Gomez, D. K., J. Sato, K. Mushiake, T. Isshiki, Y. Okinaka, and T. Nakai. 2004. PCR-based detection of betanodaviruses from cultured and wild marine fish with no clinical signs. Journal of Fish Diseases **27:** 603-608.

González, J. M., and C. A. Suttle. 1993. Grazing by marine nanoflagellates on viruses and virus-sized particles: Ingestion and digestion. Marine Ecology-Progress Series **94:** 1-10.

Graham, L. E., and L. W. Wilcox. 2000. Algae. Prentice-Hall, Inc.

Hendrix, R. W., M. C. M. Smith, R. N. Burns, M. E. Ford, and G. F. Hatfull. 1999. Evolutionary relationships among diverse bacteriophages and prophages: All the world's a phage. Proceedings of the National Academy of Sciences of the United States of America **96:** 2192-2197.

Hennes, K., and M. Simon. 1995. Significance of bacteriophages for controlling bacterioplankton growth in a Mesotrophic Lake. Applied and Environmental Microbiology **61:** 333-340.

Herniou, E., J. Martin, K. Miller, J. Cook, M. Wilkinson, and M. Tristem. 1998. Retroviral diversity and distribution in vertebrates. Journal of Virology **72:** 5955-5966.

Hidaka, T., and K. Ichida. 1976. Properties of a marine RNA-containing bacteriophage. Memoirs of the Faculty of Fisheries, Kagoshima University **25:** 77-89.

Hidaka, T. 1971. Isolation of marine bacteriophages from seawater. Bulletin of the Japanese Society of Scientific Fisheries **37:** 1199-1206.

Hoffmann, B., M. Beer, H. Schutze, and T. C. Mettenleiter. 2005. Fish rhabdoviruses: Molecular epidemiology and evolution. Current Topics in Microbiology and Immunology **292:** 81-117.

Iyer, L. A., S. Balaji, E. V. Koonin, and L. Aravind. 2006. Evolutionary genomics of nucleo-cytoplasmic large DNA viruses. Virus Research **117:** 156-184.

Jain, R., M. C. Rivera, and J. A. Lake. 1999. Horizontal gene transfer among genomes: The complexity hypothesis. Proceedings of the National Academy of Sciences of the United States of America **96:** 3801-3806.

Jiang, S. C., and J. H. Paul. 1998. Gene transfer by transduction in the marine environment. Applied and Environmental Microbiology **64:** 2780-2787.

Jiang, S. C., and J. H. Paul. 1995. Viral contribution to dissolved DNA in the marine-environment as determined by differential centrifugation and kingdom probing. Applied and Environmental Microbiology **61:** 317-325.

Kimura, H., T. Fukuba, and T. Naganuma. 1999. Biomass of thraustochytrid protoctists in coastal water. Marine Ecology-Progress Series **189:** 27-33.

Kitamura, S. I., S. J. Jung, and S. Suzuki. 2000. Seasonal change of infective state of marine birnavirus in Japanese pearl oyster *Pinctada fucata.* Archives of Virology **145:** 2003-2014.

Koonin, E. V. 1991. The phylogeny of RNA-dependent RNA polymerases of positive-strand RNA viruses. Journal of General Virology **72:** 2197-2206.

Larsen, A., T. Castberg, R. A. Sandaa, C. P. D. Brussaard, J. Egge, M. Heldal, A. Paulino, R. Thyrhaug, E. J. Van Hannen, and G. Bratbak. 2001. Population dynamics and diversity of phytoplankton, bacteria and viruses in a seawater enclosure. Marine Ecology-Progress Series **221:** 47-57.

Lawrence, J. G., G. F. Hatfull, and R. W. Hendrix. 2002. Imbroglios of viral taxonomy: Genetic exchange and failings of phenetic approaches. Journal of Bacteriology **184:** 4891-4905.

Lewin, R. A. 1963. Rod-shaped particles in *Saprospira.* Nature **198:** 103-104.

Mann, N. H., A. Cook, A. Millard, S. Bailey, and M. Clokie. 2003. Marine ecosystems: bacterial photosynthesis genes in a virus. Nature **424:** 741.

Mari, J., B. T. Poulos, D. V. Lightner, and J. R. Bonami. 2002. Shrimp Taura syndrome virus: genomic characterization and similarity with members of the genus *Cricket paralysis-like viruses.* Journal of General Virology **83:** 915-926.

Matsuyama, Y., T. Uchida, and T. Honjo. 1999. Effects of harmful dinoflagellates, *Gymnodinium mikimotoi* and *Heterocapsa circularisquama*, red-tide on filtering rate of bivalve molluscs. Fisheries Science **65:** 248-253.

Mertens, P. 2004. The dsRNA viruses. Virus Research **101:** 3-13.

Middelbøe, M. 2000. Bacterial growth rate and marine virus-host dynamics. Microbial Ecology **40:** 114-124.

Middelbøe, M., N. O. G. Jorgensen, and N. Kroer. 1996. Effects of viruses on nutrient turnover and growth efficiency of non-infected marine bacterioplankton. Applied and Environmental Microbiology **62:** 1991-1997.

Munn, C. B. 2006. Viruses as pathogens of marine organisms - from bacteria to whales. Journal of the Marine Biological Association of the United Kingdom **86:** 453-467.

Nagasaki, K., Y. Shirai, Y. Takao, H. Mizumoto, K. Nishida, and Y. Tomaru. 2005. Comparison of genome sequences of single-stranded RNA viruses infecting the bivalve-killing dinoflagellate *Heterocapsa circularisquama*. Applied and Environmental Microbiology **71:** 8888-8894.

Nagasaki, K., Y. Tomaru, N. Katanozaka, Y. Shirai, K. Nishida, S. Itakura, and M. Yamaguchi. 2004a. Isolation and characterization of a novel single-stranded RNA virus infecting the bloom-forming diatom *Rhizosolenia setigera*. Applied and Environmental Microbiology **70:** 704-711.

Nagasaki, K., Y. Tomaru, K. Nakanishi, N. Hata, N. Katanozaka, and M. Yamaguchi. 2004b. Dynamics of *Heterocapsa circularisquama* (Dinophyceae) and its viruses in Ago Bay, Japan. Aquatic Microbial Ecology **34:** 219-226.

Not, F., M. Latasa, D. Marie, T. Cariou, D. Vaulot, and N. Simon. 2004. A single species, *Micromonas pusilla* (Prasinophyceae), dominates the eukaryotic picoplankton in the western English channel. Applied and Environmental Microbiology **70:** 4064-4072.

Pappalardo, R., J. Mari, and J. R. Bonami. 1986. Tau-(tau) virus infection of *Carcinus mediterraneus* - Histology, cytopathology, and experimental transmission of the disease. Journal of Invertebrate Pathology **47:** 361-368.

Pedrós-Alió, C., J. I. Calderón-Paz, and J. M. Gasol. 2000. Comparative analysis shows that bacterivory, not viral lysis, controls the abundance of heterotrophic prokaryotic plankton. FEMS Microbiology Ecology **32:** 157-165.

Prangishvill, D., R. A. Garrett, and E. V. Koonin. 2006. Evolutionary genomics of archaeal viruses: Unique viral genomes in the third domain of life. Virus Research **117:** 52-67.

Prescott, L. M., J. P. Harley, and D. A. Klein. 1993. Microbiology, 2 ed. Wm. C. Brown.

Proctor, L. M., and J. A. Fuhrman. 1990. Viral mortality of marine bacteria and cyanobacteria. Nature **343:** 60-62.

Rappé, M. S., and S. J. Giovannoni. 2003. The uncultured microbial majority. Annual Review of Microbiology **57:** 369-394.

Rima, B. K., A. M. J. Collin, and J. A. P. Earle. 2005. Completion of the sequence of a cetacean morbillivirus and comparative analysis of the complete genome sequences of four morbilliviruses. Virus Genes **30:** 113-119.

Roizman, S. G., and P. Palese. 1996. Multiplication of Viruses: An Overview, p. 101-111. *In* B. N. Fields, D. M. Knipe and P. M. Howley [eds.], Fields Virology. Lippincott-Raven.

Schroeder, D. C., J. Oke, M. Hall, G. Malin, and W. H. Wilson. 2003. Virus succession observed during an *Emiliania huxleyi* bloom. Applied and Environmental Microbiology **69:** 2484-2490.

Shibata, A., K. Kogure, I. Koike, and K. Ohwada. 1997. Formation of submicron colloidal particles from marine bacteria by viral infection. Marine Ecology-Progress Series **155:** 303-307.

Short, S. M., and C. A. Suttle. 2002. Sequence analysis of marine virus communities reveals that groups of related algal viruses are widely distributed in nature. Applied and Environmental Microbiology **68:** 1290-1296.

Short, S. M., and C. A. Suttle. 2003. Temporal dynamics of natural communities of marine algal viruses and eukaryotes. Aquatic Microbial Ecology **32:** 107-119.

Short, C. M., and C. A. Suttle. 2005. Nearly identical bacteriophage structural gene sequences are widely distributed in both marine and freshwater environments. Applied and Environmental Microbiology **71:** 480-486.

Smayda, T. J. 1998. Ecophysiology and Bloom Dynamics of *Heterosigma akashiwo* (Raphidophyceae), p. 113-131. *In* D. M. Anderson, A. D. Cembella and G. M. Hallegraeff [eds.], Physiological Ecology of Harmful Algal Blooms. Springer-Verlag.

Smith, A. 2000. Aquatic Virus Cycles, p. 447-491. *In* C. Hurst [ed.], Viral Ecology. Academic Press.

Sritunyalucksana, K., S. Apisawetakan, A. Boon-Nat, B. Withyachumnarnkul, and T. W. Flegel. 2006. A new RNA virus found in black tiger shrimp *Penaeus monodon* from Thailand. Virus Research **118:** 31-38.

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

Steward, G. F., J. Wikner, W. P. Cochlan, D. C. Smith, and F. Azam. 1992. Estimation of virus production in the sea: 2. Field results. Marine Microbial Food Webs **6:** 79-90.

Stoderegger, K., and G. J. Herndl. 1998. Production and release of bacterial capsular material and its subsequent utilization by marine bacterioplankton. Limnology and Oceanography **43:** 877-884.

Sullivan, M. B., D. Lindell, J. A. Lee, L. R. Thompson, J. P. Bielawski, and S. W. Chisholm. 2006. Prevalence and evolution of core photosystem II genes in marine cyanobacterial viruses and their hosts. Plos Biology **4:** 1344-1357.

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

Suttle, C. A. 1992. Inhibition of photosynthesis in phytoplankton by the submicron size fraction concentrated from seawater. Marine Ecology Progress Series **87:** 105-112.

Suttle, C. A., A. M. Chan, and M. T. Cottrell. 1990. Infection of phytoplankton by viruses and reduction of primary productivity. Nature **347:** 467-469.

Suzuki, S., and M. Nojima. 1999. Detection of a marine birnavirus in wild molluscan shellfish species from Japan. Fish Pathology **34:** 121-125.

Tai, V., J. E. Lawrence, A. S. Lang, A. M. Chan, A. I. Culley, and C. A. Suttle. 2003. Characterization of HaRNAV, a single-stranded RNA virus causing lysis of *Heterosigma akashiwo* (Raphidophyceae). Journal of Phycology **39:** 343-352.

Takao, Y., K. Mise, K. Nagasaki, T. Okuno, and D. Honda. 2006. Complete nucleotide sequence and genome organization of a single-stranded RNA virus infecting the marine fungoid protist *Schizochytrium* sp. Journal of General Virology **87:** 723-733.

Takao, Y., K. Nagasaki, K. Mise, T. Okuno, and D. Honda. 2005. Isolation and characterization of a novel single-stranded RNA virus infectious to a marine fungoid protist, *Schizochytrium* sp. (Thraustochytriaceae, labyrinthulea). Applied and Environmental Microbiology **71:** 4516-4522.

Taubenberger, J. K., M. M. Tsai, T. J. Atkin, T. G. Fanning, A. E. Krafft, R. B. Moeller, S. E. Kodsi, M. G. Mense, and T. P. Lipscomb. 2000. Molecular genetic evidence of a novel morbillivirus in a long-finned pilot whale (*Globicephalus melas*). Emerging Infectious Diseases **6:** 42-45.

Tomaru, Y., N. Katanozaka, K. Nishida, Y. Shirai, K. Tarutani, M. Yamaguchi, and K. Nagasaki. 2004. Isolation and characterization of two distinct types of HcRNAV, a single-stranded RNA virus infecting the bivalve-killing microalga *Heterocapsa circularisquama*. Aquatic Microbial Ecology **34:** 207-218.

Van Bressem, M. F., K. Van Waerebeek, and J. A. Raga. 1999. A review of virus infections of cetaceans and the potential impact of morbilliviruses, poxviruses and papillomaviruses on host population dynamics. Diseases of Aquatic Organisms **38:** 53-65.

Van Etten, J. L., M. V. Graves, D. G. Muller, W. Boland, and N. Delaroque. 2002. Phycodnaviridae - large DNA algal viruses. Archives of Virology **147:** 1479-1516.

Van Hannen, E. J., G. Zwart, M. P. Van Agterveld, H. J. Gons, J. Ebert, and H. J. Laanbroek. 1999. Changes in bacterial and eukaryotic community structure after mass lysis of filamentous cyanobacteria associated with viruses. Applied and Environmental Microbiology **65:** 795-801.

Villarreal, L. P. 2005. Viruses and the Evolution of Life. ASM Press.

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

Weinbauer, M. G. 2004. Ecology of prokaryotic viruses. FEMS Microbiology Reviews **28:** 127-181.

Weinbauer, M. G., I. Brettar, and M. G. Hofle. 2003. Lysogeny and virus-induced mortality of bacterioplankton in surface, deep, and anoxic marine waters. Limnology and Oceanography **48:** 1457-1465.

Weinbauer, M. G., and C. A. Suttle. 1997. Comparison of epifluorescence and transmission electron microscopy for counting viruses in natural marine waters. Aquatic Microbial Ecology **13:** 225-232.

Weinbauer, M. G., S. W. Wilhelm, C. A. Suttle, and D. R. Garza. 1997. Photoreactivation compensates for UV damage and restores infectivity to natural marine viral communities. Applied and Environmental Microbiology **63:** 2200-2205.

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

Wilson, W. H., G. Tarran, and M. V. Zubkov. 2002. Virus dynamics in a coccolithophore-dominated bloom in the North Sea. Deep-Sea Research Part II-Topical Studies in Oceanography **49:** 2951-2963.

Wommack, K. E., and R. R. Colwell. 2000. Virioplankton: Viruses in aquatic ecosystems. Microbiology and Molecular Biology Reviews **64:** 69-114.

Wommack, K. E., J. Ravel, R. T. Hill, J. S. Chun, and R. R. Colwell. 1999. Population dynamics of Chesapeake Bay virioplankton: Total-community analysis by pulsed-field gel electrophoresis. Applied and Environmental Microbiology **65:** 231-240.

Wommack, K. E., R. T. Hill, M. Kessel, R. C.E., and R. R. Colwell. 1992. Distribution of viruses in the Chesapeake Bay. Applied and Environmental Microbiology **58:** 2965-2970.

Zanotto, P. M. D., M. J. Gibbs, E. A. Gould, and E. C. Holmes. 1996. A reevaluation of the higher taxonomy of viruses based on RNA polymerases. Journal of Virology **70:** 6083-6096.

Zeidner, G., J. P. Bielawski, M. Shmoish, D. J. Scanlan, G. Sabehi, and O. Beja. 2005. Potential photosynthesis gene recombination between *Prochlorococcus* and *Synechococcus* via viral intermediates. Environmental Microbiology **7:** 1505-1513.

Zhang, S., Z. Shi, J. Zhang, and J. R. Bonami. 2004. Purification and characterization of a new reovirus from the Chinese mitten crab, *Eriocheir sinensis*. Journal of Fish Diseases **27:** 687.

Zhong, Y., F. Chen, S. W. Wilhelm, L. Poorvin, and R. E. Hodson. 2002. Phylogenetic diversity of marine cyanophage isolates and natural virus communities as revealed by sequences of viral capsid assembly protein gene G20. Applied and Environmental Microbiology **68:** 1576-1584.

Chapter II. Genome sequence and characterization of a virus (HaRNAV) related to picorna-like viruses that infects the marine toxic bloom-forming alga *Heterosigma akashiwo*

A version of this chapter has been published

Lang, A.S., A.I. Culley, and C.A. Suttle. 2004. Genome sequence and characterization of a virus (HaRNAV) related to picorna-like viruses that infects the marine toxic bloom-forming alga *Heterosigma akashiwo*. Virology **320**: 206-217.

2.1 Introduction

For many years, viruses or virus-like particles have been reported from numerous taxa representing nearly all the classes of eukaryotic algae (reviewed in Van Etten et al. 1991, Van Etten & Meints 1999). The first report of a virus isolate that infected a marine photosynthetic protist was in 1979 (Mayer & Taylor 1979), but it was not until 10 years later that viruses infecting phytoplankton were readily isolated from seawater (Suttle et al. 1990). Subsequently, there were numerous examples of viruses isolated from the marine environment that infected eukaryotic phytoplankton (Castberg et al. 2002, Cottrell & Suttle 1991, Jacobsen et al. 1996, Nagasaki and Yamaguchi 1997, Sandaa et al. 2001, Suttle and Chan 1995, Tarutani et al. 2001). Although these viruses infect a variety of distantly related taxa, they are morphologically remarkably similar (Suttle 2000). All are large polyhedrons that contain double-stranded DNA ranging from 130 to 560 kbp and appear to belong in the family *Phycodnaviridae*.

Recently, several different types of viruses have been isolated that infect *Heterosigma akashiwo* (*Raphidophyceae*), a unicellular phototrophic marine flagellate that is common in temperate coastal waters and which forms toxic blooms that can kill fish (Taylor 1990). These viruses include HaV, which appears to belong within the *Phycodnaviridae* (Nagasaki & Yamaguchi 1997), HaNIV, which forms paracrystalline arrays within the *H. akashiwo* nucleus (Lawrence et al. 2001), and HaRNAV, a single-stranded RNA virus that assembles within the cytoplasm of infected cells (Tai et al. 2003). HaRNAV particles appear to have icosahedral symmetry and are approximately 25 nm in diameter (Tai et al. 2003). Sequence analysis of the putative RNA-dependent RNA polymerase (RdRp) domain from HaRNAV shows it is related to the picorna-like virus superfamily (Culley et al. 2003).

Here, we report the complete genome sequence of HaRNAV. Analysis of the genome reveals that this virus is related to viruses from the picorna-like superfamily of viruses (the *Picornaviridae*, *Caliciviridae*, *Dicistroviridae*, *Sequiviridae*, *Comoviridae* and *Potyviridae* families, Liljas et al. 2002), but does not belong in any of these currently defined families. The HaRNAV putative nonstructural protein domains and capsid proteins show a mosaic pattern of relationships with sequences from viruses in these families. The organization of HaRNAV structural proteins appears to be the same as in the *Dicistroviridae*, although the overall genome architecture is different from these viruses. Based on our sequence comparisons and analyses of

genome structure, we argue that HaRNAV defines a new family (*Marnaviridae*) of picorna-like viruses.

2.2 Results

2.2.1 Features of the HaRNAV genome sequence

We have determined the complete nucleotide sequence of the HaRNAV genome (GenBank accession number AY337486). The genome is 8587 nucleotides (nts) long, plus a poly(A) tail, which is in close agreement with the predicted size of 8.6 kbp, based on the analysis of the previously published denaturing gel of the HaRNAV genome (Tai et al. 2003). The genome contains a 7743-nt open reading frame (ORF) (Figure 2.1) that is predicted to encode a protein of 2581 amino acids. No other ORFs that could encode proteins larger than 60 amino acids are predicted (Figure 2.1). Assuming we have correctly predicted the start of the polyprotein, the 5' and 3' untranslated regions (UTRs) are 483 and 361 nts long, respectively, accounting for 9.8% of the genome. The protein sequence predicted by the large ORF contains conserved sequence domains from RNA viruses, and the sequences found in the HaRNAV structural proteins (see below).

We used an oligo(dT) primer as part of the 3' RACE system (see Materials and Methods) to clone the 3' end of the genome. This suggests that there is a 3' poly(A) tail, although previous experiments suggested its absence (Tai et al. 2003). A region near the 3' end of the genome (nts 8420–8445) contains 22 out of 26 bases that are U, which could form a secondary structure with the poly(A) tail thereby giving the impression that the genome does not have a poly(A) tail. The addition of DMSO to the first strand cDNA synthesis during the 3' RACE procedure (see Materials and Methods) may have helped to disrupt any secondary structure. All five of the 3' RACE clones gave the same sequence for the end of the genome.

We used a 5' RACE approach to clone the 5' end of the HaRNAV genome. Wu et al. (2002) found that secondary structure in the 5' end of the *Perina nuda* picorna-like virus (PnPV) genome interfered with the 5' RACE procedure. We encountered the same phenomenon, and sequenced nine clones to determine the 5' end of the HaRNAV sequence (Figure 2.2). This revealed a potential stem-loop structure from nts 5 to 39 in which 14 of 15 bases are capable of hybridizing and forming a loop of four bases (Figure 2.2). This is very similar to the PnPV

sequence where a predicted 13-base stem-loop structure occurs from nts 11 to 38 of the genome (Wu et al. 2002). Analysis of the HaRNAV 5' UTR with the web-based version of the RNA secondary structure prediction program mfold 3.0 (Mathews et al. 1999, Zuker et al. 1999) predicted a large amount of secondary structure in the 5' UTR (not shown) including the potential stemloop structure mentioned above. Secondary structure in this region may be important for replication of the RNA as found for poliovirus (Andino et al. 1993). Other potential secondary structures that were predicted closer to the putative start of the polyprotein are likely important as part of an internal ribosome entry site (IRES) for polyprotein translation (reviewed in Hellen & Sarnow 2001, Martinez-Salas et al. 2001, Sarnow 2003). A pyrimidine-rich stretch of sequence wherein 22 of 29 bases are pyrimidines occurs from nts 447 to 475, ending eight bases upstream of the predicted start codon of the large ORF (Figure 2.2); such pyrimidine-rich sequences are conserved in picornavirus IRESs (Hellen & Sarnow 2001, Pestova et al. 1991).

There are two notable repeats in the genome sequence. One pair of repeats involves sequences in the proposed 5' and 3' UTRs. The 136-nt region from nts 312 to 448 shares 123 identical bases with the 137-nt sequence from nts 8265 to 8401 (90% identity). These repeated sequences might have a function in RNA replication or polyprotein translation. It has been suggested that circularization of viral mRNAs may be important for aiding polyprotein translation (reviewed in Martinez-Salas et al. 2001). These repeats could be involved in this function through RNA–RNA or RNA–protein–RNA interactions. RNA secondary structure analysis with mfold predicted the sequence from bases 312 to 448 would fold on itself; therefore these repeats are theoretically capable of interacting directly. The other set of repeats is an overlapping repeat in the predicted coding region. The 95 base sequence from nts 1052 to 1147 shares 87 identical bases with the sequence between nts 1124 and 1218 (91.6% identity). This creates a 31-amino-acid self-overlapping repeat where 27 of 31 amino acids are identical (87% identity). It is difficult to speculate on a potential function for this repeated sequence because we have no indication of a possible function for that region of the polyprotein.

2.2.2 Determination of HaRNAV genome polarity

It was previously shown that the HaRNAV genome is a single-stranded RNA molecule (Tai et al. 2003). We performed separate first-strand cDNA synthesis reactions with two primers,

263P11 and 263P12 (see Materials and Methods; Table 2.1). The primer 263P12 will bind to and initiate first-strand cDNA synthesis from a positive-stranded RNA molecule, whereas 263P11 will bind to and initiate first-strand cDNA synthesis from a negative-stranded molecule. After treatment with RNase H, these first-strand cDNA reactions were used as templates for PCR with these two primers. A PCR product was obtained only with the first-strand reaction that was performed with 263P12 (Figure 2.3), showing that the genome is positive stranded.

2.2.3 Analysis of HaRNAV structural proteins

We performed N-terminal sequence analyses of protein bands (Figure 2.4; Table 2.2) from purified virus particles. These sequences were found in the amino acid sequence predicted by the large ORF (Figure 2.1), which therefore encodes the viral structural proteins. The approximate locations of the N-termini within the polyprotein sequence are shown in Figure 2.5. Table 2.2 gives the apparent molecular weights of the protein bands based on SDS-PAGE (Figure 2.4) and the predicted molecular weights based on the genome sequence (using the N-termini as guides for the boundaries between proteins in the polyprotein; Figure 2.5). The theoretical molecular weights of the proteins are close to the apparent sizes based on their migration within the gel, but four of five proteins migrated at sizes slightly larger than predicted (Table 2.2). Comparison of the sequences around the protein boundaries (Table 2.2) did not reveal a clear pattern, preventing the identification of a potential consensus protease recognition site. However, three out of five processing sites are on the N-terminal side of a serine residue, and two of the sites share five identical residues (ST-SEI). The lack of a recognizable pattern at the cleavage sites could indicate that more than one protease is involved in processing the polyprotein.

As reported previously (Tai et al. 2003), HaRNAV particles purify in different layers on a sucrose gradient, and particles from these two layers give different protein banding patterns (Figure 2.4). This difference in protein composition may explain why particles from the upper layer are noninfectious while particles in the lower band can infect *H. akashiwo* (Tai et al. 2003). There are several major differences between particles from the upper and lower layers. In the protein gel (Figure 2.4), there is a large amount of protein Band 1 in the particles from the upper layer and comparatively little of this protein in particles from the lower layer, while Bands 5 and 7 are comparatively stronger bands in particles from the lower layer. The reduction of protein

Band 1 and coincident intensification of protein Band 5 are well explained by the N-terminal sequencing results and the genome sequence. The N-terminus of protein Band 5 is located approximately 7 kDa away from the N-terminus of protein Band 1 in the C-terminal direction (i.e. downstream; Figure 2.5). Therefore, it appears that protein Band 5 may result from processing of protein Band 1 during maturation of the viral particles. There is a small amount of the Band 1 protein in the lower sucrose layer (Figure 2.4) and this protein has the same N-terminus (first six residues determined, sequence SEIVEY) as the protein Band 1 in the upper sucrose layer. It is possible that not all of protein Band 1 gets processed or that some "immature" noninfectious particles were in the lower sucrose sample. Similarly, the N-terminus of protein Band 7 is located approximately 4 kDa C-terminally from the N-terminus of protein Band 6 (Figure 2.5). Protein Band 7 may arise from secondary processing of the Band 6 protein or from an alternative cleavage of a precursor protein; both proteins are present in mature capsids. The protein from Band 3 was analyzed by mass spectrometry and the resulting peptide sequences were found in the same region of the polyprotein corresponding to protein Band 2 (Figure 2.5), indicating that Band 3 is a different version of the protein in Band 2. The region of the gel containing protein Bands 4 and 5 (Figure 2.4) was also analyzed by mass spectrometry and the resulting peptide sequences were found in the Bands 2 and 5 regions of the polyprotein (Figure 2.5).

2.2.4 Comparisons of HaRNAV proteins and putative protein domains to other virus sequences

Analysis of the HaRNAV predicted polyprotein sequence by BLAST searches of the NCBI database (Altschul et al. 1997) revealed the presence of a conserved RNA-dependent RNA polymerase (RdRp) domain, a conserved RNA virus RNA helicase domain, and conserved picorna-like virus capsid protein domains (Figure 2.5). Discussions of the individual protein and putative protein domain sequences are below. We used the putative HaRNAV RdRp protein domain sequence corresponding to conserved regions I–VIII (Koonin & Dolja 1993) in a BLAST search of the GenBank database. The most similar sequences found in this search were from a variety of viruses from the *Comoviridae*, *Sequiviridae*, and *Dicistroviridae* families. The top-scoring sequence was from tomato ringspot virus (ToRSV; *Comoviridae*), and an alignment of the HaRNAV sequence (regions I–VIII) with the corresponding region from ToRSV showed that they are 29% identical. Another high-scoring alignment was with acute bee paralysis virus (ABPV; *Dicistroviridae*), and alignment of this sequence with HaRNAV (regions I–VIII)

showed 30% identity. Alignment of the sequences from ToRSV and ABPV showed that these are 30% identical. Therefore, the putative HaRNAV RdRp domain sequence is as similar to virus sequences from two different picorna-like families as these sequences are to each other. An alignment of these three sequences is shown (Figure 2.6). For comparison, we aligned the three RdRp sequences, individually, with the RdRp sequence from tobacco etch virus (TEV: *Potyviridae*) whose RdRp sequence has been shown to tree outside the other picorna-like virus families in phylogenetic analyses (Koonin & Dolja 1993). The HaRNAV, ToRSV, and ABPV sequences are 19%, 23%, and 21% identical to the TEV sequence, respectively, which is lower than the HaRNAV, ToRSV, and ABPV sequences are to each other.

We used the HaRNAV sequence that showed similarity to a central region of a conserved RNA virus RNA helicase domain (residues 524–686 of the polyprotein) for a BLAST search of the NCBI database. Similar to the results found for the putative RdRp domain, this showed the HaRNAV sequence is most similar to sequences from viruses in the *Picornaviridae*, *Dicistroviridae*, and *Comoviridae* families. The top-scoring sequence was from human rhinovirus 14 (HRV14; *Picornaviridae*), followed by sequences from several other rhinoviruses. An alignment of the sequences over the complete region that was input into the BLAST showed the HRV14 sequence is 23% identical to HaRNAV. The next highest scoring sequence is ABPV, which is 21% identical to HaRNAV and 24% identical to HRV14. An alignment of the three sequences is shown (Figure 2.6).

Using the experimentally determined N-termini as boundaries, the individual structural proteins were used for BLAST searches of the NCBI database. A search with the sequence from residues 1776 to 1989 of the polyprotein (corresponding to protein Bands 6 and 7; Figures 2.4 and 2.5) showed that this capsid protein is most similar to capsid proteins from viruses in the *Dicistroviridae* family. There were also lower-scoring alignments with viruses from the *Picornaviridae* and *Caliciviridae* families. The best scoring match was with the VP2 protein from Cricket paralysis virus (CrPV; *Dicistroviridae*); the two sequences were 24% identical over the BLAST-aligned residues. Similarly, the capsid protein sequence from residues 2060 to 2317 (corresponding to Band 5; Figures 2.4 and 2.5) is most similar to the VP3 protein from CrPV, and the top nine scoring sequences were from insect picorna-like viruses. This search showed that this protein is also similar to capsid proteins from viruses in the *Picornaviridae* and *Sequiviridae* families. Interestingly, when the CrPV VP3 sequence is subjected to BLAST, the

HaRNAV protein scores as high as Kashmir bee virus and higher than Taura syndrome virus, both of which are members of the same family as CrPV. An alignment of the HaRNAV and CrPV sequences (Figure 2.6) shows the HaRNAV protein is 24% identical to the CrPV protein. A BLAST search with the capsid protein sequence from residues 2318 to 2581 (corresponding to protein Band 2; Figures 2.4 and 2.5) showed that this protein has less similarity to other viral proteins than the two previously discussed capsid proteins. For this protein, the initial BLAST search returns some low-scoring matches to *Dicistroviridae* capsid proteins, but the second iteration of a PSI-BLAST search finds the CrPV VP1 protein sequence with a high-scoring alignment, suggesting this is the HaRNAV VP1 homologue. A search with the sequence between residues 1990 and 2059 did not return any matches to sequences in the NCBI database.

2.2.5 Phylogenetic analyses of HaRNAV proteins and putative protein domains

We constructed phylogenetic trees in attempts to evaluate the evolutionary relationship of HaRNAV to other viruses. We first compared the relationship of the putative HaRNAV RdRp domain to RdRp sequences from picorna-like viruses. The alignments included residues 1362–1619 of the HaRNAV predicted polyprotein that represent the conserved regions I–VIII (Koonin & Dolja 1993), and the corresponding regions from the other viruses. We chose *Potyviridae* sequences [from tobacco etch virus (TEV) and barley yellow mosaic virus (BaYMV)] as the outgroup because RdRp sequences from these viruses are in a separate lineage relative to other picorna-like viruses (Koonin & Dolja 1993) with which HaRNAV has significant sequence similarity (see above). The maximum likelihood (Fig. 2.7) and neighbor-joining (not shown) trees resolved the *Caliciviridae*, *Picornaviridae*, *Comoviridae*, and *Sequiviridae* sequences into their families, but not the *Dicistroviridae*. Within the *Dicistroviridae*, the viruses we included from the *Cripavirus* genus (DCV, CrPV, TSV, ABPV, RhPV, BQCV, and HiPV) constitute a clade, but with poor support and bootstrap values. The viruses from the *Iflavirus* genus (IFV, SbV, and PnPV) in this family did not resolve into a clade. The maximum likelihood tree (Figure 2.7) does not suggest a close relationship between the HaRNAV sequence and any of the established families, although the neighbor-joining tree (not shown) placed HaRNAV within the *Cripavirus* genus of the *Dicistroviridae* family with a low bootstrap value of 53. Both trees support the placement of the HaRNAV sequence within the picorna-like group, relative to the *Potyviridae* sequences. Interestingly, if the HaRNAV sequence is excluded from the alignments and phylogenetic analyses, the *Cripavirus* clade has much higher maximum likelihood support

and neighbor-joining bootstrap values (75 and 90, respectively; not shown).

Similarly, we constructed phylogenetic trees with part of the HaRNAV putative helicase domain (residues 524–686 of the polyprotein) and the putative RdRp domain sequences concatenated into one, and the corresponding sequences from other picorna-like viruses. Sequences from the *Potyviridae* were not included because they contain helicase sequences with different conserved motifs (Koonin & Dolja 1993). These trees (not shown) did not group the HaRNAV sequence with any other virus family. In addition to the family groups supported in the RdRp analyses (Figure 2.7), these trees additionally supported independent clades for both the Cripaviruses and Iflaviruses with high (>80) maximum likelihood support and neighbor-joining bootstrap values.

The same approach was used to analyze concatenated (putative) helicase/RdRp/VP3-like sequences. Because of greater sequence divergence in the VP3 proteins across the picorna-like virus families, we were able to include fewer families in these analyses. The trees we generated with these sequences (Figure 2.8) supported the same relevant clades as the previous analyses. They also do not support a strong relationship between HaRNAV and the other viruses.

2.3 Discussion

Sequence analyses show that HaRNAV is closely related to viruses from the picorna-like superfamily of viruses. HaRNAV particles are icosahedral with a diameter of approximately 25 nm (Tai et al. 2003), a size and structure consistent with picorna-like viruses (other than the *Potyviridae*). The predicted 5' UTR contains conserved picorna-like sequences and putative structural features. The capsids comprise three major structural protein sequences that are recognizably similar to known picorna-like capsid proteins. The organization of the individual structural proteins, as indicated by sequence similarities, is similar to that found in the *Dicistroviridae*. However, the overall structure of the genome and phylogenetic analyses indicate that HaRNAV does not belong within any of the established picorna-like virus families. Therefore, we propose that HaRNAV is the first member of a new virus family (*Marnaviridae*), which most likely falls within the picorna-like superfamily. It is not surprising that this virus belongs to a previously unknown family because it is the first described ssRNA virus that infects a photosynthetic protist. It seems likely that as more effort is spent looking for these types of

viruses in microorganisms and in marine environments, more novel groups of viruses will be found. Indeed, four new putative groups of picorna-like viruses have been postulated based on RdRp gene sequences amplified from natural communities of marine viruses (Culley et al. 2003).

The HaRNAV genome structure is most like that found in potyviruses (e.g. tobacco etch virus; Allison et al. 1986) in that the putative nonstructural protein domains are located at the N-terminus and the structural proteins are at the C-terminus of a single large polyprotein encoded on a monopartite genome (Figure 2.5). However, the HaRNAV capsid structure is icosahedral, whereas potyvirus capsids are filamentous. Viruses from the *Caliciviridae* also have a similar genome structure, but they encode more than one polyprotein (e.g. feline calicivirus, Carter et al. 1992). None of the database searches or phylogenetic analyses suggested that HaRNAV is closely related to viruses from either of these families. The Cripaviruses in the *Dicistroviridae* and the unassigned insect picorna-like virus, *Acyrthosiphon pisum* virus (APV), also encode their structural proteins in the 3' region of the genome (Johnson & Christian 1998, Mari et al. 2002, Moon et al. 1998, Nakashima et al. 1999, Sasaki et al. 1998, Van der Wilk et al. 1997, van Munster et al. 2002). However, these viruses encode the capsid and nonstructural polyproteins in distinct ORFs.

The organization of the individual structural protein units within the capsid polyprotein region of HaRNAV appears to be similar to the organization in the *Dicistroviridae*. This is based on the sequence relationships between the individual HaRNAV capsid proteins and the CrPV capsid proteins (Figure 2.5), for which the crystal structure has been determined (Tate et al. 1999). N-terminal sequencing data shows that there is a cleavage generating the VP3-like protein from a larger protein present in the noninfectious particles (Figure 2.5, Table 2.2). This smaller protein fragment may be a VP4-like protein, although we have no evidence that the small protein released is associated with mature capsids, and this protein is not recognizably similar to any other sequences when used for a BLAST search of the NCBI database. In the picornaviruses, VP4 is cleaved from the N-terminus of VP0 to generate VP4 and VP2, whereas in the *Dicistroviridae*, VP4 is cleaved from the N-terminus of VP0 to generate VP4 and VP3. Similar to both of these families, HaRNAV capsids comprise three large structural protein sequences (although some proteins appear to occur in multiple bands as discussed below).

One aspect of HaRNAV capsid protein structure appears analogous to structural protein

processing in *P. nuda* picorna-like virus (PnPV). In this virus, multiple capsid protein bands are generated from the same protein region because proteins of different apparent molecular weights share the same N-termini (Wu et al. 2002). These authors speculated that these proteins shared N-termini but were processed differently at their C-termini. We found that different HaRNAV structural proteins contain sequences from the same protein region but have different N-termini. Analysis of the N-termini of the protein Bands 6 and 7 (Figure 2.4) show that these two bands comprise the HaRNAV VP2-like sequences, as indicated by similarity with the CrPV VP2 protein. Analysis by mass spectrometry of capsid proteins from infectious particles in the region of the gel containing Bands 4 and 5 (Figure 2.4) showed that this region contains a mixture of two proteins (from the VP3- and VP1-like regions of the polyprotein, Figure 2.5). The VP1-like sequences were also found in the analysis of Band 3 (Figure 2.4) by mass spectrometry. Therefore, three protein bands (2, 3, and 4; Figure 2.4) contain sequences from the VP1-like region of the polyprotein. It is possible that these proteins differ by processing at the N-terminus, C-terminus, or by another post-translational modification. Two of the three bands appear larger by SDS-PAGE (33 and 32 kDa) than predicted from the genome sequence (29 kDa), as measured from the known N-terminus of the largest band to the end of the polyprotein sequence (Table 2.2, Figure 2.5). An explanation of these observations is that amino acids are not removed from the largest version of the protein to generate the smaller versions, but rather that there are post-translational additions to the proteins. Glycosylation is one possible post-translational modification that has been observed with other algal virus structural proteins (Friess-Klebl et al. 1994, Wang et al. 1993), but experiments need to be done to test this hypothesis.

The HaRNAV protein sequences show a mosaic pattern of relationships to picorna-like virus sequences. However, sequences from the *Dicistroviridae* gave higher scoring matches when the HaRNAV structural proteins were used for database searches. This, and the apparently conserved structural protein organization within the polyprotein, may reflect a closer evolutionary relationship of HaRNAV with these viruses than with the other families of picorna-like viruses. Overall however, our analyses do not suggest that HaRNAV belongs in any of the picorna-like families. This is based on the lack of a consistent pattern of sequence relationships between HaRNAV and the other picorna-like virus families, and the overall genome organization of HaRNAV in comparison with the other families. Outside of the regions indicated in Figure 2.5, none of the predicted HaRNAV protein sequence shows any recognizable similarity to

known protein (viral or other) sequences.

We have determined and analyzed the genome sequence of the ssRNA virus, HaRNAV, infecting the marine unicellular photosynthetic alga *H. akashiwo*. To our knowledge, this is the first genome sequence reported for a positive-stranded RNA virus from an alga or other protist. Our analyses of the genome sequence and structural protein composition indicate that HaRNAV is most closely related to viruses from the picorna-like superfamily. The evidence suggests that HaRNAV is the first representative of a new virus family, which shares the defining characteristics of the picorna-like virus superfamily.

2.4 Materials and Methods

2.4.1 Purification of virus particles

H. akashiwo cultures were grown and infected with HaRNAV (isolate SOG263) as described (Tai et al. 2003). Virus particles were purified from 1.5 l of culture lysate by centrifugation. The lysate was first cleared by centrifugation at 4000 × g for 1 h. The supernatant was then centrifuged for 5 h at 108 000 × g to pellet the viruses which were then resuspended in a total volume of 200 µl of 50 mM Tris (pH 7.6). These samples were centrifuged for 2 min at 4000 × g to remove large material and the resulting supernatant layered on top of a linear 5–35% sucrose (w/v) gradient (in 50 mM Tris; pH 7.6). The gradient was centrifuged for 3 h at 50 000 × g in a Beckman (Palo Alto, USA) SW-40 rotor and the viral bands purified from the gradient as described (Tai et al. 2003).

2.4.2 Determination of HaRNAV genomic sequence

RNA was purified from the viruses using the viral RNA kit (Qiagen, Mississauga, Canada) according to the manufacturer's instructions. The first strand of cDNA synthesis was performed with Superscript II RNase H$^-$ (Invitrogen, Burlington, Canada) according to manufacturer's instructions using either oligo(dT)$_{12-18}$ (Amersham Pharmacia Biotech, Piscataway, USA) or d(N)$_{10}$T as primers. After treatment with RNase H (Invitrogen), this cDNA was used as template for PCR with Platinum *Taq* (Invitrogen) using either d(N)$_{10}$T or viral-specific primers (Table 2.1) or a combination of d(N)$_{10}$T and a viral-specific primer. PCR reactions with viral specific primers (400 nM each primer, 2 mM $MgCl_2$, 400 nM each dNTP) were run as follows: 95 °C for 60 s, followed by 30 cycles of 95 °C for 45 s, 54 °C for 45 s, and

72 °C for 1 min per kilobase of expected product, and a final incubation at 72 °C for 5 min. PCR reactions with $d(N)_{10}T$ [1 mM $d(N)_{10}T$, 400 nM viral specific primer (if applicable), 4 mM $MgCl_2$, 400 nM each dNTP] were run as follows: 95 °C for 60 s, followed by 35 cycles of 95 °C for 45 s, 40 °C for 2 min and 72 °C for 3 min, and a final cycle of 72 °C for 5 min. PCR products were purified with the QIAquick PCR purification system (Qiagen) and ligated with pGEM-T (Promega, Madison, USA) directly when only viral-specific primers were used, or digested with restriction enzymes (*Mbo*I, *Rsa*I, *Hae*III, or *Alu*I; New England Biolabs, Mississauga, Canada) before ligation with pUC19 (Vieira and Messing 1982) when $d(N)_{10}T$ was used.

The 5' and 3' ends of the viral genome were cloned using the 5' and 3' RACE systems (Invitrogen) according to manufacturer's instructions, with the following exceptions. Viral RNA was treated with 400 µg/µl proteinase K (Sigma- Aldrich, Oakville, Canada) for 60 min at 37 °C and subsequently extracted with phenol and precipitated with ethanol before 5' RACE, as described (Johnson & Christian 1998). To reduce secondary structure in the RNA, first-strand cDNA synthesis was done in the presence of 5% dimethyl sulfoxide (DMSO; Sigma-Aldrich) for 3' RACE, and in the presence of 4% DMSO and at 50 °C for the 5' RACE. Primers 263P8 and 263P15 were used as the first strand and nested viral-specific primers, respectively, for amplifying the 5' end. Primer 263P1 was used as the viral-specific primer for the 3' RACE procedure. A total of nine clones from the 5' end and five clones from the 3' end were sequenced.

DNA sequencing was carried out using the universal M13 primers for cloned fragments, and virus-specific primers for cloned fragments and PCR products. Sequencing reactions were done with Big Dye version 3.0 (Applied Biosystems, Foster City, USA) and analyzed by the University of British Columbia Nucleic Acid and Protein Service (NAPS) Facility (Vancouver, Canada). The genome sequence has been deposited in the GenBank database and assigned accession number AY337486.

2.4.3 Protein sequencing

For N-terminal protein sequencing, purified viral particles were subjected to SDS-PAGE and transferred onto a polyvinylidene difluoride membrane (Bio-Rad, Mississauga, Canada) according to the manufacturer's recommendations, and the N-terminal sequences were determined at the University of British Columbia NAPS Facility. For mass spectrometry protein

sequencing, purified viral particles were subjected to SDS-PAGE and Coomassie blue-stained protein bands were excised and sent to the University of Victoria Genome British Columbia Proteomics Center (Victoria, Canada) for analysis by nanospray-quadrupole-time of flight (ESI-Q-TOF) mass spectrometry.

2.4.4 Nucleotide and protein sequence analyses

Identification of potential coding regions and predictions of protein molecular weights were done with DNA Strider version 1.2 (Marck 1988). Analysis of RNA sequences for potential secondary structure was done with the web-based version of the program *mfold* v3.0 (Mathews et al. 1999, Zuker et al. 1999). Database searches were performed with the BLAST algorithm (Altschul et al. 1997).

2.4.5 Sequence alignments

The viruses from which sequences were used for the phylogenetic analyses are listed with their GenBank sequence accession numbers in Table 2.3. Each protein sequence group was aligned using CLUSTAL X v1.81 (Thompson et al. 1997) with the BLOSUM series protein weight matrix (Henikoff & Henikoff 1992). The complete sequence alignments are available from the authors upon request.

2.4.6 Phylogenetic tree construction and presentation

The protein sequence alignments were transformed into maximum likelihood distances using TREE-PUZZLE v5.0 (Strimmer & von Haeseler 1996) and 25000 puzzling steps. The default multiple substitution matrix chosen by TREE-PUZZLE was used [Variable Time (VT), Müller and Vingron 2000]. Trees were plotted using TreeView v1.6.5 (Page 1996). For the RdRp tree, sequences from the *Potyviridae* (TEV and BaYMV) were used for an outgroup because they are in a separate lineage relative to other picorna-like virus RdRp sequences (Koonin & Dolja 1993). Neighbor-joining trees were constructed with PAUP* v4.0 (Swofford 2000), and bootstrap values calculated based on percentages of 1000 replicates are shown on the trees.

2.5 Tables and Figures

Table 2.1 Primers used for cDNA synthesis, PCR and RT-PCR.

Primer	Sequence (5'-3')	Location	Strand based on
263P1	CTCGCTCAACAGGTACACAA	7623-7642	+
263P2	CTTCCCGCATTCAGTTCG	1986-2003	-
263P4	AAAAGTGATGATGTTTGAAGAC	2298-2319	+
263P5	CCGATGTAGAAGTGGGTAGAT	6434-6454	-
263P6	CGATTTTGTGAGCATTGGG	8139-8157	+
263P7	AGCACCCGTAACTTTTCACTGT	2048-2069	-
263P8	GGGTCTAAATCACCACTAACTG	1241-1262	-
263P9	TGGTGATTTGGCTTCTATTT	5916-5935	+
263P10	ACAACTTTCCATACCACCCTC	5043-5063	-
263P11	TGGTACTGCGTGGTTTTACT	3069-3088	+
263P12	ATTTCCGCCGATCTGATT	4281-4298	-
263P13	TACCTACGAGTGTGGAAAATG	3986-4006	+
263P14	CTCTGGTTTGTTGGCGG	7794-7810	+
263P15	TTTTCTGCCTGCTTGACG	591-608	-
263P16	GGTCCGCCGCAAACATCA	666-683	+
263P17	GCCTGTCACCAACTACAAAAT	3662-3682	-
263P18	GGTTGATTGGTGCTTGG	7954-7970	-
263P19	TGGATTCTACACGCAAAGTT	485-504	+

Table 2.2 N-terminal sequences of proteins from purified HaRNAV particles

Band[a]	Protein mol wt (kDa)[b]	N-terminal sequence[c]	Position of N terminus in polyprotein	Theoretical protein mol wt (kDa)[d]	Sequence at cut site
1	39	SEIVEYXKGEHXGGD[e]	1990	36	PTST-SEIV
2	33	SEIISESGADPTLVL[e]	2318	29	FVST-SEII
5	29	SRPDLLGAPEPFVPR[e]	2060	29	LFGY-SRPP
6	26	S(or T)ETLCN[f]	1776	24	EKLL-TETL
7	24	VDGDLASILSAPRTV[e]	1810	20	RPGE-VDGD

[a] As labeled in Figure 2.4

[b] Based on SDS-PAGE

[c] X indicates it was not possible to discern the amino acid identity at this position

[d] Based on the genome sequence data and using the determined N-termini as boundaries

[e] Only the first 15 residues of the determined sequence are shown

[f] Only the first 6 residues were determined

Table 2.3 Summary of viruses used in phylogenetic analyses

Virus	Abbreviation	Accession number
Acute bee paralysis virus	ABPV	NC_002548
Acyrthosiphon pisum virus	APV	NC_003780
Barley yellow mosaic virus	BaYMV	NC_002990
Black queen cell virus	BQCV	NC_003784
Carnation mottle virus	CarMV	NC_001265
Cowpea mosaic virus	CPMV	NC_003549
Cricket paralysis virus	CrPV	NC_003924
Drosophila C virus	DCV	NC_001834
Feline calicivirus	FCV	NC_001481
Heterosigma akashiwo RNA virus	HaRNAV	AY337486
Himetobi P virus	HiPV	NC_003782
Human rhinovirus 14	HRV	NC_001490
Human poliovirus	PV	NC_002058
Infectious flacherie virus	IFV	NC_003781
Parsnip yellow fleck virus	PYFV	NC_003628
Perina nuda picorna-like virus	PnPV	NC_003113
Rabbit hemorrhagic disease virus	RHDV	NC_001543
Rhopalosiphum padi virus	RhPV	NC_001874
Rice tungro spherical virus	RTSV	NC_001632
Sacbrood virus	SbV	NC_002066
Taura syndrome virus	TSV	NC_003005
Tobacco etch virus	TEV	NC_001555
Tomato ringspot virus	ToRSV	NC_003840

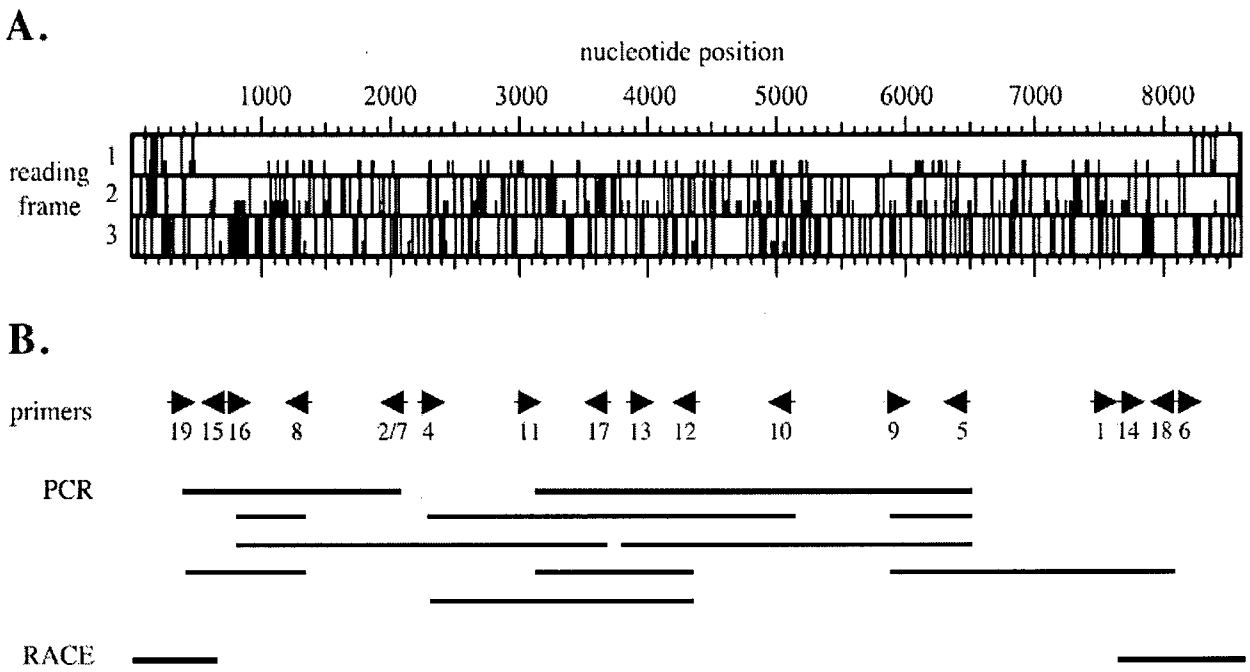

Figure 2.1 Analysis of the HaRNAV genome sequence for open reading frames, and coverage of the genome by PCR and cloning

(A) Analysis for possible open reading frames. For each reading frame (labeled on the left of the figure), potential start codons (AUG) are shown with a half-height line and stop codons (UGA, UAA, and UAG) are shown by full-height lines. Reading frame 1 contains a large open reading frame starting at position 484 that is 7743 nts long and predicted to encode a 2581-amino-acid residue protein. (B) Coverage of the HaRNAV genome with PCR, RACE, and cDNA clones. The approximate locations of primers used (Table 2.1) are shown as arrowheads that reflect the direction of priming for each primer. PCR products obtained with viral-specific primers following first-strand cDNA synthesis (PCR), the 5' and 3' RACE clones (RACE), and cDNA clones obtained by PCR involving $d(N)_{10}T$ primers (cDNA) are shown as lines. Details on the generation of the fragments/clones are given in Materials and Methods.

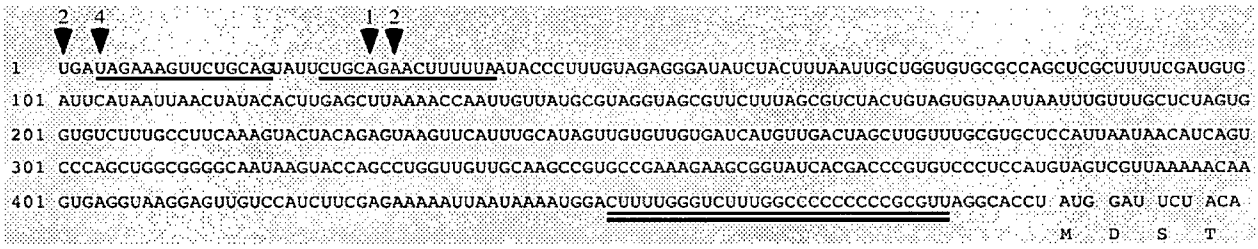

Figure 2.2 Sequence of the 5' untranslated region of the HaRNAV genome

The first 495 nts of the genome sequence are shown. The arrowheads above the sequence mark the locations found as ends in nine 5' RACE clones, and the numbers above the arrowheads indicate the number of clones that ended at each position. The location of the stem of a predicted stem-loop structure at the 5' end is underlined; 14 of the 15 underlined bases can hybridize. The large open reading frame predicted to encode the single HaRNAV polyprotein begins at position 484 and the first four predicted amino acid residues are shown below the nucleotide sequence. A pyrimidine-rich region upstream of the predicted start codon, possibly important for polyprotein translation (Pestova et al. 1991), is double-underlined.

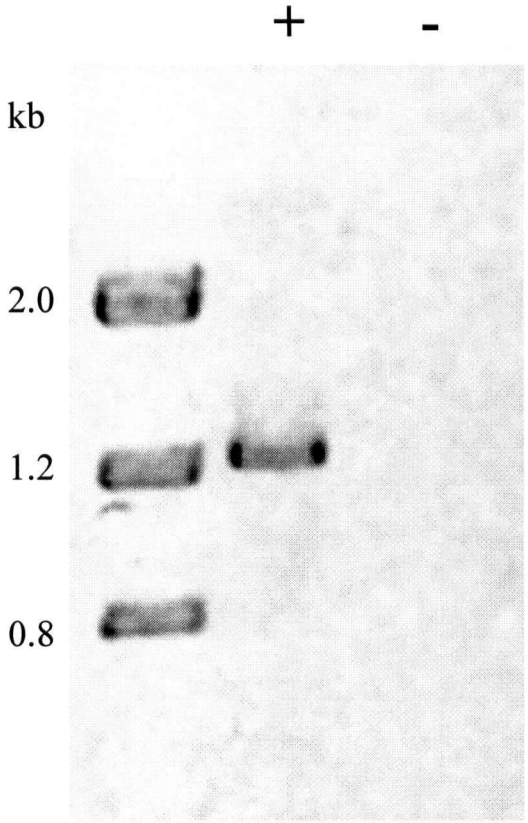

Figure 2.3 Demonstration that the HaRNAV genome is positive-stranded

First strand cDNA synthesis was performed with a primer that would bind to a positive stranded genome (+) or a primer that would bind to a negative-stranded genome (-), followed by PCR with both primers (see Materials and Methods). A product is only visible in the + lane, indicating the genome is positive stranded.

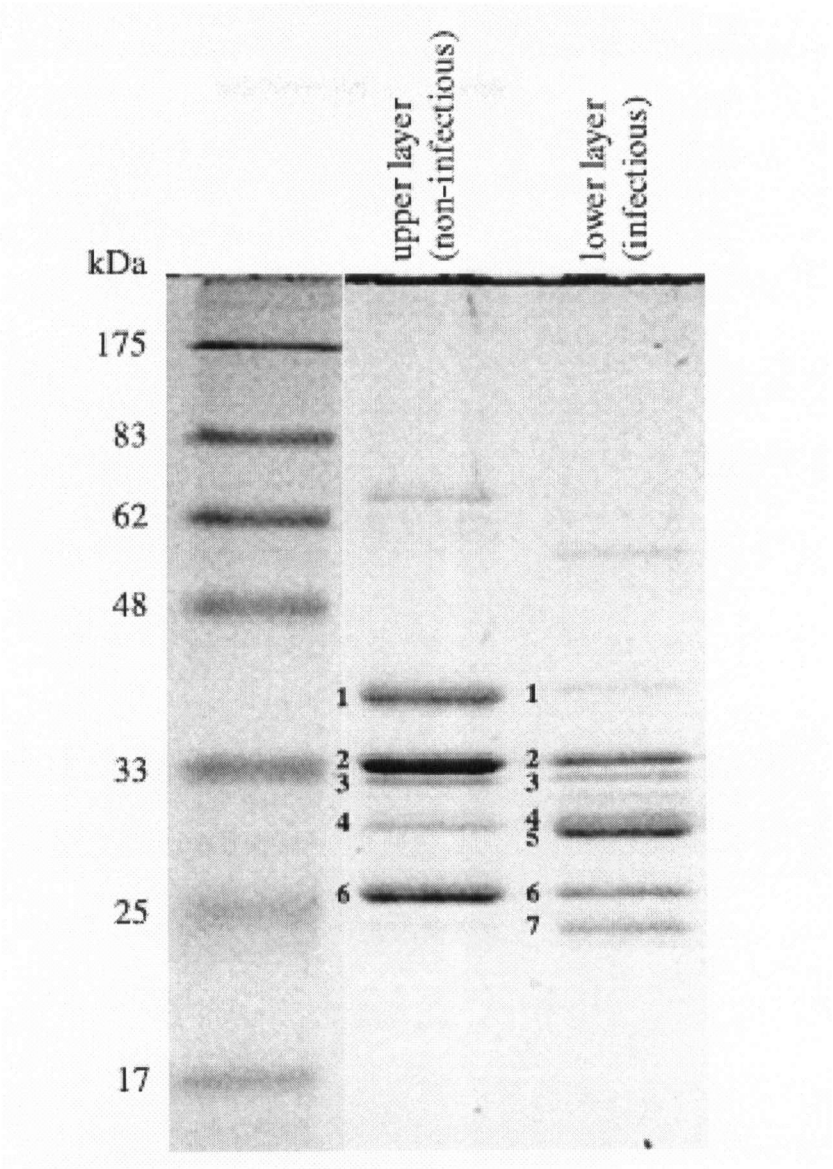

Figure 2.4 Analysis of structural proteins from HaRNAV particles

HaRNAV particles from the noninfectious (upper layer) and the infectious (lower layer) samples obtained during sucrose gradient purification were subjected to SDS-PAGE. Bands discussed in the manuscript are labeled (1– 7).

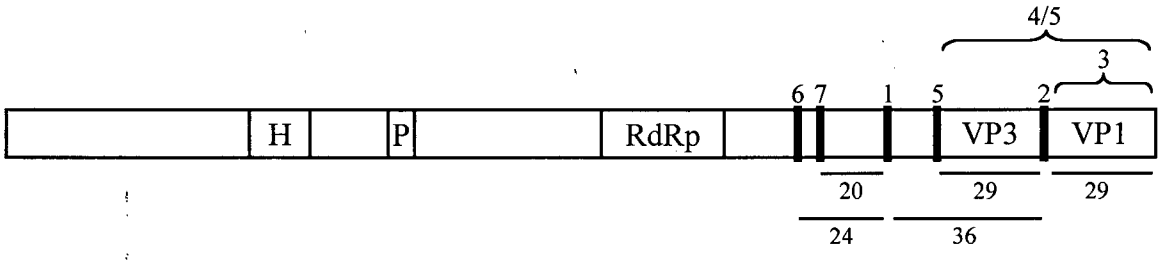

Figure 2.5 Representation of the predicted HaRNAV polyprotein

The N- and C-termini are indicated on the ends. The locations of conserved protein domains are indicated within the polyprotein box: RNA virus helicase (H), protease (P) and RNA-dependent RNA polymerase (RdRp), structural proteins (VP3 and VP1). The location of N-termini found by N-terminal sequencing of the HaRNAV structural proteins are shown by heavy vertical black lines and labeled above by their corresponding protein band number (Figure 2.4, Table 2.2). The theoretical molecular weights (kDa) of proteins between the N-termini are shown below the lines underneath the polyprotein box. The regions of the polyprotein sequence that contain the peptide sequences found by mass spectrometry sequencing of proteins are indicated by the lines above the polyprotein box and labeled with their corresponding band number from Figure 2.4.

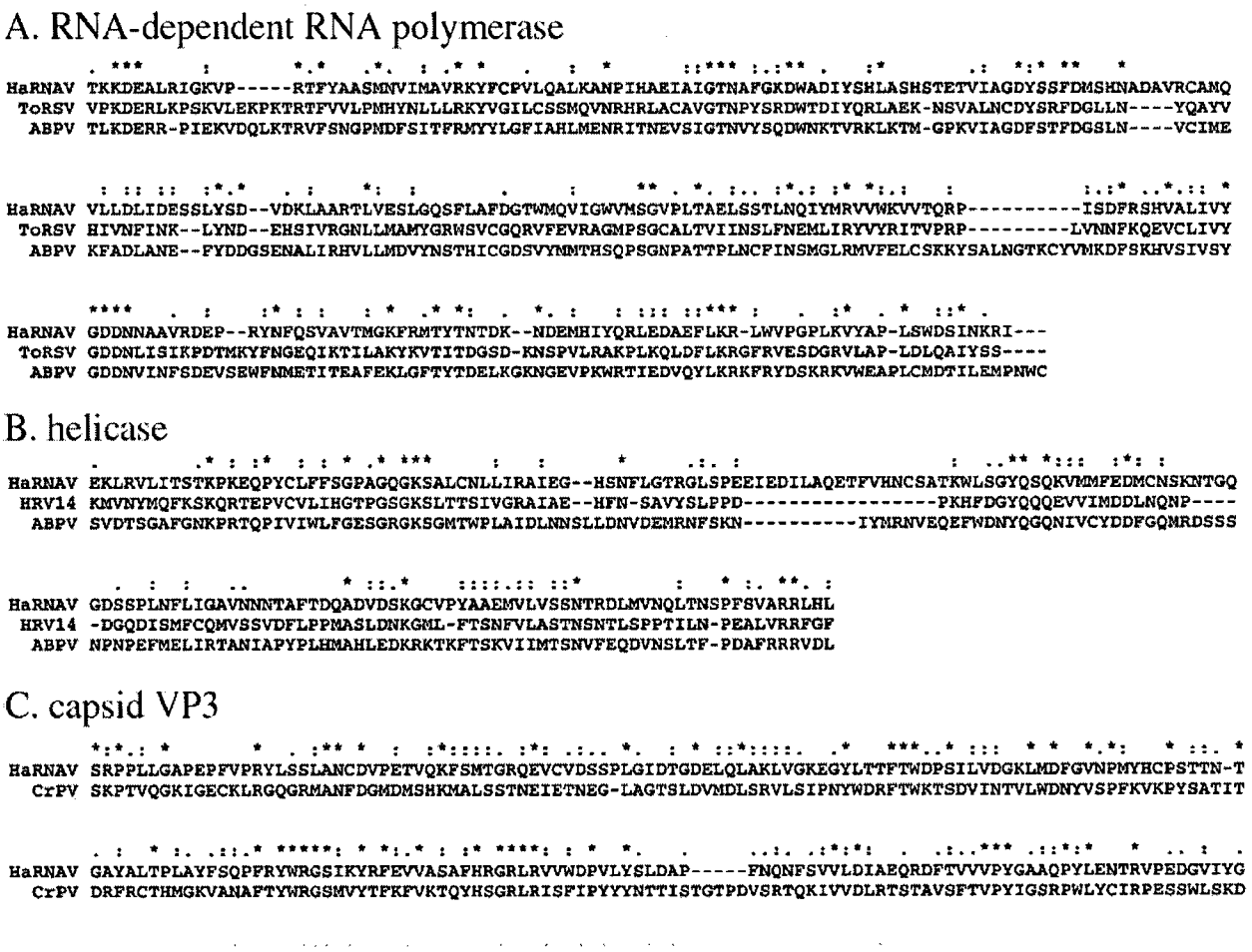

Figure 2.6 Alignment of HaRNAV sequences with sequences from other viruses

Alignments were done with CLUSTAL X v1.81 (Thompson et al. 1997), and the presence of positively scoring groups is indicated above the aligned sequences as defined in CLUSTAL: "*" indicates a fully conserved residue, ":" indicates full conservation of a strong group, and "." indicates full conservation of a weak group. The accession numbers for virus sequences used are given in Table 2.3. (A) Alignment of putative RNA-dependent RNA polymerase protein sequences (corresponding to conserved regions I – VIII; Koonin and Dolja 1993). The alignment comprises residues 1362–1619 of the predicted HaRNAV polyprotein, residues 1691–1950 of the ToRSV polyprotein, and residues 1565–1840 of the ABPV replicase polyprotein. (B) Alignment of putative RNA helicase domain sequences. The alignment comprises residues 524–686 of the predicted HaRNAV polyprotein, residues 1207–1344 of the HRV14 polyprotein, and residues 522–675 of the ABPV replicase polyprotein. (C) Alignment of viral capsid protein sequences. The HaRNAV VP3-like protein sequence (residues 2060–2317 of the predicted sequence, corresponding to protein Band 5; Figures 2.4 and 2.5) was aligned with residues 341–622 of the CrPV structural polyprotein sequence (corresponding to the VP3 capsid protein; Tate et al. 1999).

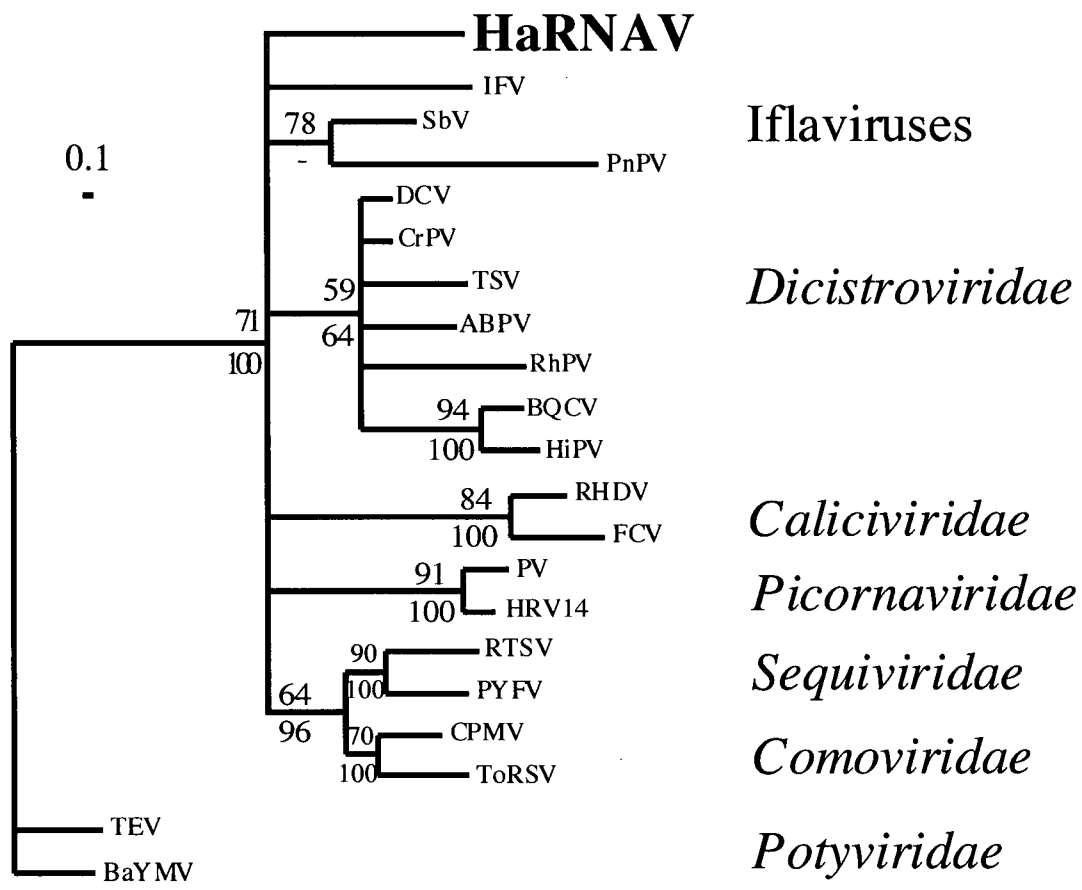

Figure 2.7 Phylogenetic analysis of RNA-dependent RNA polymerase domain protein sequences

Virus classifications are indicated, and the accession numbers and abbreviations for the viruses are listed in Table 2.3. The tree is based on maximum likelihood distances and the *Potyviridae* sequences from tobacco etch virus (TEV) and barley yellow mosaic virus (BaYMV) were used as an outgroup (outgroup clade support values are 71/100, see Materials and Methods). Support values based on 25000 puzzling steps are shown above the branches. Bootstrap values (percentages based on 1000 replicates) for branches that are supported by >50% by neighbor-joining analysis are labeled below the branches (a dash indicates there was no corresponding branch supported by >50% in the neighbor-joining tree). The maximum likelihood distance scale bar is shown.

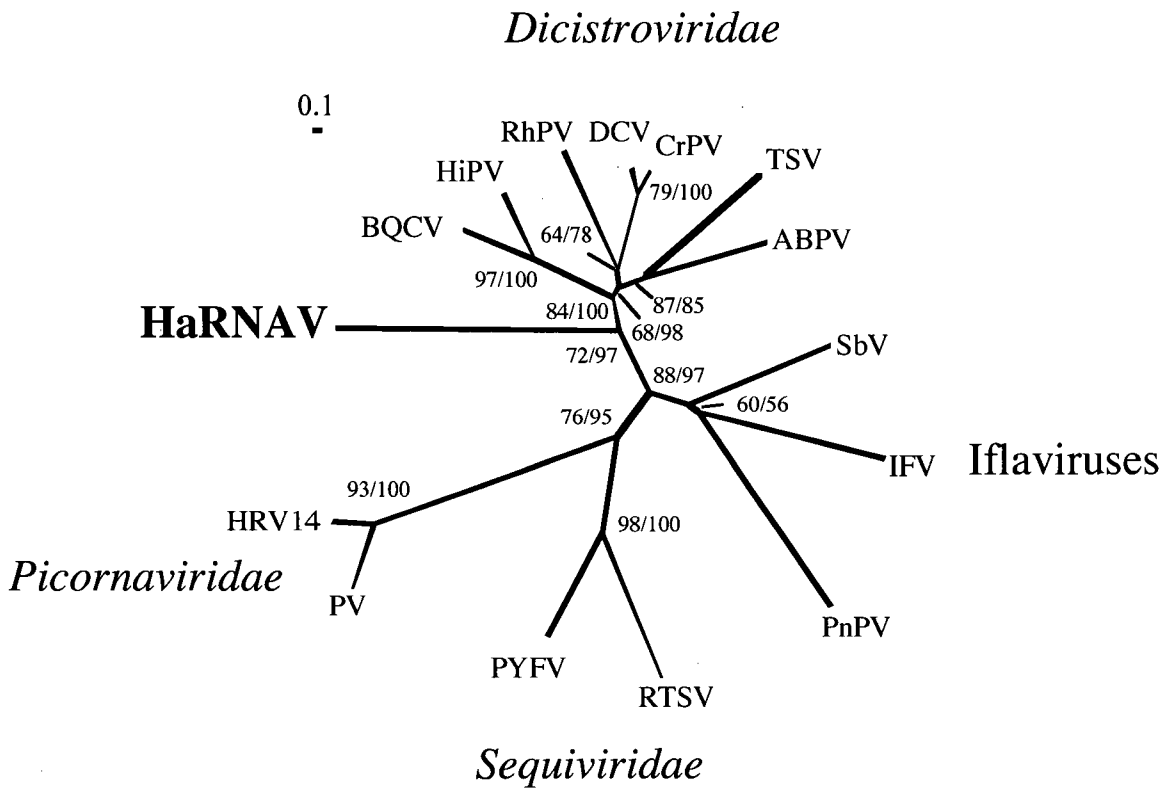

Figure 2.8 Phylogenetic analysis of concatenated (putative) helicase/RdRp/VP3-like capsid protein sequences

Virus classifications are indicated, and the accession numbers and abbreviations for the viruses are listed in Table 2.3. The unrooted tree is based on maximum likelihood distances and the maximum likelihood distance scale bar is shown. Support values based on 25000 puzzling steps are shown for the branches followed by bootstrap values (percentages based on 1000 replicates) from the neighbor-joining analysis.

2.6 References

Allison, R., R. E. Johnston, and W. G. Dougherty. 1986. The nucleotide sequence of the coding region of tobacco etch virus genomic RNA: evidence for the synthesis of a single polyprotein. Virology **154:** 9-20.

Altschul, S. F., T. L. Madden, A. A. Schaffer, J. H. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research **25:** 3389-3402.

Andino, R., G. E. Rieckhof, P. L. Achacoso, and D. Baltimore. 1993. Poliovirus RNA synthesis utilizes an RNP complex formed around the 5' end of viral RNA. EMBO Journal **12:** 3587-3598.

Carter, M. J., I. D. Milton, J. Meanger, M. Bennett, R. M. Gaskell, and P. C. Turner. 1992. The complete nucleotide sequence of a feline calicivirus. Virology **190:** 443-448.

Castberg, T., R. Thyrhaug, A. Larsen, R. A. Sandaa, M. Heldal, J. L. Van Etten, and G. Bratbak. 2002. Isolation and characterization of a virus that infects *Emiliania Huxleyi* (Haptophyta). Journal of Phycology **38:** 767-774.

Cottrell, M. T., and C. A. Suttle. 1991. Wide-spread occurrence and clonal variation in viruses which cause lysis of a cosmopolitan, eukaryotic marine phytoplankter, *Micromonas pusilla*. Marine Ecology Progress Series **78:** 1-9.

Culley, A. I., A. S. Lang, and C. A. Suttle. 2003. High diversity of unknown picorna-like viruses in the sea. Nature **424:** 1054-1057.

Friess-Klebl, A. K., R. Knippers, and D. G. Müller. 1994. Isolation and characterization of a DNA virus infecting *Feldmannia simplex* (Phaeophyceae). Journal of Phycology **30:** 653-658.

Hellen, C. U. T., and P. Sarnow. 2001. Internal ribosome entry sites in eukaryotic mRNA molecules. Genes and Development **15:** 1593-1612.

Henikoff, S., and J. G. Henikoff. 1992. Amino acid substitution matrices from protein blocks. Proceedings of the National Academy of Sciences of the United States of America **89:** 10915-10919.

Jacobsen, A., G. Bratbak, and M. Heldal. 1996. Isolation and characterization of a virus infecting *Phaeocystis pouchetii* (Prymnesiophyceae). Journal of Phycology **32:** 923-927.

Johnson, K. N., and P. D. Christian. 1998. The novel genome organization of the insect picorna-like virus Drosophila C virus suggests this virus belongs to a previously undescribed virus family. Journal of General Virology **79:** 191-203.

Koonin, E. V., and V. V. Dolja. 1993. Evolution and taxonomy of positive-strand RNA viruses: implications of comparative-analysis of amino-acid-sequences. Critical Reviews in Biochemistry and Molecular Biology **28:** 375-430.

Lawrence, J. E., A. M. Chan, and C. A. Suttle. 2001. A novel virus (HaNIV) causes lysis of the toxic bloom-forming alga *Heterosigma akashiwo* (Raphidophyceae). Journal of Phycology **37:** 216-222.

Liljas, L., J. Tate, T. Lin, P. Christian, and J. E. Johnson. 2002. Evolutionary and taxonomic implications of conserved structural motifs between picornaviruses and insect picorna-like viruses. Archives of Virology **147:** 59-84.

Marck, C. 1988. "DNA Strider": a "C" program for the fast analysis of DNA and protein sequences on the Apple Macintosh family of computers. Nucleic Acids Research **16:** 1829-1836.

Mari, J., B. T. Poulos, D. V. Lightner, and J. R. Bonami. 2002. Shrimp Taura syndrome virus: genomic characterization and similarity with members of the genus *Cricket paralysis-like viruses*. Journal of General Virology **83:** 915-926.

Martinez-Salas, E., R. Ramos, E. Lafuente, and S. L. De Quinto. 2001. Functional interactions in internal translation initiation directed by viral and cellular IRES elements. Journal of General Virology **82:** 973-984.

Mathews, D. H., J. Sabina, M. Zuker, and D. H. Turner. 1999. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. Journal of Molecular Biology **288:** 911-940.

Mayer, J. A., and F. J. R. Taylor. 1979. Virus which lyses the marine nanoflagellate *Micromonas pusilla*. Nature **281:** 299-301.

Moon, J. S., L. L. Domier, N. K. Mccoppin, C. J. D'arcy, and H. Jin. 1998. Nucleotide sequence analysis shows that *Rhopalosiphum padi* virus is a member of a novel group of insect-infecting RNA viruses. Virology **243:** 54-65.

Müller, T., and M. Vingron. 2000. Modeling amino acid replacement. Journal of Computational Biology **7:** 761-776.

Nagasaki, K., and M. Yamaguchi. 1997. Isolation of a virus infectious to the harmful bloom causing microalga *Heterosigma akashiwo* (Raphidophyceae). Aquatic Microbial Ecology **13:** 135-140.

Nakashima, N., J. Sasaki, and S. Toriyama. 1999. Determining the nucleotide sequence and capsid-coding region of Himetobi P virus: a member of a novel group of RNA viruses that infect insects. Archives of Virology **144:** 2051-2058.

Page, R. D. M. 1996. TREEVIEW: An application to display phylogenetic trees on personal computers. Computer Applications in the Biosciences **12:** 357-358.

Pestova, T. V., C. U. T. Hellen, and E. Wimmer. 1991. Translation of Poliovirus RNA: role of an essential *cis*-acting oligopyrimidine element within the 5' nontranslated region and involvement of a cellular 57-kilodalton protein. Journal of Virology **65:** 6194-6204.

Sandaa, R. A., M. Heldal, T. Castberg, R. Thyrhaug, and G. Bratbak. 2001. Isolation and characterization of two viruses with large genome size infecting *Chrysochromulina ericina* (Prymnesiophyceae) and *Pyramimonas orientalis* (Prasinophyceae). Virology **290:** 272-280.

Sarnow, P. 2003. Viral internal ribosome entry site elements: Novel ribosome-RNA complexes and roles in viral pathogenesis. Journal of Virology **77:** 2801-2806.

Sasaki, J., N. Nakashima, H. Saito, and H. Noda. 1998. An insect picorna-like virus, *Plautia stali* intestine virus, has genes of capsid proteins in the 3 ' part of the genome. Virology **244:** 50-58.

Strimmer, K., and A. von Haeseler. 1996. Quartet puzzling: A quartet maximum-likelihood method for reconstructing tree topologies. Molecular Biology and Evolution **13:** 964-969.

Suttle, C. A. 2000. Viral infection of cyanobacteria and eukaryotic algae, p. 248-286. *In* C. Hurst [ed.], Viral Ecology. Academic Press.

Suttle, C. A., and A. M. Chan. 1995. Viruses infecting the marine Prymnesiophyte *Chrysochromulina* spp.: Isolation, preliminary characterization and natural abundance. Marine Ecology Progress Series **118:** 275-282.

Suttle, C. A., A. M. Chan, and M. T. Cottrell. 1990. Infection of phytoplankton by viruses and reduction of primary productivity. Nature **347:** 467-469.

Swofford, D. 2000. PAUP*: Phylogenetic Analysis Using Parsimony and other Methods 4.0. Sinauer Associates.

Tai, V., J. E. Lawrence, A. S. Lang, A. M. Chan, A. I. Culley, and C. A. Suttle. 2003. Characterization of HaRNAV, a single-stranded RNA virus causing lysis of *Heterosigma Akashiwo* (Raphidophyceae). Journal of Phycology **39:** 343-352.

Tarutani, K., K. Nagasaki, S. Itakura, and M. Yamaguchi. 2001. Isolation of a virus infecting the novel shellfish-killing dinoflagellate *Heterocapsa circularisquama.* Aquatic Microbial Ecology **23:** 103-111.

Tate, J., L. Liljas, P. Scotti, P. Christian, T. W. Lin, and J. E. Johnson. 1999. The crystal structure of Cricket paralysis virus: the first view of a new virus family. Nature Structural Biology **6:** 765-774.

Taylor, F. J. R. 1990. Red tides, brown tides, and other harmful algal blooms: the view into the 1990s, p. 527-533. *In* E. Graneli, B. Sundstroem, L. Edler and D. M. Anderson [eds.], Toxic Marine Phytoplankton. Elsevier.

Thompson, J. D., T. J. Gibson, F. Plewniak, F. Jeanmougin, and D. G. Higgins. 1997. The CLUSTAL_X Windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Research **25:** 4876-4882.

Van der Wilk, F., A. M. Dullemans, M. Verbeek, and J. Vandenheuvel. 1997. Nucleotide sequence and genomic organization of *Acyrthosiphon pisum* virus. Virology **238:** 353-362.

Van Etten, J. L., and R. H. Meints. 1999. Giant viruses infecting algae. Annual Review of Microbiology **53:** 447-494.

Van Etten, J. L., L. C. Lane, and R. H. Meints. 1991. Viruses and virus-like particles of eukaryotic algae. Microbiology and Molecular Biology Reviews **55:** 586-620.

Van Munster, M., A. M. Dullemans, M. Verbeek, J. Van Den Heuvel, A. Clerivet, and F. Van Der Wilk. 2002. Sequence analysis and genomic organization of Aphid lethal paralysis virus: a new member of the family *Dicistroviridae*. Journal of General Virology **83:** 3131-3138.

Vieira, J., and J. Messing. 1982. The pUC Plasmids, an M13mp7-derived system for insertion mutagenesis and sequencing with synthetic universal primers. Gene **19:** 259-268.

Wang, I. N., Y. Li, Q. D. Que, M. Bhattacharya, L. C. Lane, W. G. Chaney, and J. L. Van Etten. 1993. Evidence for virus-encoded glycosylation specificity. Proceedings of the National Academy of Sciences of the United States of America **90:** 3840-3844.

Wu, C. Y., C. F. Lo, C. J. Huang, H. T. Yu, and C. H. Wang. 2002. The complete genome sequence of *Perina nuda* picorna-like virus, an insect-infecting RNA virus with a genome organization similar to that of the mammalian picornaviruses. Virology **294:** 312-323.

Zuker, M., D. H. Mathews, and D. H. Turner. 1999. Algorithms and thermodynamics for RNA secondary structure prediction: a practical guide, p. 11-43. *In* J. Barciszewski and B. F. C. Clark [eds.], RNA Biochemistry and Biotechnology. Kluwer Academic Publishers.

Chapter III. High diversity of unknown picorna-like viruses in the sea

A version of this chapter has been published

Culley, A.I., A.S. Lang, and C.A. Suttle. 2003. High diversity of unknown picorna-like viruses in the sea. Nature **320**: 1054-1057.

3.1 Introduction

Viruses are extremely abundant and geochemically significant agents of microbial mortality in the ocean (Fuhrman 1999, Wilhelm & Suttle 1999). They comprise a morphologically and genetically diverse array of pathogens, some of which infect heterotrophic bacteria and cyanobacteria, as well as photosynthetic and non-photosynthetic protists (Suttle 2000, Wommack & Colwell 2000, Mann 2003). On the basis of a variety of evidence, marine viral communities have been assumed to consist almost entirely of dsDNA viruses; consequently, little effort has been made to examine natural communities of RNA viruses. There are data however to indicate that marine RNA viruses are also important. For example, rhabdoviruses and paramyxoviruses are negative-sense, ssRNA viruses that are major pathogens of fish (Bernard & Bremont 1995) and marine mammals (Van Bressem et al. 1999), respectively. In addition, picorna-like viruses have been isolated that are pathogens of penaeid shrimp (Mari et al. 2002), seals (Smith 2000) and whales (Smith 2000).

3.2 Results and Discussion

We used available sequences in GenBank to design degenerate primers that target the highly conserved RdRp sequence in picorna-like viruses. These primers in conjunction with RT–PCR were used to assay for the presence of picorna-like viruses in the coastal waters of British Columbia, Canada. With the exception of retroviruses, all RNA viruses encode an RdRp, which is essential for replication. Within the RdRp protein sequence, several motifs have been identified which are homologous among diverse species of RNA viruses (Poch et al. 1989, Koonin & Dolja 1993). Groups based on alignments of these conserved regions are congruent with presently defined RNA virus families (Koonin 1991, Zanotto et al. 1996). Families of picorna-like viruses classified on the basis of RdRp sequence data are congruent with families of picorna-like viruses classified according to virus structure, host and epidemiology. Consequently, sequence analysis can be used to infer the relationship of sequences from natural viral communities to known families of picorna-like viruses.

Viruses were concentrated from seawater using ultrafiltration (Suttle et al. 1991) in the spring and summer of 1996, 1997 and 2000. Viral RdRp sequences were amplified from 13 of the 21 viral communities examined. Amplification occurred in samples collected from

oceanographically diverse environments, including anthropogenically influenced sites, estuarine environments, stations with a well-mixed water column and pristine, highly stratified fjords. Preliminary analysis of the environmental sequences by BLAST (Altschul et al. 1997) searches of the GenBank database gave high similarities to several picorna-like virus family RdRp sequences as well as an unidentified viral sequence from a Chinese clam homogenate (Kingsley et al. 2002). These results suggested that picorna-like viruses are prevalent in a variety of marine waters.

To confirm the amplified products originated from picorna-like viruses, a selection of the amplicons were sequenced, and along with representative sequences from known picorna-like virus families, used to construct phylogenetic trees. All sequenced environmental PCR products translated into continuous amino acid sequences that contained the signature positive-stranded ssRNA virus RdRp motif GDD (Kamer & Argos 1984). Phylogenetic analysis of the RdRp fragment amplified in this study resolved all established picorna-like families (Figure 3.1). Strikingly, none of the environmental sequences fell within established families of picorna-like viruses, but rather into four previously unknown and distantly related groups that we refer to as A, B, C, and D. Our results suggest that at least two and possibly all four of these groups represent new families of RNA viruses. Phylogenetic trees constructed with both maximum-likelihood and neighbor-joining methods group the environmental sequences outside established picorna-like virus families (Figure 3.1). However, low bootstrap support prevents us from drawing any further conclusions regarding the evolutionary relationship between groups B, C, and D. Interestingly, a single sample (JP800) collected from English Bay, which is adjacent to the Strait of Georgia and the city of Vancouver, contained representatives from three of the four novel clades, indicating that a high diversity of picorna-like viruses exists even within a single water sample.

Sequence identity among the clades of environmental picorna-like virus sequences was low, ranging from 38.9% to 54.6% nucleotide identity and 21.8% to 52.9% identity when translated to amino acids. In contrast, sequences within each clade were highly conserved and ranged from 97.7% to 100.0% and 95.9% to 100.0% identical on nucleotide and amino-acid levels, respectively. Interestingly, within group A, although no sequences were identical on a nucleotide level, six were identical when translated into amino acids; this suggests that the observed nucleotide differences may be real and not due to methodological error.

Although four groups of picorna-like viruses were discovered in this study, one of these groups (group A) contains a virus (HaRNAV, Tai et al. 2003) that causes the lysis of *Heterosigma akashiwo*, a toxic-bloom-forming alga that is responsible for major fish deaths in temperate waters (Smayda 1998). Nucleotide sequences from HaRNAV were 98.8% to 99.7% identical to sequences from group A, implying that there are a number of viruses closely related to HaRNAV that belong within this group. Interestingly, we were able to amplify sequences belonging to group A from samples collected in different locations, seasons and years, suggesting that HaRNAV-like viruses are reoccurring and widely distributed in the Strait of Georgia.

Our results indicate that there is a diverse but previously unknown community of picorna-like viruses that are persistently occurring and widespread in the ocean. The fact that sequences from four stations resulted in at least two novel families of picorna-like viruses suggests that the diversity of these viruses in the ocean is high. Branch lengths between groups B, C, and D are similar to those among families and genera of known picorna-like viruses (Figure 3.1), suggesting that these groups are representative of at least three novel genera within a new family, or, in fact, may represent three previously unknown families. However, bias is undoubtedly associated with each step of the methods used in this research, including concentration and extraction of the RNA virus community, primer design, cDNA synthesis and PCR amplification and cloning (Von Wintzingerode et al. 1997). Thus it is more than likely that these results are an underestimate of marine picorna-like virus richness (see Chapter VI for a more detailed discussion of methodological bias).

Viruses are obligate pathogens that generally remain infectious for a relatively short time in natural marine waters (Suttle & Chen 1992). The repeated amplification of picorna-like virus sequences from the same geographic location in water samples collected over a four-year period implies persistent viral production and therefore infection. The significant, high degree of identity between a sequence amplified from clams harvested from Asian waters and marine group B sequences from this study suggests that picorna-like viruses are likely to be present in a wide range of marine environments. Furthermore, the few isolates of marine picorna-like viruses infect ecologically and economically important organisms. These include HaRNAV (Tai et al. 2003), which infects the red tide-forming, fish-killing raphidophyte *Heterosigma akashiwo*, Taura syndrome virus (TSV), which infects penaeid shrimp (Mari et al. 2002), an intensely

farmed, important member of the marine food web, and San Miguel Sea Lion viruses (SMSVs), which infect pinnipeds such as the California Sea Lion (*Zalophus californianus*) (Smith et al. 1973). Ultimately these newly identified viruses should be isolated, sequenced in full and their hosts identified. In the terrestrial environment, picorna-like viruses infect a wide variety of organisms and are responsible for several important diseases; it seems likely that they will prove to be important players in marine ecosystems as well.

3.3 Materials and Methods

3.3.1 Sample collection and preparation

Viruses from 40 to 200 liters of seawater were concentrated from stations in and adjacent to the Strait of Georgia, British Columbia, between May and August during 1996, 1997 and 2000 aboard the CCGS Vector as described (Suttle et al. 1991). Two milliliters from each concentrated viral community was pelleted by ultracentrifugation and RNA extracted from resuspended pellets using Trizol-LS (Invitrogen, Burlington, Canada) as per the manufacturer's protocol.

3.3.2 RT-PCR

The degenerate primers RdRp 1 (positive-strand, 5'-GGA/GGAC/TTACAG/CCIA/GA/TTTTGAT-3') and RdRp 2 (5'-A/CACCCAACG/TA/CCG/TCTTG/CAA/GA/GAA-3') were designed from an alignment of the putative RdRp sequences from several picorna-like viruses in the NCBI database. To confirm the identity of environmental PCR products, the 454-base pair (bp) target fragment includes a highly conserved amino acid motif (GDD) characteristic of the RdRp of positive-strand RNA viruses (Kramer & Argos 1984).

Complementary DNA was synthesized with SuperScript II RNaseH$^-$ Reverse Transcriptase (Invitrogen) with the reagents provided using 5 µl of extracted RNA and the primer RdRp 2. Subsequently, PCR was performed with RdRp 1 and RdRp 2 primers. PCR products were separated on a 1.5% agarose gel and bands of the appropriate size (approximately 500 bp) were excised. Washed agarose plugs were used as the template in a second round of PCR with the RdRp primer set. Positive and negative controls were done in parallel for the entire procedure. These RdRp fragments were either directly cloned or separated using denaturing gradient gel electrophoresis (DGGE).

DGGE was conducted using 25% to 40% linear denaturant gradient, 7% to 8% linear polyacrylamide gradient gels, as described (Short & Suttle 2002). DGGE bands were excised, re-suspended and amplified in a third round of PCR with RdRp primers.

3.3.3 Cloning and sequencing

Second-round PCR fragments or re-amplified excised DGGE bands were cloned in the pGEM-T (Promega) vector using the manufacturer's protocol. Recombinants containing the cloned insert were identified using PCR with universal -21M13 (5'–GTTTTCCCAGTCACGACGTTGTA-3') and M13R (5'-CAGGAAACAGCTATGACC-3') primers. Cloned, second-round PCR products were screened by restriction endonuclease digestion. PCR products from plasmids with DGGE bands and second-round PCR inserts with unique digestion patterns were sequenced. PCR fragments were sequenced at the University of British Columbia Nucleic Acid and Protein Service Facility (Vancouver, Canada). Conserved regions of translated sequences were aligned with CLUSTAL X v1.81 (Thompson et al. 1997) and then transformed into maximum-likelihood distances using the WAG matrix (Whelan & Goldman 2001) in TREE-PUZZLE v5.0 (Schmidt et al. 2002) and 25000 puzzling steps. Neighbor-joining bootstrap values were calculated based on 1000 replicates using FITCH v.3.6 (Fitch & Margoliash 1967). The name, acronym and accession number of viruses used in Figure 3.1 are: Aichi virus (AiV), NC_001918; broad bean wilt virus 2 (BBWV2), AB013615; cowpea mosaic virus (CPMV), NC_003549; Cricket paralysis virus (CrPV), NC_003924; Drosophila C virus (DCV), NC_001834; encephalomyocarditis virus (EMCV), NC_001479; equine rhinitis B virus (ERBV), NC_003983; foot-and-mouth disease virus (FMDV), NC_004004; human rhinovirus A (HRV), NC_001490; maize chlorotic dwarf virus (MCDV), NC_003626; parsnip yellow fleck virus (PYFV), NC_003628; poliovirus (PV), NC_002058; porcine teschovirus (PTV), NC_003985; potato virus Y (PVY), NC_001616; rabbit haemorrhagic disease virus (RHDV), NC_001543; rice tungro spherical virus (RTSV), NC_001632; ryegrass mosaic virus (RGMV), NC_001814; Sapporo virus (SV), U65427; sweet potato mild mottle virus (SPMMV), NC_003797; swine vesicular exanthema virus (VESV), NC_002551; Taura syndrome virus (TSV), NC_003005; tomato ringspot virus (ToRSV), NC_003840; wheat streak mosaic virus (WSMV), NC_001886.

We determined an error rate of 0.3% to 0.9% based on five independent amplifications

using RT–PCR of a picorna-like virus isolate, implying that sequences less than 99.0% identical are probably different.

3.4 Tables and Figures

Table 3.1 Sequence details

Name	Sample Collection Date (mm/dd/yyyy)	Sampling Site	Latitude	Longitude	Depth(m)	Sample Volume (L)
FRP896-1	08/26/1996	Fraser River Plume	49° 08' W	123° 31' N	11	200
FRP896-2	08/26/1996	Fraser River Plume	49° 08' W	123° 31' N	11	200
FRP896-3	08/26/1996	Fraser River Plume	49° 08' W	123° 31' N	11	200
FRP896-4	08/26/1996	Fraser River Plume	49° 08' W	123° 31' N	11	200
FRP896-5	08/26/1996	Fraser River Plume	49° 08' W	123° 31' N	11	200
FRP896-6	08/26/1996	Fraser River Plume	49° 08' W	123° 31' N	11	200
FRP896-7	08/26/1996	Fraser River Plume	49° 08' W	123° 31' N	11	200
JP500-1	05/26/2000	Jericho Pier	49° 16' W	123° 12' N	0	40
JP700-1	07/13/2000	Jericho Pier	49° 16' W	123° 12' N	0	40
JP800-1	08/17/2000	Jericho Pier	49° 16' W	123° 12' N	0	40
JP800-2	08/17/2000	Jericho Pier	49° 16' W	123° 12' N	0	40
JP800-3	08/17/2000	Jericho Pier	49° 16' W	123° 12' N	0	40
JP800-4	08/17/2000	Jericho Pier	49° 16' W	123° 12' N	0	40
JP800-5	08/17/2000	Jericho Pier	49° 16' W	123° 12' N	0	40
JP800-6	08/17/2000	Jericho Pier	49° 16' W	123° 12' N	0	40
JP800-7	08/17/2000	Jericho Pier	49° 16' W	123° 12' N	0	40
JP800-8	08/17/2000	Jericho Pier	49° 16' W	123° 12' N	0	40
JP800-9	08/17/2000	Jericho Pier	49° 16' W	123° 12' N	0	40
JP800-10	08/17/2000	Jericho Pier	49° 16' W	123° 12' N	0	40
JP800-11	08/17/2000	Jericho Pier	49° 16' W	123° 12' N	0	40

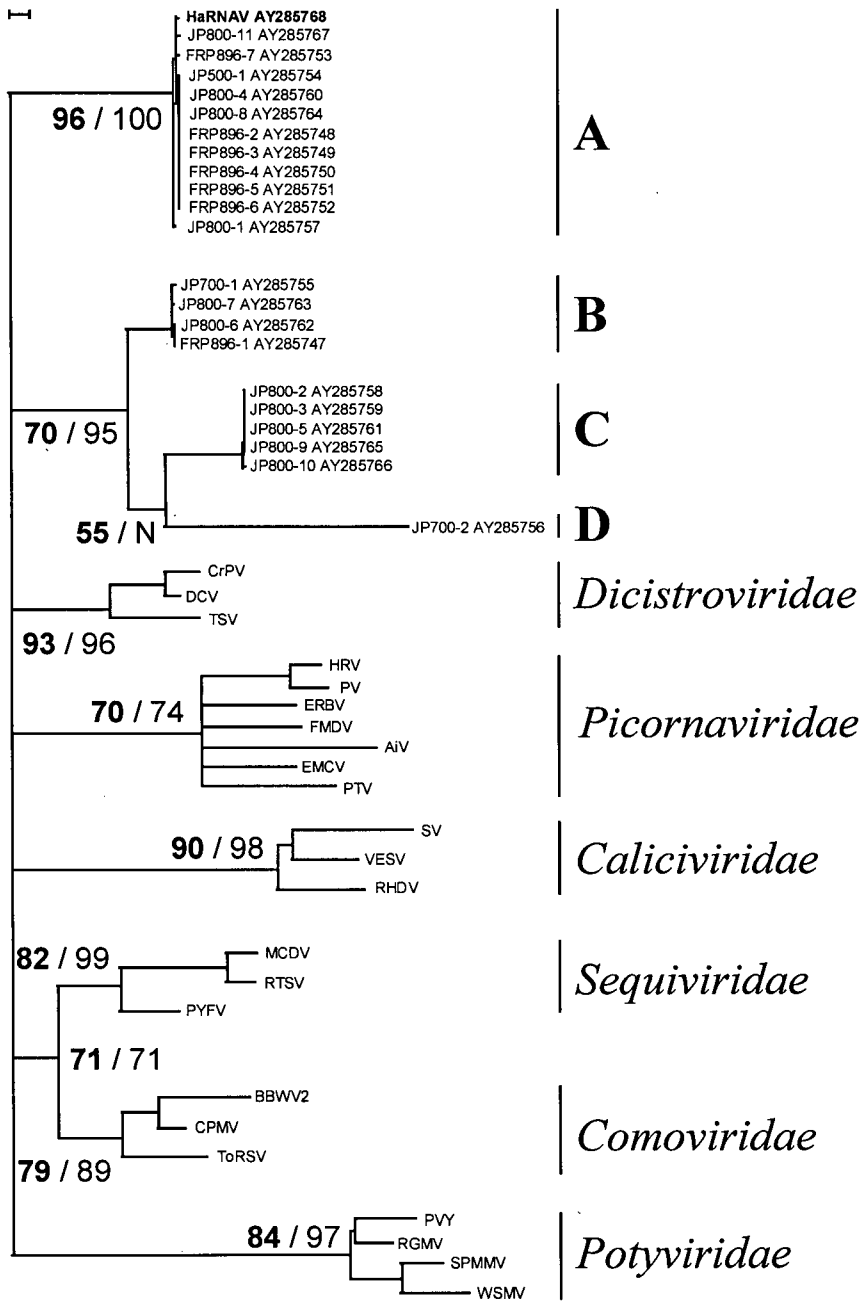

Figure 3.1 Maximum-likelihood tree of RdRp sequences from environmental amplicons and representative viruses from picorna-like virus families

(See Methods for complete virus names). Viruses from the *Potyviridae*, which contain RdRp sequences from a different lineage (Koonin & Dolja 1993), were used as an outgroup. Family names and group letters are shown. Environmental amplicons from coastal British Columbia are labeled by a two or three letter station designation, month, year, group sequence number and NCBI database accession number (SSSMYY-AA, see Table 3.1 for details). TREE-PUZZLE support values are shown for relevant nodes in boldface followed by bootstrap values based on neighbor-joining analysis. N indicates there was no corresponding node in the neighbor-joining tree. The maximum likelihood distance scale bar indicates a distance of 0.1.

3.5 References

Altschul, S. F., T. L. Madden, A. A. Schaffer, J. H. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research **25:** 3389-3402.

Bailey, L., A. J. Gibbs, and R. D. Woods. 1964. Sacbrood virus of the larval honey bee (*Apis mellifera* Linnaeus). Virology **23:** 425-429.

Bernard, J., and M. Bremont. 1995. Molecular biology of fish viruses - a review. Veterinary Research **26:** 341-351.

Fitch, W. M., and E. Margoliash. 1967. Construction of phylogenetic trees. Science **155:** 279-284.

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

Kamer, G., and P. Argos. 1984. Primary structural comparison of RNA-dependent polymerases from plant, animal and bacterial viruses. Nucleic Acids Research **12:** 7269-7282.

Kingsley, D. H., G. K. Meade, and G. P. Richards. 2002. Detection of both hepatitis a virus and Norwalk-like virus in imported clams associated with food-borne illness. Applied and Environmental Microbiology **68:** 3914-3918.

Koonin, E. V. 1991. The phylogeny of RNA-dependent RNA polymerases of positive- strand RNA viruses. Journal of General Virology **72:** 2197-2206.

Lazarowitz, S. G. 2001. Plant Viruses, p. 533-598. *In* D. M. Knipe and P. M. Howley [eds.], Fields Virology. Lippincott, Williams & Wilkins.

Liljas, L., J. Tate, T. Lin, P. Christian, and J. E. Johnson. 2002. Evolutionary and taxonomic implications of conserved structural motifs between picornaviruses and insect picorna-like viruses. Archives of Virology **147:** 59-84.

Mann, N. H. 2003. Phages of the marine cyanobacterial picophytoplankton. FEMS Microbiology Reviews **27:** 17-34.

Mari, J., B. T. Poulos, D. V. Lightner, and J. R. Bonami. 2002. Shrimp Taura syndrome virus: genomic characterization and similarity with members of the genus *Cricket paralysis-like viruses*. Journal of General Virology **83:** 915-926.

Pallansch, M. A., and R. P. Roos. 2001. Enteroviruses: Polioviruses, Coxsackieviruses, Echoviruses, and Newer Enteroviruses, p. 723-775. *In* D. M. Knipe and P. M. Howley [eds.], Fields Virology. Lippincott, Williams & Wilkins.

Poch, O., I. Sauvaget, M. Delarue, and N. Tordo. 1989. Identification of four conserved motifs among the RNA-dependent polymerase encoding elements. EMBO Journal **8:** 3867-3874.

Schmidt, H. A., K. Strimmer, M. Vingron, and A. Von Haeseler. 2002. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics **18:** 502-504.

Short, S. M., and C. A. Suttle. 2002. Sequence analysis of marine virus communities reveals that groups of related algal viruses are widely distributed in nature. Applied and Environmental Microbiology **68:** 1290-1296.

Smayda, T. J. 1998. Ecophysiology and Bloom Dynamics of *Heterosigma akashiwo* (Raphidophyceae), p. 113-131. *In* D. M. Anderson, A. D. Cembella and G. M. Hallegraeff [eds.], Physiological Ecology of Harmful Algal Blooms. Springer-Verlag.

Smith, A. 2000. Aquatic Virus Cycles, p. 447-491. *In* C. Hurst [ed.], Viral Ecology. Academic Press.

Suttle, C. A. 2000. The ecological, evolutionary and geochemical consequences of viral infection of cyanobacteria and eukaryotic algae, p. 248-286. *In* C. J. Hurst [ed.], Viral Ecology. Academic Press.

Suttle, C. A., A. M. Chan, and M. T. Cottrell. 1991. Use of ultrafiltration to isolate viruses from seawater which are pathogens of marine phytoplankton. Applied and Environmental Microbiology **57:** 721-726.

Tai, V., J. E. Lawrence, A. S. Lang, A. M. Chan, A. I. Culley, and C. A. Suttle. 2003. Characterization of HaRNAV, a single-stranded RNA virus causing lysis of *Heterosigma akashiwo* (Raphidophyceae). Journal of Phycology **39:** 343-352.

Thompson, J. D., T. J. Gibson, F. Plewniak, F. Jeanmougin, and D.G. Higgins 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Research **25:** 4876-4882.

Van Bressem, M. F., K. Van Waerebeek, and J. A. Raga. 1999. A review of virus infections of cetaceans and the potential impact of morbilliviruses, poxviruses and papillomaviruses on host population dynamics. Diseases of Aquatic Organisms **38:** 53-65.

Von Wintzingerode, F., U. B. Gobel, and E. Stackebrandt. 1997. Determination of Microbial Diversity in Environmental Samples: Pitfalls of PCR-based rRNA analysis. FEMS Microbiology Reviews **21:** 213-229.

Whelan, S., and N. Goldman. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Molecular Biology and Evolution **18:** 691-699.

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

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

Zanotto, P. M. D., M. J. Gibbs, E. A. Gould, and E. C. Holmes. 1996. A reevaluation of the higher taxonomy of viruses based on RNA polymerases. Journal of Virology **70:** 6083-6096.

Chapter IV. Metagenomic analysis of coastal RNA virus communities

A version of this chapter has been published

Culley, A.I., A.S. Lang, and C.A. Suttle. 2006. Metagenomic analysis of coastal RNA virus communities. Science **313**: 1795-1798.

4.1 Introduction

High mutation rates and short generation times cause RNA viruses to exist as dynamic populations of genetic variants that are capable of using multiple host species (Drake & Holland 1999). In the oceans, the largest ecosystem on Earth, RNA viruses infect ecologically and economically important organisms at all trophic levels including heterotrophic bacteria (Børsheim 1993, see Chapter I section 1.1.8), fish (Kim et al. 2005), crustaceans (Mari et al. 2002), and marine mammals (Rima et al. 2005). Recently, a series of previously unknown RNA viruses have been characterized that infect marine phytoplankton. These include positive-sense single-stranded (ss)RNA viruses (HaRNAV and HcRNAV) that lyse the toxic bloom-formers *Heterosigma akashiwo* and *Heterocapsa circularisquama* (Lang et al. 2004, Nagasaki et al. 2005), a positive-sense ssRNA virus (RsRNAV) that infects the diatom *Rhizosolenia setigera* (Nagasaki et al. 2004), and a double-stranded (ds)RNA virus (MpRNAV) with a genome organization similar to reoviruses that infects the cosmopolitan species *Micromonas pusilla* (Brussard et al. 2004).

Despite the apparent importance of RNA viruses to marine organisms, almost nothing is known about natural communities of RNA viruses in the sea. The most tantalizing evidence that the diversity of RNA viruses in the sea extends well beyond what has been revealed in culture comes from a study that used gene-specific primers to target a subset of picorna-like viruses (Culley et al. 2003). The work showed that these positive-sense ssRNA viruses are persistent, widespread and diverse members of marine virus communities.

Cultivation-independent genomic approaches have recently been used to characterize entire microbial (DeLong & Karl 2005, DeLong et al. 2006) and bacteriophage (Breitbart et al. 2002, Breitbart et al. 2003) assemblages from a diversity of ecosystems. For this study we used randomly reverse-transcribed whole-genome shotgun sequencing to characterize the diversity of uncultivated marine RNA virus assemblages.

4.2 Results and Discussion

Natural virus communities were concentrated from English Bay at Jericho Pier (JP) and the Strait of Georgia (SOG), British Columbia, Canada (Table 4.1). RNA was extracted from the purified virus fraction, reverse-transcribed into cDNA, and used to construct libraries

representative of the natural RNA virus communities (see section 4.3.12 and 4.3.13 for details). Few sequence fragments [37 and 19% for JP and SOG, respectively (Figure 4.1)] showed significant similarity [tBLASTx (Altschul et al. 1997) *E* value < 0.001] to sequences in the National Center for Biotechnology Information (NCBI) database and no similarity to sequences from the Sargasso Sea microbial metagenome (Venter et al. 2004). In contrast, 90% of Sargasso Sea microbial sequence fragments are significantly similar to sequences in the NCBI database (Edwards & Rohwer 2005). These results imply that most RNA viruses in the sea are distantly related to known viruses and that their genetic diversity is much less explored relative to that of the prokaryotic community.

Sequence similarity (tBLASTx *E* value < 0.001) in our samples revealed 98% of sequences belonged to positive-sense ssRNA viruses that infect eukaryotes. The one exception was a sequence with similarity to a dsRNA virus. No RNA phages were detected, supporting arguments that most marine phages have DNA genomes (Weinbauer 2004) and that the predominant hosts of marine RNA viruses are eukaryotes. In addition, no sequences were similar to retroviral or negative-sense ssRNA viruses. Our results are minimum estimates of the richness of marine viral communities because some viruses may have been excluded by our sampling methods (see section 6.1.1 for a discussion of sampling bias). Nonetheless, we observed sequences resembling those of tombusviruses (Lommel et al. 2004), umbraviruses (Taliansky et al. 2004) and nanoviruses (Vetten et al. 2004), all of which are viruses that have not previously been reported from aquatic environments (Figure 4.1 and Table 4.2). Most sequences with significant matches to known sequences (77%) were similar to viral genes with known functions, which is not surprising given the limited number of genes encoded by RNA viruses and their relatively small average genome size (Table 4.3).

The sequence fragments from the two aquatic viral communities were assembled by using a minimal mismatch percentage of 98% and an overlap of 20 base pairs (bp), the most stringent settings given the total introduced error of the RNA virus shotgun library construction method. Simulations demonstrated that these parameters correctly reassembled the genomes of different strains of the same species of RNA viruses from a random assortment of sequence fragments. After assembly, 50% of JP and 36% of SOG sequence fragments overlapped with other sequence fragments and formed contiguous segments (contigs) of overlapping sequence fragments. In the JP library, 66% of the overlapping sequence fell within four large contigs,

which were subsequently assembled into two complete viral genomes that are similar in structure (the JP genome organization schematic shown in Figure 4.2 is representative of both viruses) to each other but that differ from most other known picorna-like viruses (Figure 4.2). In contrast, over 90% of the remaining 14 contigs were formed from three sequence fragments or fewer, indicating that the JP RNA libraries are composed of two very abundant genotypes and others that were relatively rare. Similarly, the genotypic composition of the SOG library was also uneven, with 59% of the sequence fragments forming a single contig that contained most of a novel viral genome, including the 3' untranslated region (UTR), the structural proteins, and all eight conserved regions of the replicase (Koonin & Dolja 1993) (Fig. 4.2). All the remaining sequences fell into contigs composed of two or fewer fragments. Attempts to quantify the structure and diversity of the two RNA virus communities with Phage Communities from Contig Spectrum (PHACCS) (Angly et al. 2005), an online tool designed to estimate the diversity of phage communities on the basis of the frequency of overlapping sequence fragments from whole-genome shotgun libraries, failed primarily owing to the disproportionate contribution of sequence fragments from a small number of dominant genotypes to the total number of contigs in both RNA virus libraries.

The complete genomes assembled from the JP and SOG genomic libraries do not fall within any of the established families of RNA viruses. The JP genomes appear to be dicistronic single molecules of positive-sense ssRNA about 9 kbp in size (Figure 4.2). The JP genomes have characteristics similar to viruses in the proposed order *Picornavirales* (Christian et al. 2005), including the gene order of putative nonstructural genes, a poly(A) tail, a similar G + C content, and core regions of sequence similarity. However, phylogenetic analysis based on aligned RNA-dependent RNA polymerase (RdRp) amino acid sequences placed the JP genomes definitively outside the family *Dicistroviridae* (Figure 4.3), the only dicistronic family of viruses in the proposed order *Picornavirales*. Instead, the sequences fell within a well-supported clade that included HaRNAV, RsRNAV, and SssRNAV, suggesting that they share a common ancestry with viruses that infect marine protists (Figure 4.3). Phylogenies based on alignments of RdRp sequences from RNA viruses were congruent with established family assignments (Culley et al. 2003, Koonin & Dolja 1993) and hence provided a means of classifying unknown RNA virus sequences from the environment. Like the JP genomes, the SOG genome appears to be from a positive-sense ssRNA virus. BLASTp searches and phylogenetic analyses (Figure 4.3), as well

as genomic features such as a putative polymerase domain interrupted by an in-frame termination codon and the absence of obvious helicase motifs (Figure 4.2) (Lommel et al. 2004), indicated similarity to viruses in the family *Tombusviridae* and the unassigned *Umbravirus* genus, which infect flowering plants. However, unlike these viruses, the SOG genome had no detectable movement protein (on the basis of sequence similarity) and is therefore unlikely to be from a virus that infects a terrestrial plant.

In the JP sample, 97% of the significant sequence matches were to viruses in the proposed order *Picornavirales* (Christian et al. 2005) (Figure 4.1 and Table 4.2). Of these, 43% were most similar to HaRNAV, which was first isolated from British Columbia waters (Tai et al. 2003) and which is the lone genome in the database for a picorna-like, phytoplankton-infecting RNA virus. Although the sequences were divergent from HaRNAV, the results suggest that related viruses were important members of the RNA virus community at the JP site. The second most frequent top scoring matches were to picorna-like virus RdRp sequence fragments amplified from the coastal waters of British Columbia (Culley et al. 2003), followed by matches to an array of *Picornavirales* sequences, including viruses infecting higher plants (apple latent spherical virus), arthropods (Taura syndrome virus), and mammals (foot-and-mouth disease virus) (Table 4.2). Nonetheless, the sequences were notably divergent from others in the database, showing that the marine viruses were distantly related to known RNA viruses. One sequence fragment was most similar to a rotavirus sequence, indicating that dsRNA viruses were also likely present. A significant match (tBLASTx *E* value = 3×10^{-20}) to the RdRp of *Sclerophthora macrospora* virus A, an unclassified positive-sense ssRNA mycovirus with a unique genome organization (Yokoi et al. 2003), further illustrates the genetic novelty of marine RNA viruses.

In contrast, in the SOG sample, 73% of the significant sequence matches and the largest contig containing the highest number of sequence fragments were similar to sequences from the *Tombusviridae* (Figure 4.1 and Table 4.2). Known members of this family infect higher plants and have positive-sense ssRNA genomes greater than 5.5 kbp in size (Lommel et al. 2004). Also present in the community were sequences similar to viruses in the genera *Umbravirus* and *Nanovirus* (Figure 4.1 and Table 4.2).

The SOG and JP assemblages had little similarity in community composition. Picorna-

like virus RdRp sequences were amplified from the JP site but not the SOG sample (Culley et al. 2003). tBLASTx searches among sequence fragments from both libraries resulted in 7% of SOG and 8% of JP having significant similarity (E value < 0.001) with each other. Numerous factors may have affected the composition of the JP and SOG virus communities, including salinity [12 parts per thousand (ppt) for JP versus 27 ppt for SOG], interannual variability (JP and SOG samples were collected in 2000 and 2004, respectively), and depth of sampling (JP is a surface sample whereas SOG was collected at 11 m; see Table 4.1 for additional station characteristics). An indirectly shared characteristic between the samples was that 4% of JP and 9% of SOG sequence fragments had significant similarity to the same cripavirus (KBV), although these sequences had no demonstrable homology.

Our results demonstrate that marine RNA virus communities are distantly related to established groups of viruses. Both Bayesian (Altekar et al. 2004) and neighbor-joining (Swofford 2002) phylogenetic analyses of RdRp sequences strongly supported the occurrence of a distinct clade of marine picorna-like viruses. The only known viruses in this clade are HaRNAV, RsRNAV and SssRNAV (Fig. 4.3), all of which infect marine photosynthetic protists; hence, it seems likely that the environmental sequences were also from viruses that infect eukaryotic phytoplankton. Moreover, the large differences between the communities show that the RNA virus populations can differ greatly between two locations (i.e. they are not the same everywhere).

The congruence between RdRp sequences and the established taxonomy of picorna-like viruses (Figure 4.3) suggests that the environmental RdRp sequences likely originate from 10 previously unknown genera of positive-sense ssRNA viruses. A second well-supported clade included two sequences that were related to sequences from viruses in the genus *Cripavirus* (Figure 4.3), a group of viruses known only to infect arthropods. This suggests these environmental sequences may have originated from viruses that infect marine arthropods. Phylogenetic analyses of RdRp sequences from the SOG library and representative members of the family *Tombusviridae* and genus *Umbravirus* indicated that the environmental sequences did not belong within established genera (Fig. 4.3) and clearly supported the existence of other marine RNA viruses that are only distantly related to extant taxa.

Our analyses suggest the existence of a diverse group of RNA viruses that included

sequences related to viruses known to infect marine protists. Compared with the intensive sequencing required to characterize marine prokaryotic communities (Venter et al. 2004), the relatively small genome size of RNA viruses makes the construction of whole-genome shotgun libraries a realistic approach to rapidly survey the diversity of RNA virus communities. In conjunction with the isolation and the sequencing of individual RNA viruses, genomic surveys of RNA virus assemblages is an important step towards a greater understanding of the diversity and ecological impact of these pathogens in the ocean.

4.3 Materials and Methods

4.3.1 Station description

Seawater samples were collected from two stations, JP (Jericho Pier) a site in English Bay adjacent to the city of Vancouver, British Columbia and SOG (Strait of Georgia), a site in the Strait of Georgia, which separates the mainland of British Columbia from Vancouver Island (Table 4.1). Although English Bay opens directly to the Strait of Georgia, the JP station is heavily influenced by freshwater from the Fraser River, whereas the SOG site is more influenced by water from the Pacific Ocean (Leblond 1983) (Table 4.1).

4.3.2 Virus Concentration

Concentrated virus communities were produced as described by Suttle et al. (1991). Briefly, large volumes of seawater from JP (40 l) and SOG (60 l) were sequentially filtered through glass-fiber (nominal pore size 1.2 μm) and 0.45 μm pore-size Durapore PVDF (polyvinylidene fluoride) membranes (Millipore, Cambridge, Canada) to remove eukaryotic plankton and most prokaryotes. The remaining viral size particulate material in the filtrate was concentrated approximately 200-fold through a tangential flow filter cartridge (Millipore) with a 30 kDa molecular-weight cutoff. Remaining bacteria were removed by filtering each concentrate twice through a 0.22 μm pore-size Durapore PVDF membrane (Millipore). Virus-sized particles in each concentrate were pelleted by ultracentrifugation (5 h, 113,000 × g, 4 °C). Pellets were resuspended in sterile buffer (50 μM Tris HCl, pH 7.8) overnight at 4 °C.

4.3.3 RNase treatment and extraction

Before extraction, the concentrated lysates were treated with RNase (Roche, Mississauga, Canada) to remove non-encapsidated RNA (1 U RNase, 30 min incubation at 37 °C). Total

nucleic acids were extracted with the QIAmp Minelute Virus Spin Kit (Qiagen, Mississauga, Canada) according to the manufacturer's instructions. All concentrations cited are final concentrations.

4.3.4 DNase 1 treatment

Total extracted viral nucleic acids were incubated with 1 U DNase 1 (Invitrogen, Burlington, Canada) and 1× DNase 1 buffer for 15 min at room temperature to remove DNA from the sample. The reaction was terminated by adding 2.5 mM EDTA and incubating for 15 min at 65 °C.

4.3.5 Universal rRNA PCR

An aliquot of purified nucleic acids from each sample was used in a PCR reaction with universal 16S primers to confirm the absence of bacterial nucleic acids in the sample using the following conditions: 1× Platinum *Taq* buffer (Invitrogen), 1.5 mM $MgCl_2$, 0.2 mM each dNTP, 0.2 µM each universal rRNA primers GM3F (5'-AGAGTTTGATCCTGGC-3') and 907R (5'-CCGTCAATTCCTTTGAGTTT-3') (Nübel et al. 1997), 1 U of Platinum *Taq* DNA polymerase and 2 µl of template in a final volume of 50 µl. The following thermocycler conditions were used: 94 °C for 2 min, followed by 30 cycles of 94 °C for 60 s, 55 °C for 60 s and 72 °C for 90 s, and a final extension stage of 10 min at 72 °C. Aliquots of the amplification products were electrophoresed in a 1.5% agarose gel in 0.5× TBE buffer (45 mM Tris-borate, 1 mM EDTA [pH 8.0]). Gels were stained with ethidium bromide and visualized under UV illumination.

4.3.6 cDNA synthesis

The construction of a random shotgun clonal library from a community of viral RNA genomes uses some reagents and procedures found in the Superscript choice system for cDNA synthesis (Invitrogen). The overall approach is based on the Linker Amplified Shotgun Library method described at http://www.sci.sdsu.edu/PHAGE/LASL/ and first used to produce shotgun libraries of marine phages (Breitbart et al. 2002). To randomly synthesize cDNA from each extracted community of RNA virus genomes, 100 ng of random hexamers were added per sample and each reaction was heated to 70 °C for 10 min and immediately put on ice. 1× RT buffer, 10 mM DTT and 500 µM of each dNTP were added and the reaction incubated at 37 °C.

After 2 min, 1 U of Superscript II reverse transcriptase (Invitrogen) was added and cDNA synthesis was performed at 37 °C for 60 min.

4.3.7 Second-strand synthesis

Double-stranded cDNA fragments were synthesized from each first strand synthesis reaction using nick translational replacement of genomic RNA (Okayama & Berg 1982) with the following conditions. 1× second strand buffer (Invitrogen), 250 μM of each dNTP, 10 U of *E. coli* DNA Ligase (Invitrogen), 40 U of *E. coli* DNA polymerase (Invitrogen) and 2 U of *E. coli* RNase H (Invitrogen) were added directly to each first strand reaction in a final volume of 150 μl. This reaction was incubated at 16 °C for 2 h. To ensure products were blunt-ended, 10 U of T4 DNA polymerase (Invitrogen) was added to each reaction and incubated for an additional 5 min at 16 °C. Finally, EDTA was added to a final concentration of 30 mM to each sample to terminate the reaction.

4.3.8 Adapter addition

Blunted, double-stranded products from each sample were extracted with phenol:chloroform:isoamyl alcohol (25:24:1) in preparation for adapter ligation. EcoR1 (Not1) (Invitrogen) adapters were added in the following reaction: 18 μl double-stranded cDNA sample resuspended in DEPC-treated water, 1× adapter buffer, 14 mM DTT, 200 μg/ml adapters and 100 U/ml T4 DNA ligase (Invitrogen). This reaction was incubated for 18 h at 16 °C afterwards, followed by heat inactivation by a 10 min incubation at 70 °C.

4.3.9 Column chromatography

To increase the probability of cloning larger fragments, samples were size fractionated with a Sephacryl column (Invitrogen) according to the manufacturer's instructions. Fractions theoretically greater than 600 bp were EtOH precipitated and re-suspended in water in preparation for PCR.

4.3.10 Adapter-targeted PCR

Attempts to clone products directly failed due to insufficient insert concentration and therefore an amplification step was necessary. PCR was performed using a primer targeting a site on the EcoR1 (Not1) adapter. The conditions were as follows: 1× Platinum *Taq* buffer, 3 mM $MgCl_2$, 0.2 mM of each dNTP, 0.8 μM of primer

(5'-CGGCCGCGTCGAC-'3), and 2.5 U of Platinum *Taq* DNA polymerase in a final volume of 50 µl. To reduce the chances of PCR-generated errors in early cycles (Rohwer et al. 2001), each PCR for each size fraction was divided into 5 smaller reactions of 10 µl each and thermo-cycled with the following conditions: 94 °C for 75 s, followed by 25 cycles of 94 °C for 45 s, 55 °C for 45 s and 72 °C for 150 s, and a final extension for 10 min at 72 °C. Completed reactions were pooled and the products purified with a PCR Minelute cleanup kit (Qiagen).

4.3.11 Cloning & Sequencing

Products from each purified reaction were cloned with a TOPO TA Cloning Kit (Invitrogen) as per the manufacturer's protocol. Clones were analyzed by PCR with the vector-specific primers T3 (5'-ATTAACCCTCACTAAAGGGA-3') and T7 (5'-TAATACGACTCACTATAGGG-3'). After electrophoresis, the remaining PCR products from reactions demonstrating inserts greater than 600 bp were purified with a PCR Minelute cleanup kit (Qiagen) and sequenced using Applied Biosystems BigDye v3.1 Terminator Chemistry. Sequencing services were provided by University of British Columbia's Nucleic Acid and Protein Service Facility (Vancouver, Canada).

4.3.12 Sequence fragment classification

Clones were sequenced with the T3 forward primer, resulting in 247 sequences in JP and 108 in SOG. These sequences were compared against GenBank with tBLASTx (Altschul et al. 1997) and the Sargasso Sea environmental metagenome (Venter et al. 2004) with BLASTx. A sequence was considered significantly similar if BLAST *E* values were < 0.001. Based on the sequence with the tBLASTx hit with the lowest *E* value, sequences were classified into one of three biological categories, viral, bacterial or unknown. tBLASTx searches between the two libraries was performed with the stand-alone BLAST tools available at http://www.ncbi.nlm.nih.gov/BLAST/download.shtml.

Despite no amplification in either sample with universal rRNA primers (see above) BLAST searches showed that some clones were significantly similar to bacterial sequences (11% of the JP sequences and 37% of the SOG clones). A similar percentage of bacteria-like sequences were present in shotgun RNA libraries from the RNA virus community of human feces (Zhang et al. 2006). It is possible that in seawater some RNA and DNA is bound to virus-size particles, but

is resistant to enzyme digestion. Almost all of these "contaminating" sequences had significant similarity to bacterial 16S or 23S ribosomal RNA sequences with *E* values < 1 e^{-100}. Furthermore, only sequences similar to bacteria had significant similarity with sequences in the Sargasso Sea metagenome, while sequences classified as viral and unknown showed no similarity to sequences in this database whatsoever. Sequences from JP and SOG identified as bacterial had an average % G+C distinctly higher than the average % G+C for sequences in the unknown and viral categories. Because our objective was to characterize the RNA virus community alone, all sequences identified as bacterial based on the above criteria, were removed from subsequent analyses. This resulted in 216 and 61 sequences classified as viral or unknown in the JP and SOG libraries, respectively.

4.3.13 Contig assembly

Sequence fragments were assembled into overlapping segments using Sequencher v4.5 (Gene Codes, Ann Arbor, U.S.A.) based on a minimum match percentage of 98 and a minimum bp overlap of 20. These are the most stringent conditions feasible given the error introduced during the process of library construction. The distribution of contigs for the JP community was: one contig of 23 sequence fragments, one contig of 20, one contig of 18, one contig of 11, one contig of 6, two contigs of 4, one contig of 3, 9 contigs of two and 109 contigs of one. The contig distribution for the SOG community was: one contig of 13 sequence fragments, one contig of 5, two contigs of two and 39 contigs of one. In a simulation with sequence fragments from a variety of taxonomic levels of picorna-like viruses, we found that these assembly conditions correctly assembled genomes of different strains of RNA viruses belonging to the same species.

4.3.14 Bias

See Chapter VI 6.1.1 *Bias* for a discussion of bias in the library construction method

4.3.15 Phylogenetic analyses

The other viruses used in phylogenetic analyses are listed in Table 4.4. Translated sequences of viruses were aligned using CLUSTAL X v1.83 with the Gonnet series protein matrix (Thompson et al. 1997). Alignments were transformed into likelihood distances with Mr Bayes v3.1.1 (Altekar et al. 2004) and 250000 generations. Neighbor-joining trees were constructed with PAUP v4.0 (Swofford 2002), and bootstrap values calculated based on

percentages of 10000 replicates.

4.3.16 cDNA synthesis for picorna-like RdRp RT-PCR and DGGE

The degenerate primers RdRp 1 (5'-GGA/GGAC/TTACAG/CCIA/GA/TTTTGAT-3') and RdRp 2 (5'-A/CACCCAACG/TA/CCG/TCTTG/CAA/GA/GAA-3') described in Culley et al. (2003) (Chapter III) were used to assay JP and SOG libraries for picorna-like viruses. cDNA was synthesized by combining RNA virus template, 0.1 µmol primer RdRp 2, 0.5 mM of each dNTP and incubating for 5 min. at 65 °C. After cooling on ice, 1× first strand buffer, 5 mM DTT, 40 U RNaseOUT and 200 U of Superscript III RT (Invitrogen) were added and the reaction held at 50 °C for 50 min followed by 70 °C for 15 min to inactivate the enzyme. To degrade the remaining RNA template, 1 U of RNase H was added and the sample incubated at 37 °C for 20 min.

4.3.17 PCR with degenerate primers

The cDNA template was used in PCR with RdRp 1 and RdRp 2 primers. Each reaction consisted of 1× Platinum *Taq* buffer, 3.0 mM $MgCl_2$, 0.2 mM each dNTP, 1.0 µM each RdRp 1 and RdRp 2, and 1 U Platinum *Taq* DNA polymerase in a final volume of 50 µl. The following thermocycler conditions were used: 94 °C for 75 s, followed by 40 cycles of 94 °C for 45 s, 50 °C for 45 s and 72 °C for 1 min, and a final extension stage of 5 min at 72 °C. PCR products were separated on a 1.5% agarose gel in 0.5× TBE buffer. To produce enough product for DGGE, bands of the appropriate size (approximately 500 bp) were excised with a sterile pipette, suspended in 0.5× TBE and heated to 80 °C for 10 min. Aliquots of washed agarose plugs were used in a second round of PCR with the RdRp primer set using slightly more stringent PCR conditions (1.5 mM $MgCl_2$) and 25 cycles with the above thermocycling protocol. Positive and negative controls were done in parallel for the entire procedure.

4.3.18 DGGE

Picorna-like virus RdRp fragments were separated with DGGE as described by Short and Suttle (Short & Suttle 2003). We ran gels with 30 to 50% linear denaturing gradients and 7 to 8% polyacrylamide for 16 h in 1× TAE buffer (40 mM Tris, 20 mM sodium acetate, 1 mM EDTA [pH 8.5]) at 80 V and 60 °C in a D-code electrophoresis system (Bio-Rad Laboratories, Hercules, U.S.A.). The gels were then stained in 0.1× SYBR Gold (Molecular Probes, Eugene, U.S.A.) for

4 h. Selected bands were excised, re-suspended in 1× TAE, heated to 80 °C for 10 min and amplified in a third round of PCR with RdRp primers as described above. PCR products from each reaction were purified with a PCR Minelute cleanup kit and cloned with a TOPO TA Cloning Kit (Invitrogen) as per the manufacturer's protocol. Inserts were sequenced in one direction.

4.3.19 Accession numbers

Sequences have been deposited in GenBank with accession numbers DX420985-DX421142 and DQ439712-DQ439732.

4.4 Tables and Figures

Table 4.1 Characterization of sampling sites. The location is given in degree decimal format. A chlorophyll a value was not available (n.a.) for the SOG sample. We did not observe a bloom at either station during sampling.

Parameter	JP	SOG
Date (mm/dd/yyyy)	06/29/2000	07/13/2004
Location (Latitude, Longitude)	49.27, -123.20	49.86, -124.60
Depth (m)	< 1	11
Temperature (°C)	18	14
Salinity (ppt)	12	27
Volume collected (L)	40	60
Chlorophyll a (μgL^{-1})	3.0	n.a.
Tide	Ebb	Flood

Table 4.2 Identification of the top tBLASTx matches (*E* value < 0.001, n = 92) of environmental sequences from JP and SOG libraries with the Genbank database. A number in bold indicates the highest percentage of matches in each sample, and (-) indicates the virus family, genus or species was not present.

Family/Genus	JP (%) total)	SOG(%) total)	Virus species	JP(%) total)	SOG(%) total)
Comoviridae	2	-	Apple latent spherical	1	-
			Bean pod mottle	1	-
Dicistroviridae	31	9	Acute bee paralysis	1	-
			Aphid lethal paralysis	5	-
			Black queen cell	5	-
			Cricket paralysis	2	-
			Drosophila C	6	-
			Kashmir bee	4	9
			Plautia stali intestine	1	-
			Rhopalosiphum padi	4	-
			Solenopsis invicta	1	-
			Taura syndrome	1	-
Marnaviridae	**43**	-	*Heterosigma akashiwo* RNA	**43**	-
Nanoviridae	-	9	Subterranean clover stunt	-	9
Picornaviridae	2	-	Foot-and-mouth disease	2	-
Reoviridae	2	-	Human rota-	2	-
Tombusviridae	-	**72**	Hibiscus chlorotic ringspot	-	**27**
			Galinsoga mosaic	-	18
			Tobacco necrosis A	-	18
			Pea stem necrosis	-	9
Umbravirus	-	10	Groundnut rosette	-	10
Unclassified	20	-	*Sclerophthora macrospora* A	1	-
			unid. chinese clam assoc.	6	-
			unid. picorna-like. grp B	4	-
			unid. picorna-like. grp C	10	-

Table 4.3 Classification of significant tBLASTx matches (*E* value < 0.001, n = 92) to viral sequences into protein categories.

Protein classification	% of total viral hits
RNA-dependent RNA polymerase	**39**
Capsid	33
Unidentified structural	16
Unidentified nonstructural	7
Helicase	3
RNA binding protein	1
Replication initiator protein	1

Table 4.4 Sequences used in phylogenetic analyses.

Virus Group	Virus Acronym	Full Name	NCBI Accession #
Cheravirus	ALSV	Apple latent spherical virus	NC_003787
	CRLV	Cherry rasp leaf virus	NC_006271
Comoviridae	BBWV1	Broad bean wilt virus 1	NC_005289
	CPMV	Cowpea mosaic virus	NC_003549
	TRSV	Tobacco ringspot virus	NC_005097
Dicistroviridae	ABPV	Acute bee paralysis virus	NC_002548
	CrPV	Cricket paralysis virus	NC_003924
	DCV	*Drosophila* C virus	NC_001834
	PSIV	*Plautia stali* intestine virus	NC_003779
	TSV	Taura syndrome virus	NC_003005
	TrV	Triatoma virus	NC_003783
Iflavirus	DWV	Deformed wing virus	NC_004830
	KV	Kakugo virus	NC_005876
	VDV	*Varroa destructor* virus	NC_006494
Marnaviridae	HaRNAV	*Heterosigma akashiwo* RNA virus	NC_005281
Picornaviridae	EMCV	Encephalomyocarditis virus	NC_001479
	ERBV	Equine rhinitis B virus 1	NC_003983
	FMDV	Foot-and-mouth disease virus A	NC_011450
	HRV	Human rhinovirus 89	NC_001617
	PV	Poliovirus	NC_002058
Sadwavirus	SDV	Satsuma dwarf virus	NC_003785
	NIMV	Navel orange infectious mottling virus	AB_022887
Sequiviridae	PYFV	Parsnip yellow fleck virus	NC_003628
	RTSV	Rice tungro spherical virus	NC_001632
Tombusviridae	CaRMV	Carnation mottle virus	NC_001265
	CRSV	Carnation ringspot virus	NC_003530
	MCMV	Maize chlorotic mottle virus	NC_003627
	OCSV	Oat chlorotic stunt virus	NC_003633
	PMV	Panicum mosaic virus	NC_002598
	PoLV	Pothos latent virus	NC_000939
	TBSV	Tomato bushy stunt virus	NC_001554
	TNV-A	Tobacco necrosis virus A	NC_001777

Virus Group	Virus Acronym	Full Name	NCBI Accession #
Umbravirus	CMoMV	Carrot mottle mimic virus	NC_001726
	GRV	Groundnut rosette virus	NC_003603
	PEMV-2	Pea enation mosaic virus-2	NC_003853
	TBTV	Tobacco bushy top virus	NC_004366
Unclassified	SssRNAV	*Schizochytrium* single-stranded RNA virus	NC_007522
	RsRNAV	*Rhizosolenia setigera* RNA virus	AB243297
	B	Unidentified picorna-like virus JP700-1	AY_285755
	C	Unidentified picorna-like virus JP800-1	AY_285758
	D	Unidentified picorna-like virus JP700-2	AY_285756
Environmental	JP-A	Environmental sequence JP.418.600-5465	DQ439729
	JP-B	Environmental sequence JP.418.600-4289	DQ439728
	5d	Environmental sequence JP.418C.600-5	DX421064
	6d	Environmental sequence JP.418C.600-6	DX421065
	9d	Environmental sequence JP.418C.600-9	DX421066
	11d	Environmental sequence JP.418C.600-11	DX421067
	16d	Environmental sequence JP.418C.600-16	DX421068
	20d	Environmental sequence JP.418C.600-20	DX421069
	32	Environmental sequence JP.418D.600-32	DX421081
	62	Environmental sequence JP.418A.600-62	DX420998
	162	Environmental sequence JP.418D.600-162	DX421094
	1743	Environmental sequence JP.418.600-1743	DQ439724
	SOG-A	Environmental sequence SOG.658.704-3093	DQ439732
	399	Environmental sequence SOG.658.704-399	DX421139

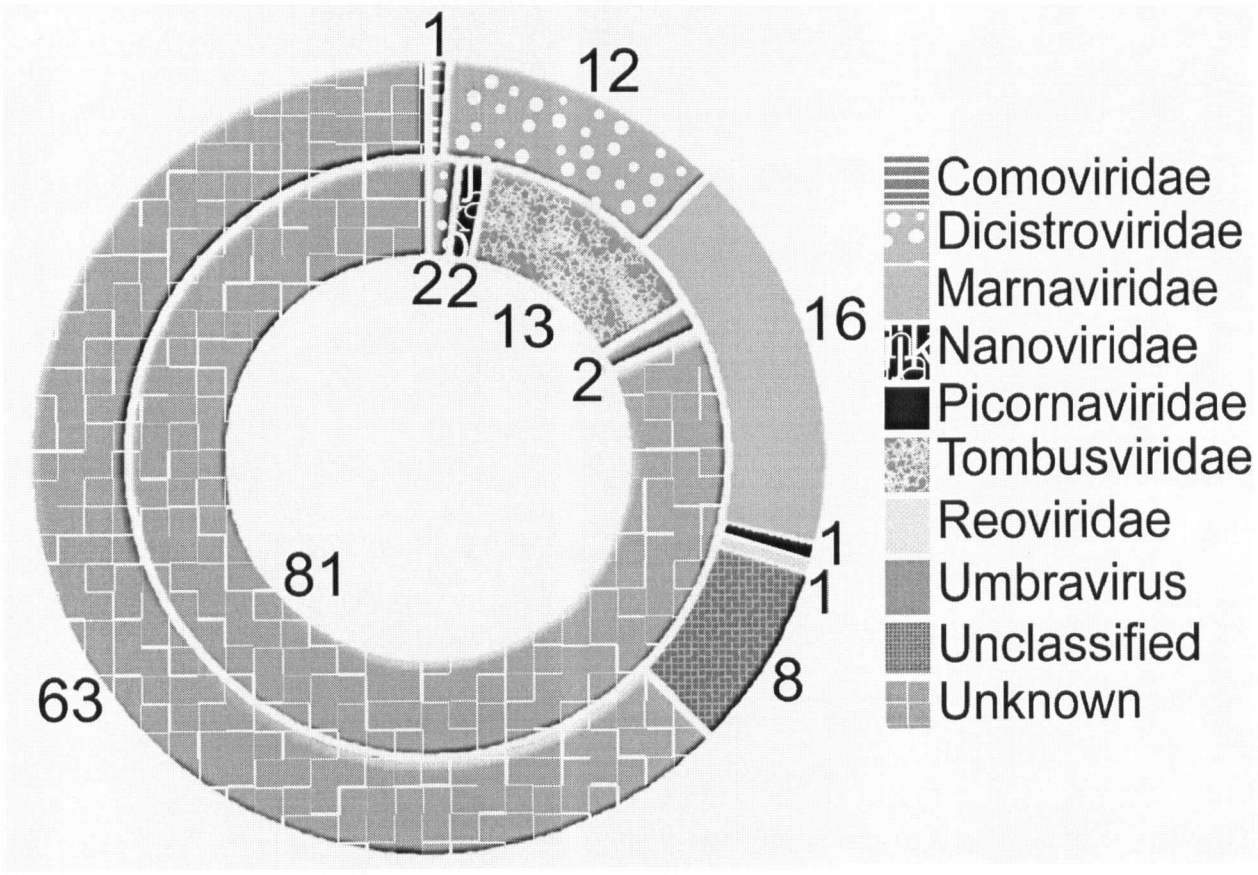

Figure 4.1 Composition of the JP (outer circle, n = 216) and the SOG (inner circle, n = 61) libraries

The top tBLASTx matches of sequences from JP and SOG with the GenBank non-redundant database (*E* value < 0.001) are categorized by taxonomic group. Virus families or genera are in different shades of grey. The *Comoviridae*, *Dicistroviridae*, *Marnaviridae*, and *Picornaviridae* are families in the proposed order *Picornavirales* (Christian et al. 2005). The percent values for each virus group in each library are shown. The identification of the individual viruses from each taxonomic group can be found in Table 4.2.

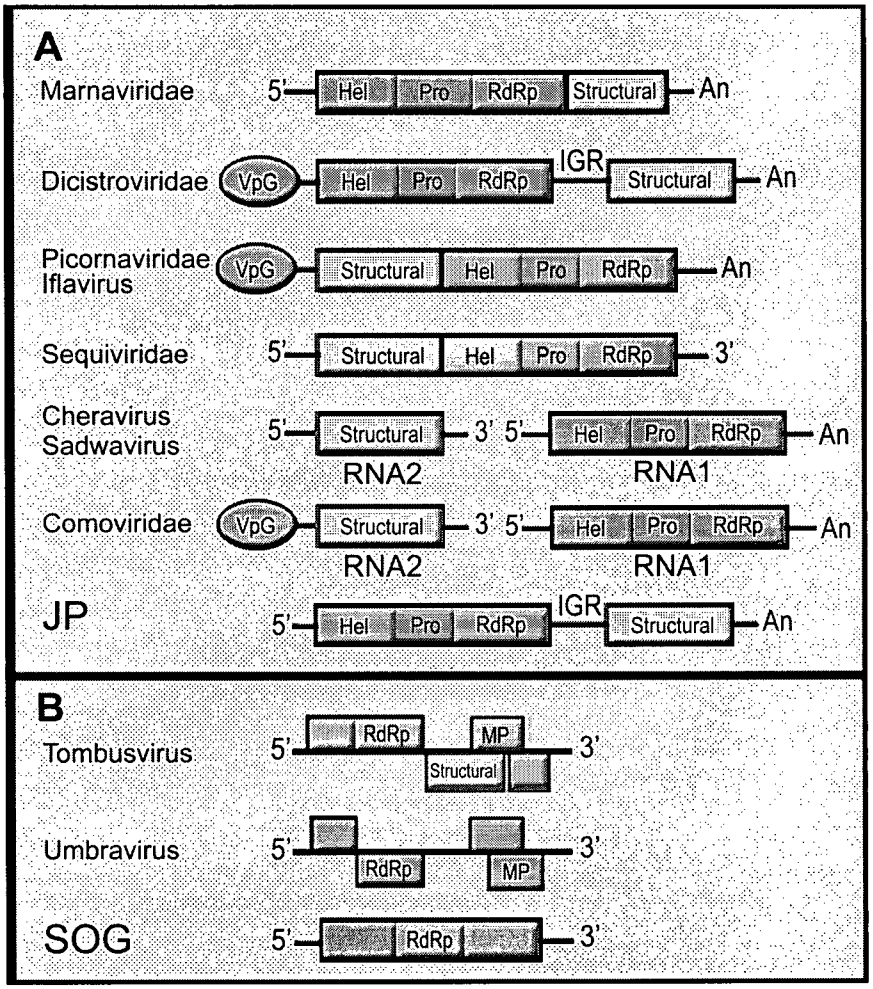

Figure 4.2 Comparison of the general genomic organization of the RNA virus genomes assembled from the JP and SOG libraries (JP and SOG) with representative viruses from the (A) proposed order *Picornavirales* (Christian et al. 2005) and the (B) family *Tombusviridae* and genus *Umbravirus*.

Genomes are shown from 5′ to 3′, where conserved RNA virus protein domains are labeled as Hel =for helicase, Pro, = protease, RdRp, = RNA-dependent RNA polymerase; IGR, = intergenic region, MP, = movement protein; and An, indicates the presence of a poly(A) tail. Figure 4.2-A genomes are approximately 10 kbp in size while Figure 4.2-B genomes are approximately 5 kbp in size. The "JP" schematic represents the genome organization of both assembled RNA viruses JP-A and JP-B. The characteristic read-through stop codon of the *Tombusviridae* replicase (represented by a divided RdRp) and the –1 frame shift of the *Umbravirus* replicase (represented by a staggered RdRp) are also shown (B). Unlabeled regions in gray refer to sequence that code for protein of unknown function.

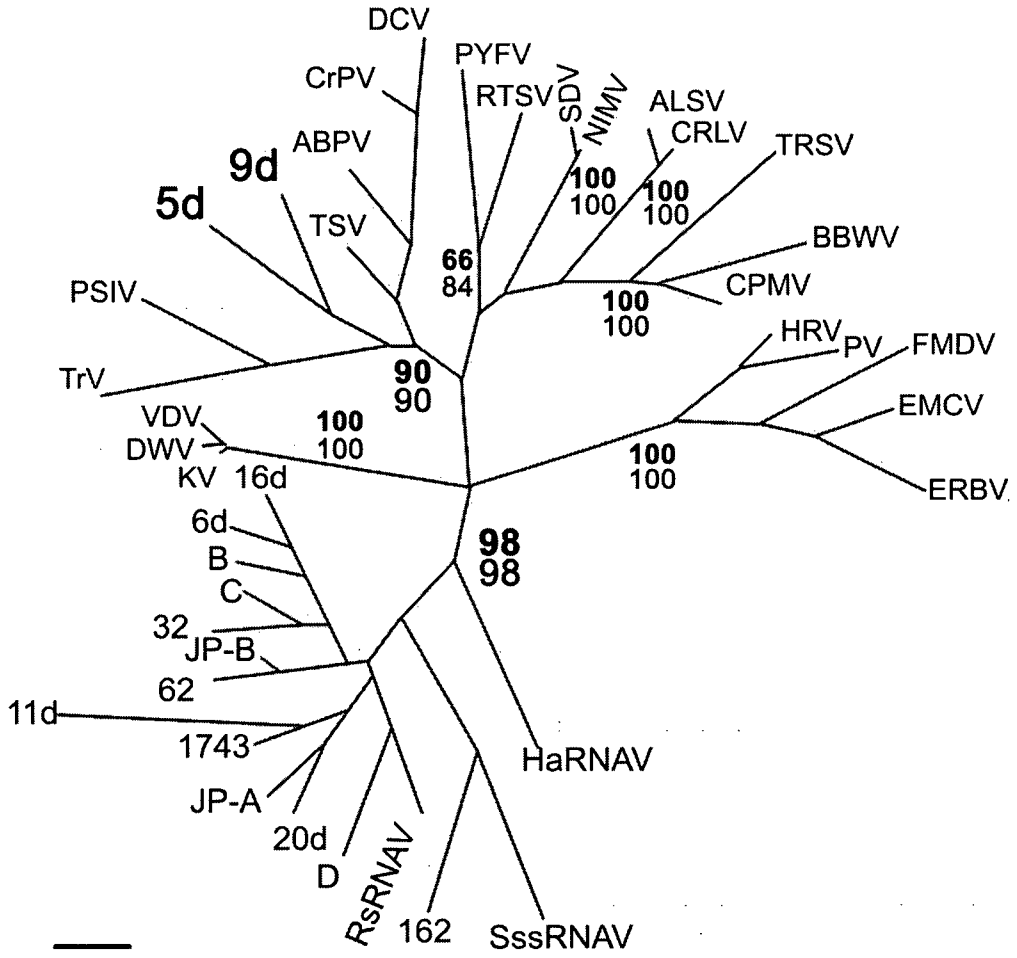

Figure 4.3 Bayesian maximum likelihood trees of aligned RdRp amino acid sequences from the JP RNA virus community and representative members of the proposed order *Picornavirales* (Christian et al. 2005)

Bayesian clade credibility values are shown for relevant nodes in boldface followed by bootstrap values based on neighbor-joining analysis. JP-A and JP-B are from the assembled environmental genomes. The Bayesian scale bar indicates a distance of 0.1. Environmental sequence numbers followed by a "d" are from excised denaturing gradient gel electrophoresis bands. See Table 4.4 for complete virus names, virus classification and sequence accession numbers.

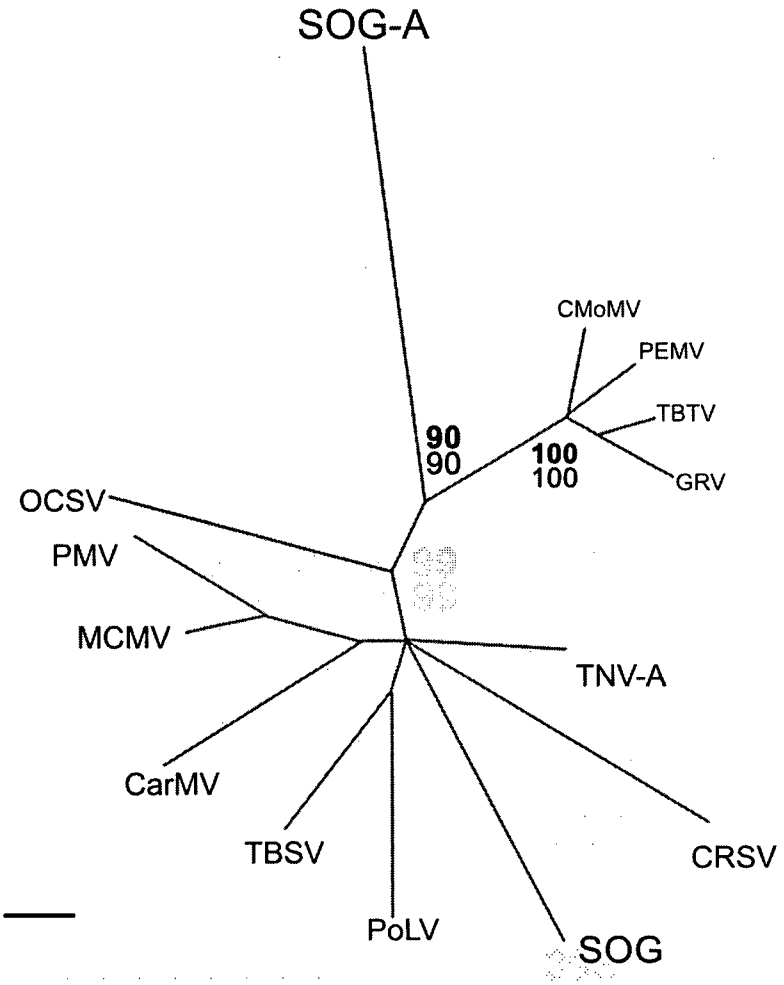

Figure 4.4 Bayesian maximum likelihood trees of aligned RdRp amino acid sequences from the SOG virus library and representative viruses from the *Tombusviridae* and *Umbravirus* genus

Bayesian clade credibility values are shown for relevant nodes in boldface followed by bootstrap values based on neighbor-joining analysis. SOG is from the assembled environmental genome. The Bayesian scale bar indicates a distance of 0.1. See Table 4.4 for complete virus names, virus classification and sequence accession numbers.

4.5 References

Altekar, G., S. Dwarkadas, J. P. Huelsenbeck, and F. Ronquist. 2004. Parallel metropolis coupled Markov chain Monte Carlo for bayesian phylogenetic inference. Bioinformatics **20:** 407-415.

Altschul, S. F., T. L. Madden, A. A. Schaffer, J. H. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research **25:** 3389-3402.

Angly, F., B. Rodriguez-Brito, D. Bangor, P. Mcnairnie, M. Breitbart, P. Salamon, B. Felts, J. Nulton, J. Mahaffy, and F. Rohwer. 2005. PHACCS, an online tool for estimating the structure and diversity of uncultured viral communities using metagenomic information. BMC Bioinformatics **6**.

Børsheim, K. Y. 1993. Native marine bacteriophages. FEMS Microbiology Letters **102:** 141-159.

Breitbart, M., B. Felts, S. Kelley, J. M. Mahaffy, J. Nulton, P. Salamon, and F. Rohwer. 2004. Diversity and population structure of a near-shore marine-sediment viral community. Proceedings of the Royal Society of London Series B-Biological Sciences **271:** 565-574.

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

Brussaard, C. P. D., A. A. M. Noordeloos, R. A. Sandaa, M. Heldal, and G. Bratbak. 2004. Discovery of a dsRNA virus infecting the marine photosynthetic protist *Micromonas pusilla*. Virology **319:** 280-291.

Christian, P., Fauquet, C.M., Gorbalenya, A.E., King, A.M.G., Knowles, N., Legall, O., Stanway, G. 2005. A proposed *Picornavirales* Order *In* C. M. Fauquet [ed.], Microbes in a changing world. International Unions of Microbiological Societies.

Culley, A. I., A. S. Lang, and C. A. Suttle. 2003. High diversity of unknown picorna-like viruses in the sea. Nature **424:** 1054-1057.

Delong, E. F., and D. M. Karl. 2005. Genomic perspectives in microbial oceanography. Nature **437:** 336-342.

Delong, E. F., C. M. Preston, T. Mincer, V. Rich, S. J. Hallam, N. U. Frigaard, A. Martinez, M. B. Sullivan, R. Edwards, B. R. Brito, S. W. Chisholm, and D. M. Karl. 2006. Community genomics among stratified microbial assemblages in the ocean's interior. Science **311:** 496-503.

Drake, J. W., and J. J. Holland. 1999. Mutation rates among RNA viruses. Proceedings of the National Academy of Sciences of the United States of America **96:** 13910-13913.

Edwards, R. A., and F. Rohwer. 2005. Viral metagenomics. Nature Reviews Microbiology **3:** 504-510.

Kim, D. H., H. K. Oh, J. I. Eou, H. J. Seo, S. K. Kim, M. J. Oh, S. W. Nam, and T. J. Choi. 2005. Complete nucleotide sequence of the hirame rhabdovirus, a pathogen of marine fish. Virus Research **107:** 1-9.

Koonin, E. V., and V. V. Dolja. 1993. Evolution and taxonomy of positive-strand RNA viruses - implications of comparative-analysis of amino-acid-sequences. Critical Reviews in Biochemistry and Molecular Biology **28:** 375-430.

Lang, A. S., A. I. Culley, and C. A. Suttle. 2004. Genome sequence and characterization of a virus (HaRNAV) related to picorna-like viruses that infects the marine toxic bloom-forming alga *Heterosigma akashiwo*. Virology **320:** 206-217.

Leblond, P. H. 1983. The Strait of Georgia - functional-anatomy of a coastal sea. Canadian Journal of Fisheries and Aquatic Sciences **40:** 1033-1063.

Lommel, S. A., Martelli, G.P., Rubino, L. Russo, M. 2004. Tombusviridae, p. 907-936. *In* C. M. Fauquet, Mayo M.A., Maniloff, J., Desselberger, U., Ball, L.A. [eds.], Virus Taxonomy. Eight Report of the International Committee on Taxonomy of Viruses. Elsevier.

Mari, J., B. T. Poulos, D. V. Lightner, and J. R. Bonami. 2002. Shrimp Taura syndrome virus: genomic characterization and similarity with members of the genus *Cricket paralysis-like viruses*. Journal of General Virology **83:** 915-926.

Nagasaki, K., Y. Shirai, Y. Takao, H. Mizumoto, K. Nishida, and Y. Tomaru. 2005. Comparison of genome sequences of single-stranded RNA viruses infecting the bivalve-killing dinoflagellate *Heterocapsa circularisquama*. Applied and Environmental Microbiology **71:** 8888-8894.

Nagasaki, K., Y. Tomaru, N. Katanozaka, Y. Shirai, K. Nishida, S. Itakura, and M. Yamaguchi. 2004. Isolation and characterization of a novel single-stranded RNA virus infecting the bloom-forming diatom *Rhizosolenia setigera*. Applied and Environmental Microbiology **70:** 704-711.

Nübel, U., F. Garciapichel, and G. Muyzer. 1997. PCR primers to amplify 16S rRNA genes from cyanobacteria. Applied and Environmental Microbiology **63:** 3327-3332.

Okayama, H., and P. Berg. 1982. High-efficiency cloning of full-length cDNA. Molecular and Cellular Biology **2:** 161-170.

Rima, B. K., A. M. J. Collin, and J. A. P. Earle. 2005. Completion of the sequence of a cetacean morbillivirus and comparative analysis of the complete genome sequences of four morbilliviruses. Virus Genes **30:** 113-119.

Rohwer, F., V. Seguritan, D. H. Choi, A. M. Segall, and F. Azam. 2001. Production of shotgun libraries using random amplification. Biotechniques **31:** 108-118.

Short, S. M., and C. A. Suttle. 2003. Temporal dynamics of natural communities of marine algal viruses and eukaryotes. Aquatic Microbial Ecology **32:** 107-119.

Suttle, C. A., A. M. Chan, and M. T. Cottrell. 1991. Use of ultrafiltration to isolate viruses from seawater which are pathogens of marine phytoplankton. Applied and Environmental Microbiology **57:** 721-726.

Swofford, D. 2000. PAUP*: Phylogenetic Analysis Using Parsimony and other Methods 4.0. Sinauer Associates, Inc.

Tai, V., J. E. Lawrence, A. S. Lang, A. M. Chan, A. I. Culley, and C. A. Suttle. 2003. Characterization of HaRNAV, a single-stranded RNA virus causing lysis of *Heterosigma akashiwo* (Raphidophyceae). Journal of Phycology **39:** 343-352.

Taliansky, M. E., Robinson, D.J., Waterhouse, P.M., Murant, A.F., De Zoeten, G.A., Falk, B.W., Gibbs, M.J. 2004. Umbravirus, p. 901-906. *In* C. M. Fauquet, Mayo M.A., Maniloff, J., Desselberger, U., Ball, L.A. [eds.], Virus Taxonomy. Eight Report of the International Committee on Taxonomy of Viruses. Elsevier.

Thompson, J. D., T. J. Gibson, F. Plewniak, F. Jeanmougin, and D.G. Higgins 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Research **25:** 4876-4882.

Venter, J. C., K. Remington, J. F. Heidelberg, A. L. Halpern, D. Rusch, J. A. Eisen, D. Y. Wu, I. Paulsen, K. E. Nelson, W. Nelson, D. E. Fouts, S. Levy, A. H. Knap, M. W. Lomas, K. Nealson, O. White, J. Peterson, J. Hoffman, R. Parsons, H. Baden-Tillson, C. Pfannkoch, Y. H. Rogers, and H. O. Smith. 2004. Environmental genome shotgun sequencing of the Sargasso Sea. Science **304:** 66-74.

Vetten, H. J., Chu, P.W.G., Dale, J.L., Harding, R., Hu, J., Katul, L., Kojima, M., Randles, J.W., Sano, Y., Thomas, J.E. 2004. Nanoviridae, p. 343-352. *In* C. M. Fauquet, Mayo M.A., Maniloff, J., Desselberger, U., Ball, L.A. [eds.], Virus Taxonomy. Eight Report of the International Committee on Taxonomy of Viruses. Elsevier.

Weinbauer, M. G. 2004. Ecology of prokaryotic viruses. FEMS Microbiology Reviews **28:** 127-181.

Yokoi, T., S. Yamashita, and T. Hibi. 2003. The nucleotide sequence and genome organization of *Sclerophthora macrospora* virus A. Virology **311:** 394-399.

Zhang, T., M. Breitbart, W. H. Lee, J. Q. Run, C. L. Wei, S. W. L. Soh, M. L. Hibberd, E. T. Liu, F. Rohwer, and Y. J. Ruan. 2006. RNA viral community in human feces: Prevalence of plant pathogenic viruses. PLoS Biology **4:** 108-118.

Chapter V. The complete genomes of three viruses assembled from shotgun libraries of marine RNA virus communities

A version of this chapter will be submitted for publication

Culley, A.I., A.S. Lang, and C.A. Suttle. The complete genomes of three viruses assembled from shotgun libraries of marine RNA virus communities.

5.1 Introduction

Based on a variety of evidence, marine viral communities have been assumed to consist entirely of dsDNA bacteriophages (Weinbauer 2004) (see Chapter I, section 1.1 for more details). However, RNA viruses of every classification have been isolated from the ocean, although the *in situ* RNA virus community remains largely uncharacterized. For example, the identities and hosts of the most abundant marine RNA viruses from any location are still unknown. Although there are several examples of RNA viruses that infect marine animals (e.g. Smith 2000), these organisms represent a very small portion of the organisms in the sea; therefore it is unlikely that viruses infecting these organisms make up a significant fraction of the natural RNA virioplankton. It is more likely that the dominant RNA viruses infect the diverse and abundant marine protists. For example, RNA viruses have recently been isolated that infect a diatom (Nagasaki et al. 2004), dinoflagellate (Tomaru et al. 2004), and marine fungoid protist (Takao et al. 2005).

Previous research investigated the diversity of marine picorna-like viruses, a "superfamily" of positive-sense single-stranded (ss)RNA viruses that have similar genome features and several conserved protein domains (Chapter III, Culley et al. 2003). Analysis of RNA dependent RNA polymerase (RdRp) sequences amplified from marine virus communities demonstrated that picorna-like viruses are present and persistent in a diversity of marine environments. Furthermore, phylogenetic analyses showed that none of the environmental sequences fell within established virus families (Chapter III, Culley et al. 2003).

In a recent study, reverse-transcribed whole-genome shotgun libraries were used to characterize two marine RNA virus communities (Chapter IV, Culley et al. 2006). Positive-sense ssRNA viruses that are distant relatives of known RNA viruses dominated the libraries. One RNA virus library (JP) was characterized by a diverse, monophyletic clade of picorna-like viruses, while the second library (SOG) was dominated by viruses related to members of the family *Tombusviridae* and genus *Umbravirus* (Chapter IV, Culley et al. 2006). Moreover, in both libraries, a high percentage of sequence fragments contributed to a handful of contiguous segments of sequence (contig). Specifically, in the SOG sample 59% of the sequence fragments that formed overlapping contigs fell into one segment. Similarly, 66% of JP sequence fragments contributed to only four contigs (Chapter IV, Culley et al. 2006). Using a PCR-based approach to

increase the amount of sequence for each dominant contig resulted in the assembly of three complete viral genomes. This contribution analyzes these three marine RNA virus genomes and investigates their similarities and differences of representative genotypes with established viral taxa.

5.2 Results and Discussion

5.2.1 Jericho Pier site

The JP-A genome is a single molecule of linear positive-sense ssRNA, 9212 nt in length. The genome has a 568 nt 5' untranslated region (UTR) followed by 2 predicted open reading frames (ORFs) of 5131 nt (ORF 1, nt position 569 to 5700) and 3044 nt (ORF 2, nt position 5841 to 8885) separated by an intergenic region (IGR) of 139 nt (Figure 5.1). ORF 2 is followed by a 3' UTR of 327 nt (nt position 8886 to 9213) and a poly(A) tail (Figure 5.1). The base composition of JP-A is 27.1% A, 19.4% C, 22.0% G, and 31.6% U; this results in a % G+C of 41%, a percentage similar to other polycistronic picorna-like viruses (Table 5.1).

Comparison of the protein sequence predicted to be encoded by ORF 1 of JP-A to known viral sequences shows that it contains conserved sequence motifs characteristic of a type III viral helicase (aa residues 641 to 756), a 3C-like cysteine protease (aa residues 1288 to 1314) and a type I RdRp (aa residues 1561 to 1802) (Koonin and Dolja 1993, Figure 5.1). BLASTp (Altschul et al. 1997) searches of the NCBI database with the ORF 1 sequence showed significant inferred amino-acid sequence similarities (*E* value < 0.001) to nonstructural protein motifs of several viruses, including members of the families *Dicistroviridae* (Drosophila C virus), *Marnaviridae* (HaRNAV), *Comoviridae* (Cowpea mosaic virus) and the unassigned genus Iflavirus (Kakugo virus). The top matches for ORF 1 were to RsRNAV [*E* value = $3e^{-119}$, identities = 302/908 (33%)], a newly sequenced, unclassified positive-sense ss RNA virus that infects the widely distributed diatom *Rhizosolenia setigera* (Nagasaki et al. 2004), HaRNAV [*E* value = $2e^{-32}$, identities = 156/624 (25%)] and Drosophila C virus [*E* value = $1e^{-29}$, identities = 148/603 (24%)], a positive-sense ssRNA virus that infects fruit flies. Comparison of the protein sequence predicted to be encoded by ORF 2 of JP-A to known viral sequences shows that it has significant similarities to the structural proteins of viruses from the families *Dicistroviridae* (Drosophila C virus), *Marnaviridae* (HaRNAV), and the genus Iflavirus (Varroa destructor virus 1). The

sequences that are most similar to ORF 2 of JP-A were the structural protein regions of RsRNAV [*E* value = $6e^{-78}$, identities = 212/632 (33%)], HaRNAV [*E* value = $6e^{-68}$, identities = 187/607 (30%)] and SssRNAV [*E* value = $2e^{-49}$, identities = 241/962 (25%)].

JP-B is also likely from a positive-sense ssRNA virus. The 8839 bp genome consists of a 5' UTR of 766 nt followed by two predicted ORFs of 4850 nt (ORF 1, nt position 767 to 5617) and 2786 nt (ORF 2, nt position 5843 to 8629) separated by an IGR of 224 nt (nt position 5618 to 5842, Figure 5.1). The 3' UTR is 209 followed by a poly(A) tail. The base composition of the genome was A, 30.8%; C, 17.9%; G, 19.7%; U, 31.6%. Like JP-A, this is % G+C value of 38% is comparable to the % G+C observed in other dicistronic picorna-like viruses (Table 5.1)

The position of core sequence motifs conserved among positive-sense ssRNA viruses and BLAST searches of the NCBI database with the translated JP-B genome suggest that nonstructural proteins are encoded by ORF 1, and the structural proteins are encoded by ORF 2. We identified conserved sequence motifs in ORF 1 characteristic of a type III viral helicase (aa residues 587 to 700), a 3C-like cysteine protease (aa residues 1141 to 1168) and a type I RdRp (aa residues 1402 to 1667) (Koonin and Dolja 1993) (Figure 5.1). BLASTp (Altschul et al. 1997) searches of the NCBI database showed that ORF 1 has significant similarities (*E* value < 0.001) to nonstructural genes from positive-sense ssRNA viruses from a variety of families, including the *Comoviridae* (Peach rosette mosaic virus), *Dicistroviridae* (Taura syndrome virus), *Marnaviridae* (HaRNAV), *Sequiviridae* (Rice tungro spherical virus) and *Picornaviridae* (Avian encephalomyelitis virus). The top scoring sequences [*E* value = $2e^{-69}$, identities = 232/854 (27%)] were to a RdRp sequence fragment from RsRNAV and a partial picorna-like virus RdRp from an unidentified virus [*E* value = $2e^{-40}$, identities = 85/150 (56%)] amplified from the same JP station during an earlier study (Chapter III, Culley et al. 2003). Significant similarities to ORF 2 include the structural genes of viruses from the families *Dicistroviridae* (*Rhopalosiphum padi* virus) *Marnaviridae* (HaRNAV) and *Picornaviridae* (Human parechovirus 2), as well as the unclassified genus Iflavirus (Ectropis obliqua picorna-like virus). The top scoring sequences were to the capsid protein precursor regions of RsRNAV [*E* value = $9e^{-88}$, identities = 244/799 (30%)] and HaRNAV [*E* value = $8e^{-60}$, identities = 180/736 (24%)] and SssRNAV [*E* value = $1e^{-40}$. identities = 156/588 (26%)].

Several viruses in the family *Dicistroviridae* have genomes that contain internal ribosome

entry sites (IRESs) (Jan and Sarnow 2002, Nishiyama et al. 2003, Cevallos and Sarnow 2005, Czibener et al. 2005), which raises the question of whether an IRES was present in JP-A or JP-B given their apparently similar dicistronic genome organization. Structures within the IRES position the genome on the ribosome actuating elongation even in the absence of known canonical initiation factors (Jan and Sarnow 2002). For example, TSV, a marine dicistrovirus, has an IRES located in the IGR that directs the synthesis of the structural proteins (Cevallos & Sarnow 2005). Although secondary structure elements characteristic of dicistrovirus IGR-IRESs in the JP genomes (Hatakeyama et al. 2006) were not located in the JP genomes, both genome sequences have extensive predicted secondary structure in the 5' UTRs and IGRs, suggestive of IRES function. Moreover, start codons in a favorable Kozak context, [i.e. conserved sequences upstream of the start codon that are thought to play a role in initiation of translation (Kozak 1986)] were not found in the JP genomes. However, IRES structures can vary greatly between viruses and there is clearly large evolutionary distance among these viruses (see below), and therefore predicted IRES elements must be confirmed experimentally in dicistronic constructs. It seems likely that these viruses use similar mechanisms to initiate translation of the ORF 2 genes.

We used RT-PCR to assess the distribution and persistence of the JP-A and JP-B viruses *in situ*. Amplification with specific primers that target each of these viruses occurred in samples from throughout the Strait of Georgia, the West coast of Vancouver Island and in every season and tidal state at Jericho pier (Figure 5.2, Table 5.2). These results suggest that JP-A and JP-B are ubiquitous and can be detected in marine and estuarine waters.

It has long been recognized that several other groups of small, positive-sense ssRNA viruses share many characteristics with viruses in the family *Picornaviridae*. Recently, Christian et al. (2005) proposed creating an order (the *Picornavirales*) of virus families (*Picornaviridae*, *Dicistroviridae*, *Marnaviridae*, *Sequiviridae* and *Comoviridae*) and unassigned genera (*Iflavirus*, *Cheravirus,* and *Sadwavirus*) that have picornavirus-like characteristics. Viruses in the proposed order have genomes with a covalently attached protein to the 5' end and a 3' poly(A) tail, a conserved order of non-structural proteins (helicase-VpG-proteinase-RdRp), regions of high sequence similarity in the helicase, proteinase and RdRp, post translational protein processing during replication and an icosahedral capsid with a unique "pseudo-T3" symmetry and only infect eukaryotes.

Although the capsid morphology, presence of a 5' terminal protein and replication strategy is unknown, signature genomic features and phylogenetic analysis suggest that the JP viruses fall within the proposed order *Picornavirales*. Both JP genomes encode the conserved core aa motifs and have the non-structural gene order characteristic of viruses in the Picornavirales. Furthermore, both JP genomes have a poly(A) tail and G+C content commensurate with viruses in the *Picornavirales*. Bayesian trees (Altekar et al. 2004) based on alignments of conserved RdRp domains (Koonin and Dolja 1993) (Figure 5.3) as well as concatenated (putative) Hel/RdRp/VP3 capsid-like protein sequences (Figure 5.4) of the JP genomes and representative members of the *Picornavirales*, resolves established taxa within the *Picornavirales* and provides strong support for a clade comprised of viruses (HaRNAV, RsRNAV and SssRNAV) that infect marine protists. Within this clade, RsRNAV, JP-A and JP-B have the most characteristics in common. For example, they have the same order of structural and non-structural genes, they are polycistronic and phylogenetic analyses indicate they are (relatively) closely related. Whether JP-A and JP-B infect host organisms related to *Rhizosolenia setigera* remains unclear, although the inclusion of the JP genomes within this clade and the fact that protists are the most abundant eukaryotes in the sea suggest that both JP viruses likely have a protist host.

5.2.2 Strait of Georgia site

The SOG genome has features characteristic of a positive-sense ssRNA virus. The genome is 4449 bp long and comprised of a 5' UTR of 211 bp followed by three putative ORFs (nt position 212-1229, nt position 1232 – 2863 and nt position 2864 – 4231) and is terminated with a 3' UTR of 218 bp. A poly(A) tail was not detected. Another putative ORF located at nt position 49 to 786 is in an alternative reading frame relative to the ORFs discussed above (Figure 5.1). The G+C content of the SOG genome is 52%.

We identified only the eight conserved motifs of the RdRp (Koonin and Dolja 1993) in the SOG genome (aa residues pos 563 to 817, nt positions 1687- 2451) (Figure 5.1). tBLASTx (Altschul et al. 1997) searches with the remainder of the genome sequence showed no significant matches (E value < 0.001) to sequences in the NCBI database (including the five environmental metagenomes currently deposited). BLASTp searches with the putative RdRp resulted in significant similarities (E value < 0.001) to RdRp sequences from positive-sense ssRNA viruses

from the family *Tombusviridae* and the unassigned genus *Umbravirus*. The sequence with the most similarity to SOG was from Olive latent virus 1 [*E* value = $3e^{-66}$, identities = 180/508 (35%)]. This virus belongs to the genus *Necrovirus* in the family *Tombusviridae* that has a host range restricted to higher plants (Lommel et al. 2004). SOG is also significantly similar to the Carrot mottle mimic virus sequence [*E* value = $6e^{-66}$, identities = 178/492 (36%)], a member of the unclassified genus *Umbravirus* whose known members infect only flowering plants.

Although the SOG putative RdRp sequence has similarity to the RdRp of viruses from the family *Tombusviridae* and genus *Umbravirus*, the remaining SOG sequence has no detectable similarity to any other known sequence. A Bayesian maximum likelihood tree based on alignments of the SOG RdRp with the available *Umbravirus* sequences and representative members of the *Tombusviridae* indicates that the SOG genome forms a well supported clade (Bayesian clade support value of 100) with the single member of the genus *Avenavirus*, OCSV (Figure 5.5). Additionally, the presence of an amber stop codon (nt position 1230-1232) separating ORF 1 and 2 of the SOG genome (Figure 5.1) resembles the in-frame termination codon characteristic of the replicase of viruses in 7 of the 8 genera of the *Tombusviridae* (White & Nagy 2004). This division of the replicase of the *Tombusviridae* by a termination codon is thought to be part of a translational read though gene expression strategy (White & Nagy 2004). Other similarities to the *Tombusviridae* include the absence of an obvious helicase motif and the 5' proximal relative position of the RdRp (Lommel et al. 2004) within the gene. However, unlike viruses in the *Tombusviridae*, there is no recognizable sequence for conserved movement or capsid proteins in the SOG genome. The absence of a movement protein suggests that the SOG virus does not infect a higher plant. Our inability to identify structural genes may indicate that, like the umbraviruses, the SOG genome does not encode capsid proteins. However it is more likely that the structural proteins of the SOG genome have no sequence similarity to those currently in the NCBI database.

Our analyses suggest that a persistent and possibly dominant population of novel dicistronic picorna-like viruses is an important component of the RNA virioplankton in coastal waters. Nevertheless, as exemplified in the SOG genome, other marine RNA virus assemblages appear to contain viruses whose detectable sequence similarity with established groups of viruses is limited to only the most conserved RdRp genes. As we work towards the ultimate goal of understanding the ecological role of marine viruses, the next challenge with these data, and in

marine virus metagenomic research in general, is to affiliate each assembled genome with a specific virion and host.

5.3 Materials and Methods

5.3.1 Station description

The shotgun libraries were constructed from seawater samples collected from two stations, JP (Jericho Pier) a site in English Bay adjacent to the city of Vancouver, British Columbia and SOG (Strait of Georgia), located in the central Strait of Georgia next to Powell River, B.C. (Figure 5.2, Chapter IV section 4.3.1 and Table 4.1).

The locations of the stations where one or both of the JP genomes were detected are shown in Figure 5.2. Details for each station are listed in Table 5.2 and Table 4.1. In summary, samples were collected from sites throughout the Strait of Georgia, including repeated sampling from the JP site during different seasons, and from the West coast of Vancouver Island in Barkley Sound.

5.3.2 Virus concentration method

Concentrated virus communities were produced as described by Suttle et al. (1991). Twenty to 60 liters of seawater from each station were pre-filtered through glass fiber (nominal pore size 1.2 μm) and 0.45 μm pore-size Durapore PVDF (polyvinylidene fluoride) membranes (Millipore, Cambridge, Canada) respectively, to remove particulates, including eukaryotic plankton and most prokaryotes. This filtrate was subsequently concentrated approximately 200 fold through a tangential flow filter cartridge (Millipore) with a 30 kDa molecular cutoff, essentially resulting in the concentration of the 2 to 450 nm size fraction of seawater. Remaining bacteria were removed by filtering each viral concentrate two times through a 0.22 μm Durapore PVDF membrane (Millipore). Virus-sized particles in each virus concentrate were pelleted via ultracentrifugation (5 h at 113 000 × g at 4 °C). Pellets were resuspended in sterile buffer (50 μm Tris chloride) and left to resuspend overnight at 4 °C.

5.3.3 Whole-genome shotgun library construction

A detailed description of the whole genome shotgun library construction protocol can be found in Culley et al. (2006, Chapter IV). Briefly, before extraction concentrated viral lysates

were treated with RNase (Roche, Mississauga, Canada) and then extracted with a QIAamp Minelute Virus Spin Kit (Qiagen, Mississauga, Canada) according to the manufacturer's instructions. An aliquot of each extract was used in a PCR reaction with universal 16S primers to ensure samples were free of bacteria. To isolate the RNA fraction samples were treated with DNase 1 (Invitrogen, Burlington, Canada) and used as templates for reverse transcription with random hexamer primers. Double-stranded cDNA fragments were synthesized from ssDNA with Superscript III reverse transcriptase (Invitrogen) using nick translational replacement of genomic RNA (Okayama and Berg 1982). After degradation of overhanging ends with T4 DNA polymerase (Invitrogen), adapters were attached to the blunted products with T4 DNA ligase (Invitrogen). Subsequently, excess reagents were removed and cDNA products were separated by size with a Sephacryl column (Invitrogen). To increase the amount of product for cloning, size fractions greater than 600 bp were amplified with primers targeting the adapters. Products from each PCR reaction were purified and cloned with the TOPO TA Cloning system (Invitrogen). Clones were screened for inserts by PCR with vector-specific primers. Insert PCR products greater than 600 bp were purified and sequenced at the University of British Columbia's Nucleic Acid and Protein Service Facility (Vancouver, Canada). Sequence fragments were assembled into overlapping segments using Sequencher v 4.5 (Gene Codes, Ann Arbor, U.S.A.) based on a minimum match percentage of 98 and a minimum bp overlap of 20. Sequences were compared against the NCBI database with tBLASTx (Altschul et al. 1997). A sequence was considered significantly similar if BLAST *E* values were < 0.001. The details for viruses used in phylogenetic analyses are listed in Table 5.3. Virus protein sequences were aligned using CLUSTAL X v 1.83 with the Gonnet series protein matrix (Thompson et al. 1997). Alignments were transformed into likelihood distances with Mr Bayes v3.1.1 (Altekar et al. 2004) and 250000 generations. Neighbor-joining trees were constructed with PAUP v4.0 (Swofford 2002), and bootstrap values calculated based on percentages of 10000 replicates.

5.3.4 5' and 3' RACE

The 5' and 3' ends of the environmental viral genomes were cloned using the 5' and 3' RACE systems (Invitrogen) according to manufacturer's instructions. 3' RACE with the SOG genome required the addition of a poly(A) tract with poly (A) polymerase (Invitrogen) before cDNA synthesis. cDNA was synthesized directly from extracted viral RNA from the appropriate library. Three clones of each 5' and 3' end were sequenced.

5.3.5 PCR

5.3.5.1 Closing gaps in the assembly

PCR with primers targeting specific regions of the two JP environmental genomes were used to verify the genome assembly, increase sequencing coverage and reconfirm the presence of notable genome features. The template for these reactions was the amplified and purified PCR product from the JP and SOG shotgun libraries. Table 5.4 lists the sequence and genome position of primers used. The standard PCR conditions were reactions with 1 U of Platinum *Taq* DNA polymerase (Invitrogen) in 1× Platinum *Taq* buffer, 1.5 mM $MgCl_2$, 0.2 mM of each dNTP, and 0.2 μM of each primer (Table 5.4), in a final volume of 50 μl. Thermocycler conditions were, activation of the enzyme at 94 °C for 1 min 15 s, followed by 30 cycles of denaturation at 94 °C for 45 s, annealing at 50 °C for 45 s and extension at 72 °C for 1 minute. The reaction was terminated after a final extension stage of 5 min at 72 °C. PCR products were purified with a PCR Miniature cleanup kit (Qiagen) and sequenced directly with both primers.

5.3.5.2 Environmental screening

To assess the temporal and geographic distribution of the JP genomes, extracted RNA from viral concentrates were screened with Superscript III One-step RT-PCR System with Platinum *Taq* DNA Polymerase (Invitrogen) with primers JP-A 5 and 6 and JP-B 6 and 7 (Table 5.4). The template for the reactions was DNase 1 treated viral RNA extracted with a QIAamp Minelute Virus Spin Kit (Qiagen) according to the manufacturer's instructions. Each reaction consisted of RNA template, 1× reaction mix, 0.2 μM of each primer, 1 μl RT/Platinum *Taq* mix in a volume of 50 μl. Reactions were incubated 30 min at 50 °C, then immediately heated to 94 °C for 45 s, followed by 35 cycles of denaturation at 94 °C for 15 s, annealing at 50 °C for 30 s and extension at 68 °C for 1 min. After a final extension step at 68 °C for 5 min, RT-PCR products were analyzed by agarose gel electrophoresis. Products were sequenced to verify the correct target had been amplified.

5.4 Tables and Figures

Table 5.1 Comparison of base composition between polycistronic picorna-like viruses

Genome	A	C	G	U	% G+C
JP-A	**27.1**	**19.4**	**22.0**	**31.6**	**41**
JP-B	**30.8**	**17.9**	**19.7**	**31.6**	**38**
ABPV	35.7	15.4	20.1	28.9	36
ALPV	31.3	19.4	19.2	30.2	39
BQCV	29.2	18.5	21.6	30.6	40
CrPV	32.61	18.4	20.9	28.1	39
DCV	29.9	16.3	20.4	33.4	37
HiPV	29.2	18.7	20.9	31.2	39
KBV	33.8	17.5	20.2	28.6	38
PSIV	31.3	17.0	19.4	32.3	36
RhPV	30.0	18.6	20.2	31.2	39
RsRNAV	31.2	16.7	19.5	32.5	36
SINV-1	32.9	18.3	20.5	28.2	39
SssRNAV	24.2	26.1	23.6	26.0	50
TSV	28.0	20.2	23.0	28.8	43
TrV	28.7	16.1	19.8	35.4	36
Average	30.4	18.4	20.7	30.5	39

Table 5.2 JP genome survey sample sites. A "+" indicates amplification and "-" indicates no amplification occurred. "n.a." indicates the data is not available and "S" means the sample was taken from the surface.

Station Name	Station location (B.C., Canada)	Date (mm/dd/yy)	Location (Lat.,Long.)	Depth (m)	Temp (°C)	Salinity (ppt)	**JP-A** PCR	**JP-B** PCR
JP	Jericho Pier	04/28/00	49.27, -123.20	S	9	26	+	+
JP	Jericho Pier	06/15/00	49.27, -123.20	S	14	12	+	+
JP	Jericho Pier	06/29/00	49.27, -123.20	S	17	12	+	+
JP	Jericho Pier	07/06/00	49.27, -123.20	S	16	13	+	+
JP	Jericho Pier	07/13/00	49.27, -123.20	S	18	8	-	-
JP	Jericho Pier	07/27/00	49.27, -123.20	S	18	11	+	+
JP	Jericho Pier	08/17/00	49.27, -123.20	S	18	18	+	+
JP	Jericho Pier	09/14/00	49.27, -123.20	S	15	19	+	+
JP	Jericho Pier	09/21/00	49.27, -123.20	S	15	16	-	+
JP	Jericho Pier	09/28/00	49.27, -123.20	S	14	21	+	+
JP	Jericho Pier	11/23/00	49.27, -123.20	S	8	27	+	+
JP	Jericho Pier	02/15/01	49.27, -123.20	S	7	27	+	+
JP	Jericho Pier	06/14/01	49.27, -123.20	S	15	13	+	+
SEC	Sechelt Inlet	07/06/03	49.69, -123.84	4	13	26	-	+
TEA	Teakearne Inlet	07/07/03	50.19, -124.85	5	13	28	+	-
QUA	Quadra Island	07/07/03	50.19, -125.14	3	13	28	-	-
ARR	Arrow Pass	07/09/03	50.72, -126.67	2	10	31	+	+
IEC	Imperial Eagle Channel	06/20/99	48.87, -125.21	7	n.a.	n.a.	+	-
TRE	Trevor Channel	06/28/99	48.97, -125.16	S	n.a.	n.a.	+	+
BAM	Bamfield Inlet	07/06/99	48.81, -125.16	S	n.a.	n.a.	+	+
NUM	Numukamis Bay	07/12/99	48.90, -125.01	8	n.a.	n.a.	+	+

Table 5.3 Virus sequence details

Virus Group	Virus Acronym	Full Name	NCBI Accession #
Cheravirus	ALSV	Apple latent spherical virus	NC 003787
	CRLV	Cherry rasp leaf virus	NC 006271
Comoviridae	BBWV1	Broad bean wilt virus 1	NC 005289
	CPMV	Cowpea mosaic virus	NC 003549
	TRSV	Tobacco ringspot virus	NC 005097
Dicistroviridae	ABPV	Acute bee paralysis virus	NC 002548
	BQCV	Black queen cell virus	NC 003784
	CrPV	Cricket paralysis virus	NC 003924
	DCV	Drosophila C virus	NC 001834
	HiPV	Himetobi P virus	NC 003782
	PSIV	Plautia stali intestine virus	NC 003779
	TSV	Taura syndrome virus	NC 003005
	TrV	Triatoma virus	NC 003783
Iflavirus	DWV	Deformed wing virus	NC 004830
	IFV	Infectious flacherie virus	NC 003781
	KV	Kakugo virus	NC 005876
	PnPV	*Perina nuda* picorna-like virus	NC 003113
	SbV	Sacbrood virus	NC 002066
	VDV	Varroa destructor virus	NC 006494
Marnaviridae	HaRNAV	*Heterosigma akashiwo* RNA virus	NC 005281
Picornaviridae	AiV	Aichi virus	NC 001918
	EMCV	Encephalomyocarditis virus	NC 001479
	ERBV	Equine rhinitis B virus 1	NC 003983
	FMDV	Foot-and-mouth disease virus A	NC 011450
	HAV	Hepatitis A virus	NC 001489
	HPeV	Human parechovirus	NC 001897
	HRV	Human rhinovirus 14	NC 001490
	PV	Poliovirus	NC 002058
	PTV	Porcine teschovirus 1	NC 003985

Virus Group	Virus Acronym	Full Name	NCBI Accession #
Sadwavirus	SDV	Satsuma dwarf virus	NC 003785
	NIMV	Navel orange infectious mottling virus	AB022887
Sequiviridae	MCDV	Maize chlorotic dwarf virus	NC 003626
	PYFV	Parsnip yellow fleck virus	NC 003628
	RTSV	Rice tungro spherical virus	NC 001632
Tombusviridae	CaRMV	Carnation mottle virus	NC 001265
	CRSV	Carnation ringspot virus	NC 003530
	MCMV	Maize chlorotic mottle virus	NC 003627
	OCSV	Oat chlorotic stunt virus	NC 003633
	PMV	Panicum mosaic virus	NC 002598
	PoLV	Pothos latent virus	NC 000939
	TBSV	Tomato bushy stunt virus	NC 001554
	TNV-A	Tobacco necrosis virus A	NC 001777
Umbravirus	CMoMV	Carrot mottle mimic virus	NC 001726
	GRV	Groundnut rosette virus	NC 003603
	PEMV-2	Pea enation mosaic virus-2	NC 003853
	TBTV	Tobacco bushy top virus	NC 004366
Unclassified	RsRNAV	*Rhizosolenia setigera* RNA virus	AB243297
	SssRNAV	*Schizochytrium* single-stranded RNA virus	NC 007522

Table 5.4 PCR primers used to complete the three genome sequences. Primers JP-A 5 and 6 and JP-B 6 and 7 (shown in bold) were used in the environmental survey.

Genome	Primer	Sequence (5'-3')	Location (bp)	Strand primer is based on
JP-A	1	TTATTGCTAAGGCTGAAAGTCT	2596-2617	+
	2	ATCCATTTTCTACCAACTTCAC	3467-3484	-
	3	TCGTCGGGAAGATGGC	3764-3779	+
	4	GAAGCCTGCCACATCAAT	4285-4300	-
	5	**ATGGTGGCAGTATGGTCG**	**5552-5569**	+
	6	**CACTGGTATTCTTTGATTTTGAT**	**6165-6185**	-
	7	TTGTGGATGATTCTGAACTTG	6881-6901	+
	8	AAAATCGTCTCCAGCAGC	7863-7878	-
	9	TTGCTCCTTATGCTCCTCA	7943-7961	+
	10	GAAGGTTCTGGTGTTTATTTGTA	8881-8901	-
JP-B	1	CAATCATACCCCTGAGTTTAGA	213-234	+
	2	AGTCTCAACAACACCCAAGC	1058-1077	-
	3	CCCGATTTTCTGTATGTTTTAG	1397-1418	+
	4	ACCAACGACCAACTTAGCC	2076-2094	-
	5	GCGAAATGAAAAGGAGAAG	2646-2664	+
	6	**CGCTCTCGGACATAACAAA**	**3150-3168**	-
	7	**CCGTTTTCCGTTACATTGA**	**3666-3684**	+
	8	TTTTACCAACCTTAGCCTTCT	4240-4260	-
	9	GCTTCTTACTAAATCAATCCTTCTA	5521-5545	+
	10	GCTAAAGTACAACCATAGAAAAATG	6416-6440	-
SOG	1	ATACTTCTTCCCGCATCAG	378-398	+
	2	TCCTTGGAATCGCTTGTTGT	771-790	-
	3	CGTCGGGTCGTCTAAAAC	1021-1040	+
	4	CAGGCTTCTGAGGTGTGG	1464-1481	-
	5	GACTCCAACACAACAAATCG	2716-2737	+
	6	GAGACAGGACAAGCGTTATG	3160-3179	-

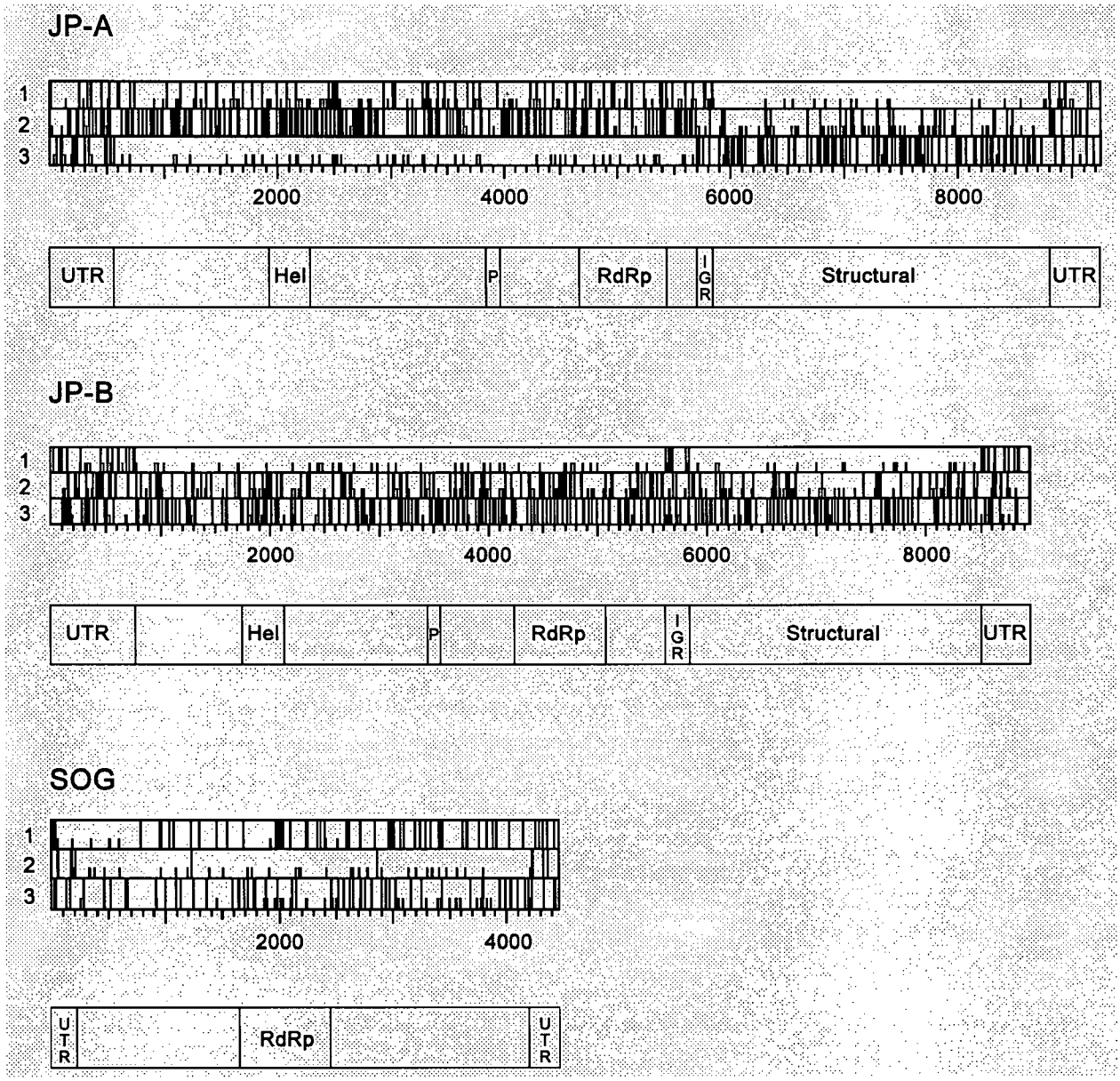

Figure 5.1 Analysis of genomes for possible open reading frames.

In the ORF maps created with DNA Strider (Marck 1988), for each reading frame, potential start codons (AUG) are shown with a half-height line and stop codons (UGA, UAA, and UAG) are shown by full-height lines. Putative genes (Hel = helicase, Pro = protease, RdRp = RNA-dependent RNA polymerase) and genomic features (UTR = untranslated region, IGR = intergenic region) are noted below each genome. See text for more detail

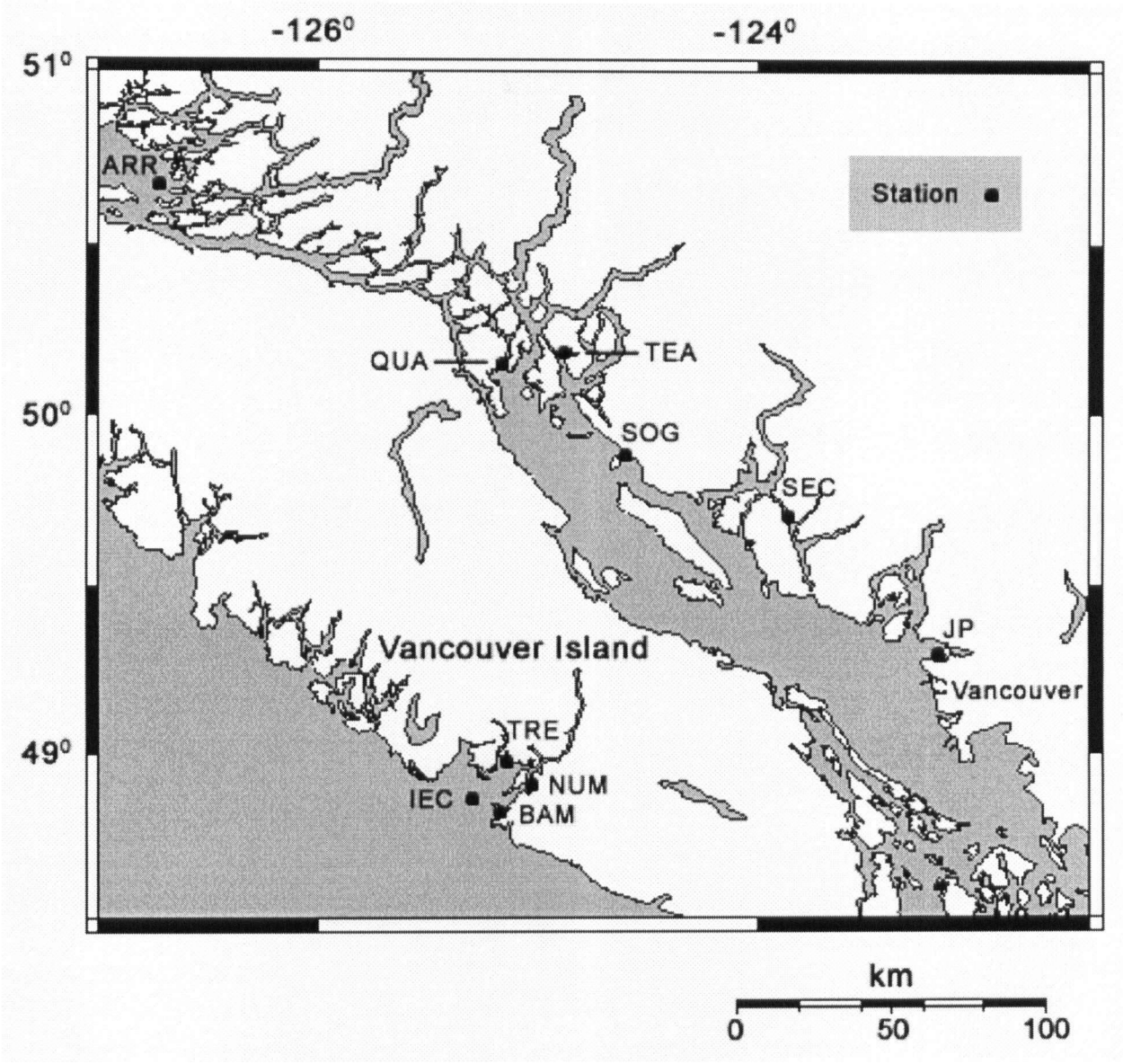

Figure 5.2 Map of the Strait of Georgia, British Columbia, Canada with station locations

JP-A and JP-B were detected at 5/9 stations. The SOG station was not assayed for JP-A or JP-B. See Table 5.2 for additional information.

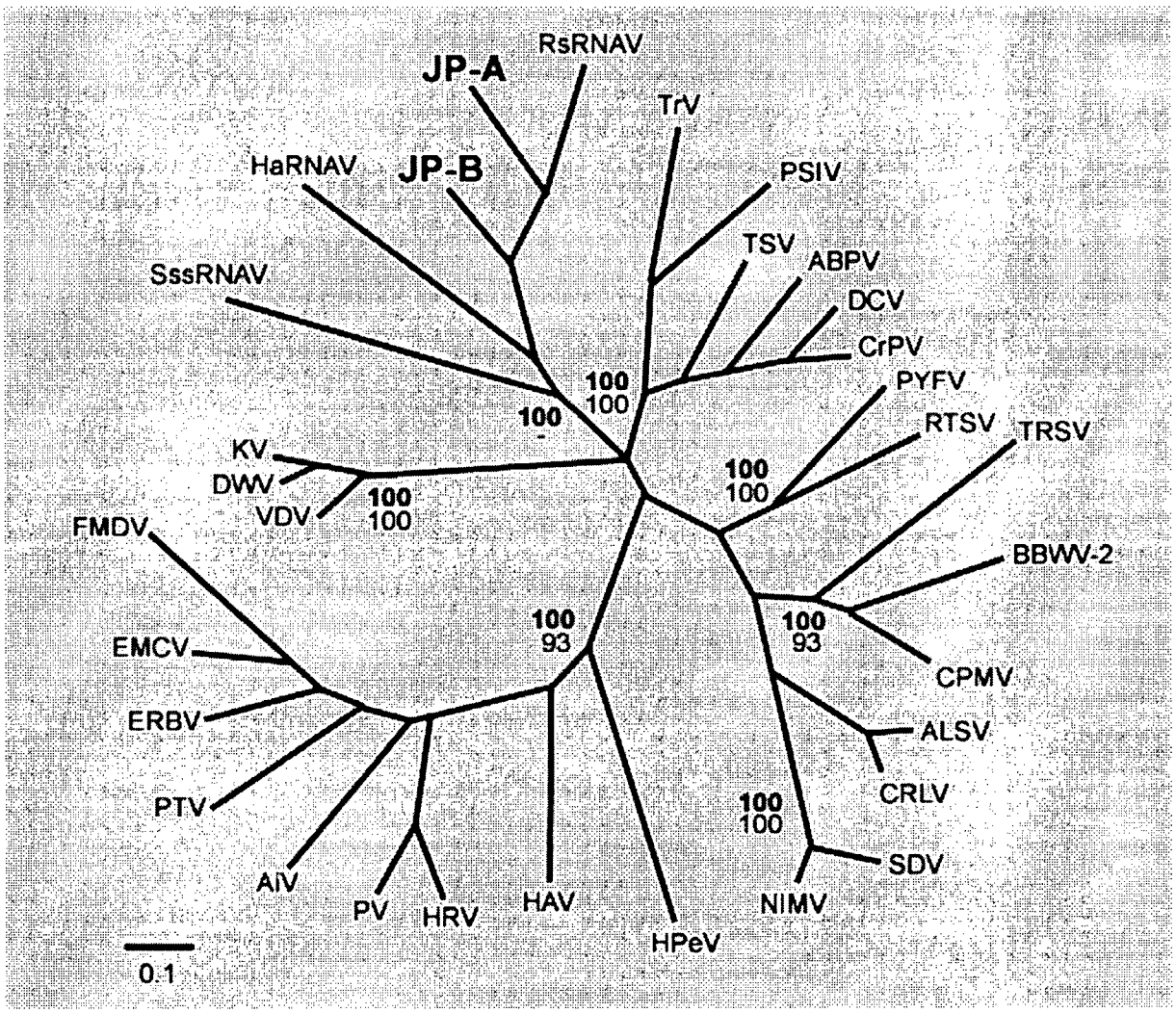

Figure 5.3 Bayesian maximum likelihood trees of aligned RdRp amino acid sequences from JP-A and JP-B and representative members of the proposed order *Picornavirales*.

Bayesian clade credibility values are shown for relevant nodes in boldface followed by bootstrap values based on neighbor-joining analysis. The Bayesian scale bar indicates a distance of 0.1. See Table 5.3 for complete virus names and accession numbers.

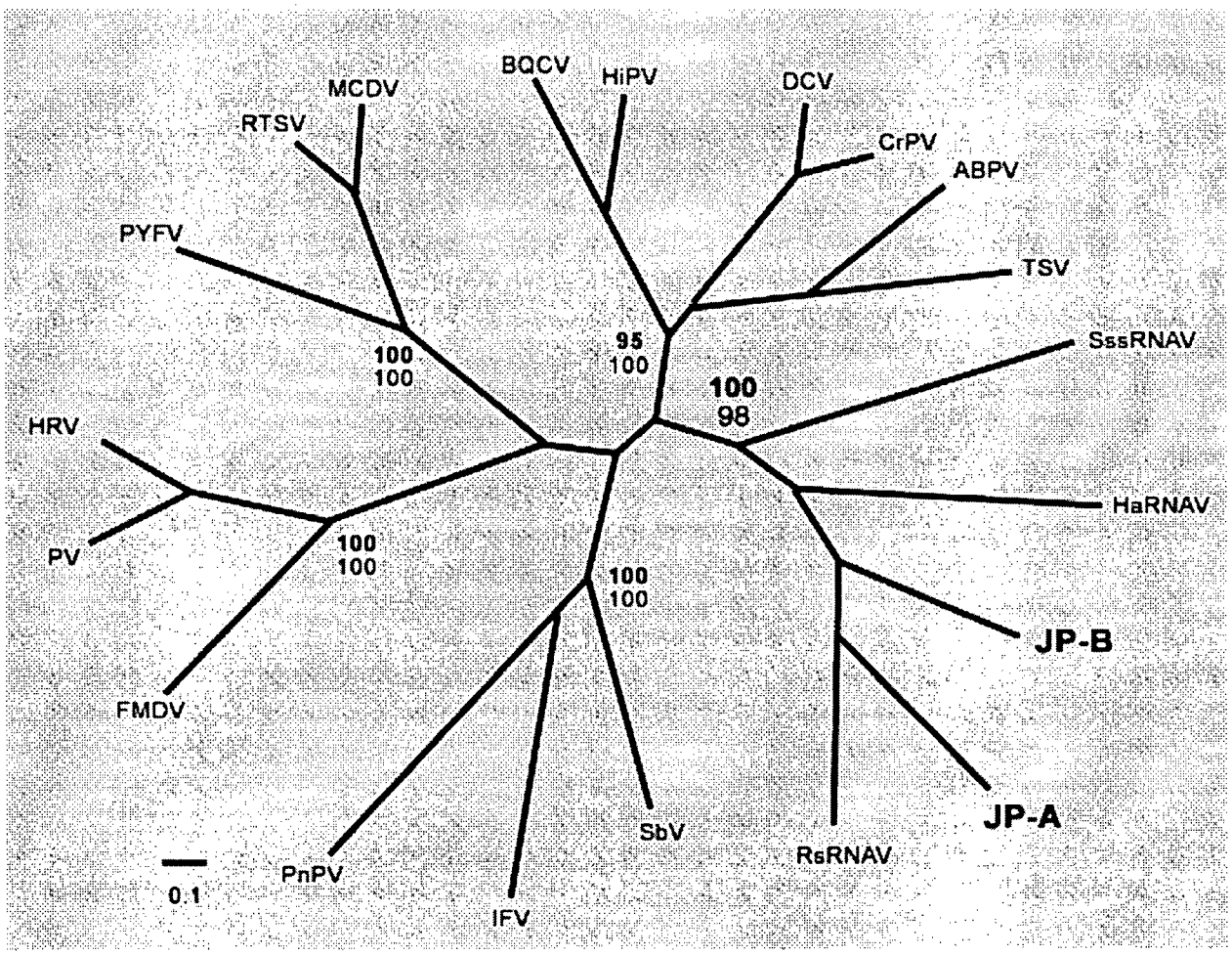

Figure 5.4 Bayesian maximum likelihood trees of aligned concatenated helicase, RdRp and VP3-like capsid amino acid sequences from JP-A and JP-B and other picorna-like viruses

Bayesian clade credibility values are shown for relevant nodes in boldface followed by bootstrap values based on neighbor-joining analysis. The Bayesian scale bar indicates a distance of 0.1. See Table 5.3 for complete virus names and accession numbers.

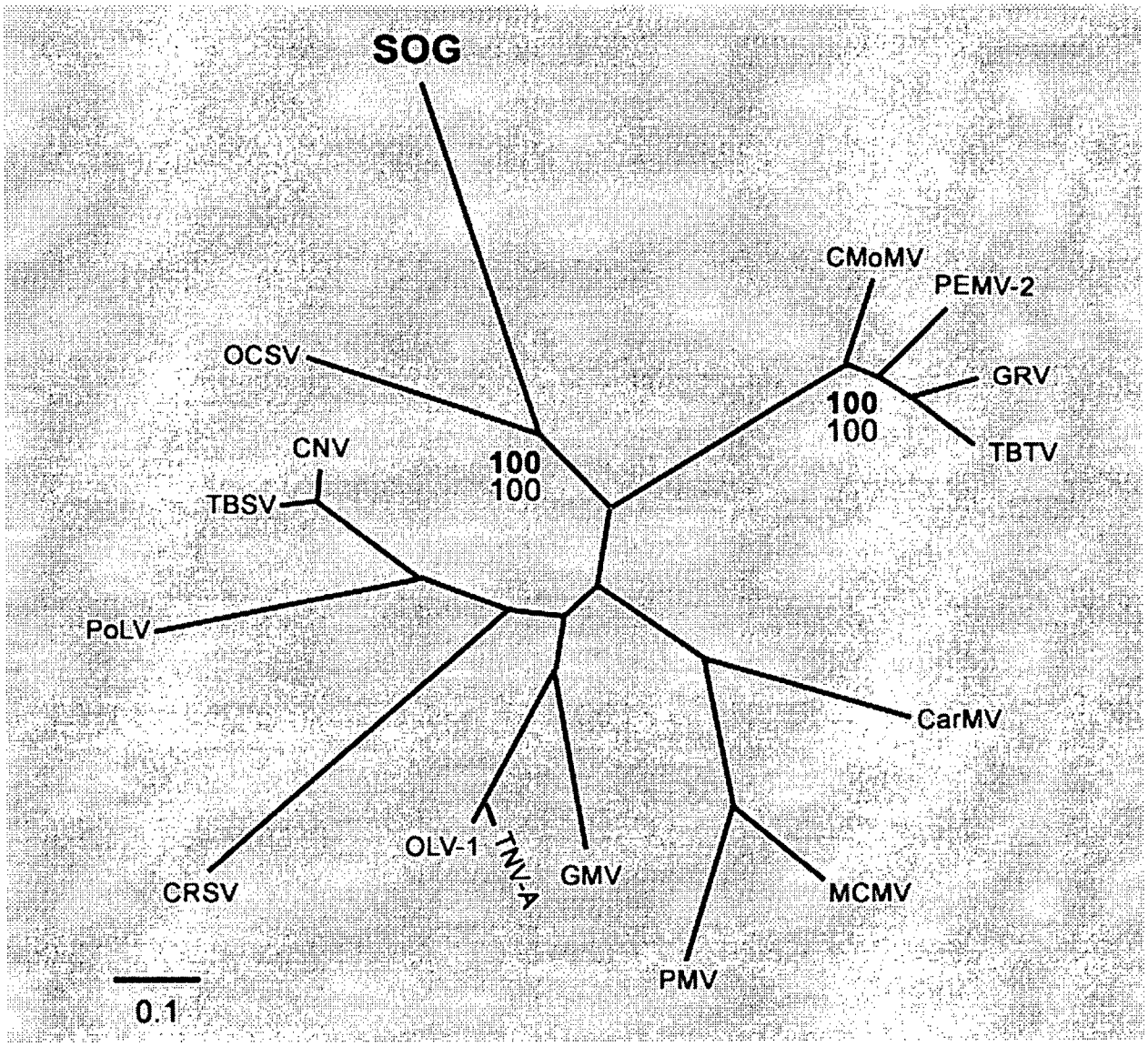

Figure 5.5 Bayesian maximum likelihood trees of aligned RdRp amino acid sequences from SOG and members of the family *Tombusviridae* and unassigned genus *Umbravirus*

Bayesian clade credibility values are shown for relevant nodes in boldface followed by bootstrap values based on neighbor-joining analysis. The Bayesian scale bar indicates a distance of 0.1. See Table 5.3 for complete virus names and accession numbers.

5.5 References

Altekar, G., S. Dwarkadas, J. P. Huelsenbeck, and F. Ronquist. 2004. Parallel metropolis coupled Markov chain Monte Carlo for bayesian phylogenetic inference. Bioinformatics **20:** 407-415.

Altschul, S. F., T. L. Madden, A. A. Schaffer, J. H. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research **25:** 3389-3402.

Cevallos, R. C., and P. Sarnow. 2005. Factor-independent assembly of elongation-competent ribosomes by an internal ribosome entry site located in an RNA virus that infects penaeid shrimp. Journal of Virology **79:** 677-683.

Christian, P., Fauquet, C.M., Gorbalenya, A.E., King, A.M.G., Knowles, N., Legall, O., Stanway, G. 2005. A proposed *Picornavirales* Order *In* C. M. Fauquet [ed.], Microbes in a changing world. International Unions of Microbiological Societies.

Culley, A. I., A. S. Lang, and C. A. Suttle. 2003. High diversity of unknown picorna-like viruses in the sea. Nature **424:** 1054-1057.

Culley, A. I., A. S. Lang, and C. A. Suttle. 2006. Metagenomic analysis of coastal RNA virus communities. Science **312:** 1795-1798.

Czibener, C., D. Alvarez, E. Scodeller, and A. V. Gamarnik. 2005. Characterization of internal ribosomal entry sites of Triatoma virus. Journal of General Virology **86:** 2275-2280.

Hatakeyama, Y., N. Shibuya, T. Nishiyama, and N. Nakashima. 2004. Structural variant of the intergenic internal ribosome entry site elements in dicistroviruses and computational search for their counterparts. RNA **10:** 779-786.

Jan, E., and P. Sarnow. 2002. Factorless ribosome assembly on the internal ribosome entry site of Cricket paralysis virus. Journal of Molecular Biology **324:** 889-902.

Koonin, E. V., and V. V. Dolja. 1993. Evolution and taxonomy of positive-strand RNA viruses - implications of comparative-analysis of amino-acid-sequences. Critical Reviews in Biochemistry and Molecular Biology **28:** 375-430.

Kozak, M. 1986. Point Mutations Define a Sequence Flanking the Aug Initiator Codon That Modulates Translation by Eukaryotic Ribosomes. Cell **44:** 283-292.

Lommel, S. A., Martelli, G.P., Rubino, L. Russo, M. 2004. Tombusviridae, p. 907-936. *In* C. M. Fauquet, Mayo M.A., Maniloff, J., Desselberger, U., Ball, L.A. [ed.], Virus Taxonomy. Eight Report of the International Committee on Taxonomy of Viruses. Elsevier.

Marck, C. 1988. "DNA Strider": a "C" program for the fast analysis of DNA and protein sequences on the Apple Macintosh family of computers. Nucleic Acids Research **16:** 1829-1836.

Nagasaki, K., Y. Tomaru, N. Katanozaka, Y. Shirai, K. Nishida, S. Itakura, and M. Yamaguchi. 2004. Isolation and characterization of a novel single-stranded RNA virus infecting the bloom-forming diatom *Rhizosolenia setigera.* Applied and Environmental Microbiology **70:** 704-711.

Nishiyama, T., H. Yamamoto, N. Shibuya, Y. Hatakeyama, A. Hachimori, T. Uchiumi, and N. Nakashima. 2003. Structural elements in the internal ribosome entry site of Plautia stali intestine virus responsible for binding with ribosomes. Nucleic Acids Research **31:** 2434-2442.

Okayama, H., and P. Berg. 1982. High-efficiency cloning of full-length cDNA. Molecular and Cellular Biology **2:** 161-170.

Smith, A. 2000. Aquatic Virus Cycles, p. 447-491. *In* C. Hurst [ed.], Viral Ecology. Academic Press.

Suttle, C. A., A. M. Chan, and M. T. Cottrell. 1991. Use of ultrafiltration to isolate viruses from seawater which are pathogens of marine phytoplankton. Applied and Environmental Microbiology **57:** 721-726.

Swofford, D. 2000. PAUP*: Phylogenetic Analysis Using Parsimony and other Methods 4.0. Sinauer Associates, Inc.

Takao, Y., K. Nagasaki, K. Mise, T. Okuno, and D. Honda. 2005. Isolation and characterization of a novel single-stranded RNA virus infectious to a marine fungoid protist, *Schizochytrium* sp. (Thraustochytriaceae, labyrinthulea). Applied and Environmental Microbiology **71:** 4516-4522.

Thompson, J. D., T. J. Gibson, F. Plewniak, F. Jeanmougin, and D.G. Higgins 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Research **25:** 4876-4882.

Tomaru, Y., N. Katanozaka, K. Nishida, Y. Shirai, K. Tarutani, M. Yamaguchi, and K. Nagasaki. 2004. Isolation and characterization of two distinct types of HcRNAV, a single-stranded RNA virus infecting the bivalve-killing microalga *Heterocapsa circularisquama*. Aquatic Microbial Ecology **34:** 207-218.

White, K. A., and P. D. Nagy. 2004. Advances in the molecular biology of tombusviruses: gene expression, genome replication, and recombination. Progress in Nucleic Acid Research and Molecular Biology **78:** 187-226.

Chapter VI. Conclusions

6.1 Concluding remarks

6.1.1 Recapitulation

The primary aim of this dissertation was to characterize RNA viruses in the ocean, a previously undescribed component of the marine microbial community. I approached this task from three perspectives, a single isolate (HaRNAV, Chapter II), a specific taxon of viruses (*Picornavirales*, Chapter III) and two assemblages of viral genomes (stations JP and SOG, Chapter IV, Chapter V). These data are some of the first characterizations of the richness of the *in situ* marine RNA virus community.

6.1.2 Bias

The methods developed in this dissertation are based on reverse transcription (RT) of RNA into cDNA and polymerase chain reaction (PCR) amplification of mixed templates. PCR in combination with improvements in cloning, sequencing and bioinformatics have resulted in, among other noteworthy contributions, the identification and classification of thousands of microbes that are not in culture (Rappé & Giovannoni 2003). However, bias associated with every step of a PCR-based assay, including sample collection, nucleic acid extraction, PCR amplification and sequence analysis of PCR amplicons, can contribute to an inaccurate portrayal of the community under examination (Von Wintzingerode et al. 1997). Formatting requirements, particularily in Chapter III and IV, prevented the inclusion of a comprehensive discussion of methodological bias. I have therefore attempted to address this topic in the following sections.

6.1.2.1 Bias associated with sample collection and extraction of RNA

The introduction of bias can occur during routine sample collection. For example, during the production of viral concentrates from seawater samples are typically pre-filtered to remove host organisms (Suttle et al 1991). However, pre-filtration, can result in the removal of viruses larger than the filter pore-size, viruses adsorbed to the material captured on the filter and the destruction of viruses that are particularly delicate. Moreover, marine virus decay rates can be on the order of hours (Weinbauer 2004) and thus the extended time (often greater than 4 hr) required to concentrate viruses from large volumes of seawater may alter the community composition. Once the viral concentrate has been collected, extraction of nucleic acids from environmental samples requires a delicate balance between lysing the most recalcitrant members

of the community and avoiding damage (e.g. through shearing or degradation) to the extracted nucleic acids. RNA presents an additional challenge because it is less stable than DNA and susceptible to degradation by RNases (Von Wintzingerode et al. 1997). Furthermore difficulties can arise during the isolation of RNA due to DNA resistant to removal by enzymatic degradation.

6.1.2.2 Bias associated with RT

Reverse transcriptase synthesizes DNA from an RNA template. The enzyme lacks 3′ to 5′ proofreading exonuclease activity, contributing to its relatively high error rate in comparison to other DNA polymerases (Yang et al. 2002). For example, Superscript II reverse transcriptase (Invitrogen) is approximately 13 times more error prone than Platinum *Taq* polymerase (Invitrogen) (Roberts et al. 1988). Factors such as the concentration of target RNA, the amount of template secondary structure and priming conditions including annealing temperature can significantly affect the precision, efficiency and production of the RT reaction (Stalhberg et al. 2001).

6.1.2.3 Bias associated with PCR

Amplification of environmental targets with PCR can result in differential amplification, the formation of chimeras and heteroduplexes and artifacts from DNA polymerase error, among other biases (Von Wintzingerode et al. 1997, Kanagawa 2003). Polz and Cavanaugh (1998) found that PCR with degenerate primers did not maintain the original ratio of template after 25 cycles and that templates with GC-rich priming sites were preferentially amplified. Moreover, Suzuki and Giovannoni (1996) observed that in reactions with greater than 35 cycles, a 1:1 ratio of products occurred regardless of the initial ratio of target sequences. PCR can also produce artifacts. Chimeras, molecules formed from parts of two different sequences, can comprise approximately 10% of PCR amplified products (Choi et al. 1994) and appear to increase in frequency with cycle number (Von Wintzingerode et al. 1997). The DNA polymerase in PCR can mis-incorporate bases during amplification resulting in sequencing artifacts. Eckert and Kunkel (1991) calculated *Taq* error rates as high as 3×10^{-3} and Acinas et al (2005) identified *Taq* DNA polymerase error as the primary cause of artifacts during the construction of rRNA clone libraries.

6.1.2.4 Bias associated with cloning

The cloning of amplified products can be another significant source of bias. Factors that may lead to cloning bias include the expression of deleterious genes (a significant concern with phages), a decrease in cloning efficiency with increasing insert size, the formation of heteroduplexes and inappropriate antibiotic resistance (Kanagawa 2003). For example, Rainey et al. (1994) observed significant differences in community composition between clone libraries constructed with different cloning systems from the same sample. Although PCR-based methods can be effective in characterizing the richness of a community as well as the identity and phylogeny of its members, estimates of evenness are dubious due to the biases discussed above.

In Chapter III, a degenerate RT-PCR assay targeting a region of the RdRp conserved among picorna-like viruses resulted in the discovery of a diverse array of picorna-like viruses from samples collected from the Strait of Georgia (Figure 3.1). The advantages of this method are that it is relatively simple and inexpensive (compared to a metagenomic approach for example). However, as discussed above, methodological biases limit the application of this approach to the examination of the richness of an RNA viral community only. Nevertheless, targeting an evolutionary informative, conserved gene with degenerate RT-PCR is an approach that can be adapted to interrogate the virioplankton for additional taxa of viruses such as reoviruses, luteoviruses and retroviruses.

6.1.2.5 Bias associated with WGS library construction

The whole-genome shotgun (WGS) library method employed in Chapter IV was based on the linker amplified shotgun library (LASL) protocol described in Breitbart et al. (2002). The LASL method appears to produce a random representation of the original template with an error rate below 1% and no detectable chimeras (Rohwer et al. 2001). A test LASL from a mixed community of previously sequenced vibriophages demonstrated that the average number of clones that contributed to any base is relatively constant over the entire length of the genomes in the original sample (Seguritan et al. 2003). Furthermore, a linear relationship existed between size of sequence overlap and number of contigs ($r^2 = 0.93$), suggesting that sequence fragments where generated from the original templates randomly (http://www.sci.sdsu.edu/PHAGE/LASL/index.htm). Of the 60450 bp generated in the test library, 332 erroneous bases were detected resulting in an error rate of (0.55%). Moreover, no

chimeras were produced in approximately 1000 sequences (http://www.sci.sdsu.edu/PHAGE/LASL/index.htm). The performance of LASL under the test conditions above is impressive, however future experiments should include an evaluation of the method with a more complex viral community and to a greater depth of sequencing.

Research published by Zhang et al. (2006) used a method similar to the WGS approach in chapter IV to examine the RNA viral community in human feces. After extracting total RNA from the viral fraction, RT was conducted with random primers, followed by strand-displacement second-strand synthesis. The double-stranded DNA products were digested with an endonuclease, followed by adapter ligation and PCR with primers targeting sites on the adapters. Amplicons between 500 and 1000 bp were then cloned and sequenced (Zhang et al. 2006). However, no analysis of bias was described.

Several different methods have been developed to accurately interrogate the transcriptosome of single cells (Peano et al. 2006). These methods share several of the same challenges as the examination of natural communities of RNA viruses, starting with very low concentrations of starting RNA template. Surprisingly, those methods that include an exponential amplification step after the addition of primer sites to the target template (like the WGS method described in chapter IV) generally introduced less bias than other approaches such as linear RNA amplification (Iscove et al. 2002, Subkhankulova & Livesey 2006).

Were funding and access to virus isolates no obstacle, a direct way to evaluate error and bias in the RNA virus WGS method would be to construct a test library from a mixed community of characterized RNA virus isolates, including retro-, double-stranded, positive- and negative-sense single-stranded representatives, in known proportions, although it is uncertain whether the results of this exercise would be applicable to an unknown assortment of RNA viral genomes. Nevertheless, the relatively small genome size of RNA viruses makes the construction of whole-genome shotgun libraries a realistic approach to rapidly survey the diversity of marine RNA virus communities. Additionally, this method may be modified to characterize RNA virus populations from a variety of samples such as blood, aerosols and sewage effluent.

6.1.3 Significance of the research

A significant finding of this work was the identification of marine RNA viruses with positive-sense single-stranded genomes that were distantly related to established virus taxa. This

research suggests that there is a persistent and widespread clade of novel picorna-like viruses in the Strait of Georgia, which is comprised of environmental sequences and RNA virus isolates that infect a diversity of protist hosts, for example, HaRNAV infects a photoautotrophic raphidophyte (Tai et al. 2003), while SssRNAV infects a heterotrophic thraustochytrid (Takao et al. 2005, Figure 6.1). In contrast to the marine DNA virus community, which appears to be composed primarily of phages (Edwards & Rohwer 2005), the viral sequences identified in this study are homologous to sequences from RNA viruses that infect eukaryotes, however, a majority of the viral sequences have no significant homology to any known viruses. Because methodological bias was not quantified, it is possible that the marine RNA virus community includes presently unidentifiable RNA bacteriophages. Nevertheless, virus abundance is linked to host abundance; hence, if the marine RNA virioplankton is primarily composed of viruses that infect eukaryotes, it is likely that RNA viruses comprise a small percentage of the total virioplankton. However, as exemplified by the recent isolation of RsRNAV (Nagasaki et al. 2004) and a suite of RNA viruses that infect multicellular organisms from shrimp (Mari et al. 2002) to salmonids (Bernard & Bremont 1995) to sea lions (Smith et al. 2000), it may be that the greatest impact of RNA viruses is that they affect the abundance and population structure of ecologically important marine organisms.

The preceding chapters have provided a preliminary glimpse into a previously uncharacterized component of marine microbial communities and may be of interest to a wide range of scientists. For example, virologists will be interested by the existence of previously unknown families of RNA viruses; biological oceanographers will be interested because previously RNA viruses have not been considered important players in the ocean; microbial ecologists should be intrigued by the discovery of an avenue of previously unexplored microbial diversity in the ocean; scientists interested in emergent pathogens will likely find the discovery of picorna-like viruses of unknown pathogenicity exciting; phycologists will be interested to learn that picorna-like viruses are pathogens of algae. However, more data and better techniques are required before we can examine important topics such as the role of marine RNA viruses in biogeochemical cycling and the evolutionary interrelationships among taxa of RNA viruses. Future research should include the refinement of methods used to characterize RNA virus communities (e.g. virus community collection and metagenomic library construction), determination of the composition and structure of RNA virus assemblages from a greater

diversity of aquatic environments, and the continued sequencing of RNA virus isolates. From these data, quantitative molecular techniques can be used to investigate the dynamics of individual RNA virus-host systems *in situ* and may ultimately lead to a broader understanding of marine RNA virus ecology.

6.2 Figure

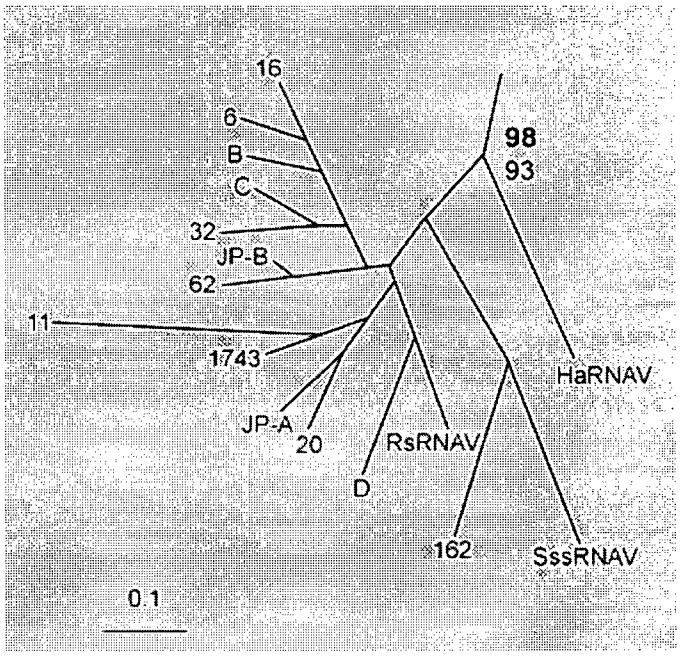

Figure 6.1 Clade of marine picorna-like virus RdRp sequences from Figure 4.3.

Bayesian clade credibility values are shown for relevant nodes in boldface followed by bootstrap values based on neighbor-joining analysis. JP-A and JP-B are from the assembled environmental genomes. The Bayesian scale bar indicates a distance of 0.1. See Table 4.4 for complete virus names and sequence accession numbers.

6.3 References

Acinas, S. G., R. Sarma-Rupavtarm, V. Klepac-Ceraj, and M. F. Polz. 2005. PCR-induced sequence artifacts and bias: Insights from comparison of two 16S rRNA clone libraries constructed from the same sample. Applied and Environmental Microbiology **71:** 8966-8969.

Bernard, J., and M. Bremont. 1995. Molecular biology of fish viruses - a review. Veterinary Research **26:** 341-351.

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

Choi, B. K., B. J. Paster, F. E. Dewhirst, and U. B. Göbel. 1994. Diversity of cultivable and uncultivable oral spirochetes from a patient with severe destructive periodontitis. Infection and Immunity **62:** 1889-1895.

Eckert, K. A., and T. A. Kunkel. 1990. High fidelity DNA-synthesis by the *Thermus-aquaticus* DNA-polymerase. Nucleic Acids Research **18:** 3739-3744.

Edwards, R. A., and F. Rohwer. 2005. Viral metagenomics. Nature Reviews Microbiology **3:** 504-510.

Iscove, N. N., M. Barbara, M. Gu, M. Gibson, C. Modi, and N. Winegarden. 2002. Representation is faithfully preserved in global cDNA amplified exponentially from sub-picogram quantities of mRNA. Nature Biotechnology **20:** 940-943.

Kanagwa, T. 2003. Bias and artifacts in multi-template polymerase chain reaction (PCR). Journal of Bioscience and Bioengineering **96:** 317-323.

Malboeuf, C. M., S. J. Isaacs, N. H. Tran, and B. Kim. 2001. Thermal effects on reverse transcription: Improvement of accuracy and processivity in cDNA Synthesis. BioTechniques **30:** 1074-1085.

Mari, J., B. T. Poulos, D. V. Lightner, and J. R. Bonami. 2002. Shrimp Taura syndrome virus: genomic characterization and similarity with members of the genus *Cricket paralysis-like viruses*. Journal of General Virology **83:** 915-926.

Nagasaki, K., Y. Tomaru, N. Katanozaka, Y. Shirai, K. Nishida, S. Itakura, and M. Yamaguchi. 2004. Isolation and characterization of a novel single-stranded RNA virus infecting the bloom-forming diatom *Rhizosolenia setigera*. Applied and Environmental Microbiology **70:** 704-711.

Peano, C., M. Severgnini, I. Cifola, G. De Bellis, and C. Battaglia. 2006. Transcriptome amplification methods in gene expression profiling. Expert Review of Molecular Diagnostics **6:** 465-480.

Polz, M. F., and C. M. Cavanaugh. 1998. Bias in template-to-product ratios in multi-template PCR. Applied and Environmental Microbiology **64:** 3724-3730.

Rainey, F. A., N. Ward, L. I. Sly, and E. Stackerbrandt. 1994. Dependence on the taxon composition of clone libraries for PCR amplified, naturally occurring 16S rDNA, on the primer pair and the cloning system used. Experientia **50:** 796-797.

Rappé, M. S., and S. J. Giovannoni. 2003. The uncultured microbial majority. Annual Review of Microbiology **57:** 369-394.

Roberts, J. D., K. Bebenek, and T. A. Kunkel. 1988. The accuracy of reverse-transcriptase from HIV-1. Science **242:** 1171-1173.

Rohwer, F., V. Seguritan, D. H. Choi, A. M. Segall, and F. Azam. 2001. Production of shotgun libraries using random amplification. BioTechniques **31:** 108-118.

Seguritan, V., I. W. Feng, F. Rohwer, M. Swift, and A. M. Segall. 2003. Genome sequences of two closely related *Vibrio parahaemolyticus* phages, VP16T and VP16C. Journal of Bacteriology **185:** 6434-6447.

Smith, A. 2000. Aquatic Virus Cycles, p. 447-491. *In* C. Hurst [ed.], Viral Ecology. Academic Press.

Stahlberg, A., J. Hakansson, X. Xian, H. Semb, and M. Kubista. 2004. Properties of the reverse transcription reaction in mRNA quantification. Clinical Chemistry **50:** 509-515.

Subkhankulova, T., and F. J. Livesey. 2006. Comparative evaluation of linear and exponential amplification techniques for expression profiling at the single-cell level. Genome Biology **7:** R18.

Suttle, C. A., A. M. Chan, and M. T. Cottrell. 1991. Use of ultrafiltration to isolate viruses from seawater which are pathogens of marine phytoplankton. Applied and Environmental Microbiology **57:** 721-726.

Suzuki, M. T., and S. J. Giovannoni. 1996. Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. Applied and Environmental Microbiology **62:** 625-630.

Tai, V., J. E. Lawrence, A. S. Lang, A. M. Chan, A. I. Culley, and C. A. Suttle. 2003. Characterization of HaRNAV, a single-stranded RNA virus causing lysis of *Heterosigma akashiwo* (Raphidophyceae). Journal of Phycology **39:** 343-352.

Takao, Y., K. Nagasaki, K. Mise, T. Okuno, and D. Honda. 2005. Isolation and characterization of a novel single-stranded RNA virus infectious to a marine fungoid protist, *Schizochytrium* sp. (Thraustochytriaceae, labyrinthulea). Applied and Environmental Microbiology **71:** 4516-4522.

Von Wintzingerode, F., U. B. Gobel, and E. Stackebrandt. 1997. Determination of microbial diversity in environmental samples: pitfalls of PCR-based rRNA analysis. FEMS Microbiology Reviews **21:** 213-229.

Weinbauer, M. G. 2004. Ecology of prokaryotic viruses. FEMS Microbiology Reviews **28:** 127-181.

Zhang, T., M. Breitbart, W. H. Lee, J. Q. Run, C. L. Wei, S. W. L. Soh, M. L. Hibberd, E. T. Liu, F. Rohwer, and Y. J. Ruan. 2006. RNA viral community in human feces: Prevalence of plant pathogenic viruses. PloS Biology **4:** 108-118.