RNA viruses as major contributors to Antarctic virioplankton

Jaclyn A. Miranda, Alexander I. Culley,† Christopher R. Schvarcz and Grieg F. Steward*
Department of Oceanography, Center for Microbial Oceanography: Research and Education, University of Hawaii at Manoa, 1950 East-West Road, Honolulu, HI 96822, USA.

Summary
Early work on marine algal viruses focused exclusively on those having DNA genomes, but recent studies suggest that RNA viruses, especially those with positive-sense, single-stranded RNA (+ssRNA) genomes, are abundant in tropical and temperate coastal seawater. To test whether this was also true of polar waters, we estimated the relative abundances of RNA and DNA viruses using a mass ratio approach and conducted shotgun metagenomics on purified viral samples collected from a coastal site near Palmer Station, Antarctica on six occasions through a summer phytoplankton bloom (November–March). Our data suggest that RNA viruses contributed up to 65% of the total virioplankton (8–65%), and that, as observed previously in warmer waters, the majority of RNA viruses in these Antarctic RNA virus metagenomes had +ssRNA genomes most closely related to viruses in the order Picornavirales. Assembly of the metagenomic reads resulted in five novel, nearly complete genomes, three of which had features similar to diatom-infecting viruses. Our data are consistent with the hypothesis that RNA viruses influence diatom bloom dynamics in Antarctic waters.

Introduction
Viruses influence the structure of marine plankton communities and food web dynamics by selectively infecting and lysing cells, thereby diverting organic carbon from higher trophic levels to decomposers (Fuhrman 1999; Wommack and Colwell, 2000; Suttle 2007). For many years, studies of the ecology and diversity of viruses in seawater focused almost exclusively on double-stranded (ds) DNA-containing viruses that infect bacteria (Mann et al., 2003; Weinbauer 2004) or eukaryotic phytoplankton (Van Etten et al., 2002). Most metagenomic analyses of the viral fraction of seawater specifically target the dsDNA-containing virions in seawater (e.g., Breitbart et al., 2002; Angly et al., 2006; Bench et al., 2007; Williamson et al., 2008; Steward and Preston, 2011; Hurwitz and Sullivan, 2013). The taxonomic distribution of the small percentage of identifiable reads in these metagenomes suggests that the pool of DNA-containing viruses in seawater is dominated by bacteriophages. More recent work suggests that other types of viruses, namely those containing single-stranded (ss) DNA (Angly et al., 2006; Rosario and Breitbart, 2011; Labonte and Suttle, 2013; Hopkins et al., 2014) or ss- or dsRNA (Culley et al., 2003; Culley et al., 2003; Culley and Steward, 2007; Steward et al., 2013), may also be abundant and contribute significantly to marine plankton ecology.

In contrast to the dsDNA viruses, which appear to be numerically dominated by those infecting bacteria (Edwards and Rohwer, 2005), marine RNA viruses appear to be dominated by those that infect eukaryotes (Lang et al., 2009). Despite the much lower concentrations of eukaryotic relative to prokaryotes in the ocean, a recent study in coastal tropical waters provided evidence that these RNA viruses were just as abundant as DNA viruses (Steward et al., 2013). Since there has been only one study in one location so far, it is not clear whether the inferred high abundance of RNA viruses is a general phenomenon.

We hypothesized that viruses, and RNA viruses in particular, would play an important role in the summer phytoplankton bloom in Antarctic waters. The typical bloom in the Western Antarctic Peninsula (WAP) is initiated by diatoms, which reach their highest concentrations in early to mid-summer (Ducklow et al., 2007). The diatom populations then decline with a transition to dominance by photosynthetic flagellates, which can generate a second peak in chlorophyll (Moline et al., 2004). One possible
driver of the succession in phytoplankton is the increased mortality of diatoms when they reach high population densities. While many studies have investigated grazing as a source of phytoplankton mortality in the Antarctic (Burkill et al., 1995; Calbet et al., 2005; Ross et al., 2008), very few studies have looked at the role of viruses (Guixa-Boixereu et al., 2002; Evans and Brussaard, 2012). Viruses could be a major contributor to succession by driving density-dependent infection and mass lysis of phytoplankton blooms (Bratbak et al., 1993; Nagasaki et al., 1994). Of the 15 or more diatom viruses isolated to date, all contain ssRNA or ssDNA; none are known yet that contain dsDNA (Tomaru and Nagasaki, 2011; Tomaru et al., 2012; Tomaru et al., 2013a,b), suggesting that RNA viruses may be particularly relevant to the ecology of the early season phytoplankton blooms in the WAP.

In this paper, we present the first metagenomic analysis of RNA viral communities in coastal polar waters and estimate their abundance relative to that of dsDNA viruses. We also analyse the changes in viral community composition at 5 m depth through the course of the summer phytoplankton bloom near Palmer Station, Antarctica and present five near-complete RNA viral genomes assembled from the metagenomes.

Results

Hydrographic context

Temperature at our sampling depth of 5 m increased from 
\(-1.1^\circ C\) at the first sampling (November 13, 2010) to a high of 2.3°C in the middle of the time series (January 20, 2011), then dropped to, and remained at c. 1°C ± 0.5°C for the remainder of the samplings (Fig. 1A). Salinity was negatively correlated with temperature \( (r = -0.596, P < 0.001, n = 30) \) and ranged from 33.9 PSU at the start of the study to a minimum of 32.5 PSU at the end of January (January 28, 2011). Chlorophyll \textit{a} concentrations at 5 m depth ranged from 0.76 to 29 mg m\(^{-3}\). Values were relatively low in late austral spring through early summer with a few minor peaks. Chlorophyll \textit{a} concentrations then exhibited larger oscillations during mid-summer, with several more prominent peaks (> 10 \( \mu \text{g Chl} \text{ m}^{-3} \)) observed on January 7 and 13 and February 7, 2011 (Fig. 1B).

Buoyant density and relative abundance of RNA viruses

After banding concentrated viruses in buoyant density gradients, local peaks of DNA and RNA were always observed in the fractions with densities in the range of 1.34–1.51 g ml\(^{-1}\), but RNA concentrations in much lower density fractions were occasionally up to half of the peak values (Supporting Information Fig. S1). For each sample, the single fraction in this range with the highest RNA content was in the narrow range of 1.44–1.47 g ml\(^{-1}\), except for sample E_20110217, which had a peak at 1.40 g ml\(^{-1}\). Second peaks in nucleic acid concentrations of similar magnitude also appeared in fractions with higher densities, typically around 1.60 g ml\(^{-1}\). The yield of purified RNA in the former buoyant density range (i.e., the virus-like RNA) varied from 0.05 to 0.64 ng per liter of water filtered, with the lowest mass recovered at the first sampling early in the season and the highest in mid-season when waters were warmest and chlorophyll \textit{a} concentrations were most dynamic (Fig. 1). Using genome size-normalized RNA to DNA mass ratios for nucleic acids in the viral buoyant density range, we estimated that RNA viruses varied from a low of 8% (range 2–15%) on November 13, 2010 at the beginning of the season to a high of 65% (range 30–79%) on January 27, 2011 in mid-aural summer (Table 1) where the ranges reflect the application of more or less conservative assumptions as described in the Experimental Procedures section. Excluding the lowest value at the first sampling, the contribution of RNA viruses varied from 30% to 65% with an average of around 50% of the detected viruses.

Metagenomic assembly analysis

To verify the viral nature of the ‘virus-like’ RNA and to see how the viral community changed over time, we prepared metagenomic libraries using a single fraction from each sample selected from within the typical RNA virus peak.
Table 1. Relative abundance of RNA viruses in seawater samples collected at Station B throughout the 2010-2011 season.

<table>
<thead>
<tr>
<th>Sample name</th>
<th>Collection date</th>
<th>Volume (L)</th>
<th>RNA copies ($\times 10^3$)</th>
<th>DNA copies</th>
<th>Mean (%)</th>
<th>Inferred % RNA Viruses</th>
</tr>
</thead>
<tbody>
<tr>
<td>A_20101113</td>
<td>13 November 2010</td>
<td>150</td>
<td>1.3</td>
<td>15</td>
<td>8</td>
<td>2–15</td>
</tr>
<tr>
<td>B_20101218</td>
<td>18 December 2010</td>
<td>50</td>
<td>3.7</td>
<td>1</td>
<td>64</td>
<td>29–78</td>
</tr>
<tr>
<td>C_20110107</td>
<td>7 January 2011</td>
<td>45</td>
<td>5.5</td>
<td>7.2</td>
<td>43</td>
<td>15–60</td>
</tr>
<tr>
<td>D_20110127</td>
<td>27 January 2011</td>
<td>50</td>
<td>5.9</td>
<td>3.2</td>
<td>65</td>
<td>30–79</td>
</tr>
<tr>
<td>E_20110217</td>
<td>17 February 2011</td>
<td>45</td>
<td>3.0</td>
<td>6.8</td>
<td>30</td>
<td>9–47</td>
</tr>
<tr>
<td>F_20110303</td>
<td>3 March 2011</td>
<td>45</td>
<td>3.8</td>
<td>4.3</td>
<td>47</td>
<td>17–64</td>
</tr>
</tbody>
</table>

The calculated percent contributions of RNA viruses are relative to the total. Ranges are extreme low and high estimates as described in the Experimental Procedures section. Sample name indicates the sample name given to the metagenome in this paper, where the simple prefix encodes the order of collection (A-F) and replaces the prefix ‘RNAV_PAL’ found in the GenBank Sequence Read Archive (SRA).

The region (1.44–1.47 g ml$^{-1}$). The number of quality-controlled, de-replicated reads varied from 19,150 to 78,321 per sample for a total of 365,902 reads (Table 2). The majority of reads formed contigs (86–98%) in each separate library, and the maximum contig length ranged from 3,996 to 8,187 base pairs (bp) in the individual library assemblies. After sequencing the 3’ end of PAL128, this contig was extended to 8660 bp.

Rarefaction curves produced by Metavir (Roux et al., 2011) with a threshold of 75% suggest that sample A_20101113 had the highest richness, while B_20101218 and F_20110303 had the lowest (Supporting Information Fig. S2A). The number of sequences per sample varied, but normalizing the data to compare an equal number of sequences from each sample resulted in the same trends (Supporting Information Fig. S2A).

Only 15% of the quality-filtered individual reads were classified as viral sequences, and most (82.7%) had no similarity to a known sequence prior to assembly (Supporting Information Fig. S2B). After assembly, nearly half of the total reads (47.4%) were classified as viral sequences based on top BLASTx hits. Of the sequences classified as viral, 97.8% were classified as +ssRNA viruses, most of which were affiliated with viruses within the order Picornavirales (Table 3). Only 0.01% of the sequences were classified as dsDNA viruses and 0.04% classified as dsRNA viruses.

Taxonomic classification of filtered reads and assembled contigs from each individual library revealed some variability throughout the season (Table 3). All libraries contained contigs similar to +ssRNA viruses from unclassified and uncultured members of the Picornavirales and members from the genus Bacillarnavirus. Bacillarnavirus-like sequences dominated in libraries B_20101218, C_20110107, and E_20110217 (> 90% of the sequences) and were also abundant in libraries D_20110127 and F_20110303 (c. 50% of the sequences). Library A_20101113 was the only library to contain contigs similar to members from the genus Labyrinthiviruses, and the Dicistroviridae family was found only in libraries A_20101113, C_20110107, and D_20110127 and in small percentages. Based on filtered reads and assembled contigs, library A_20101113 appears to contain the highest diversity of representatives within the Picornavirales. Overall, the taxonomic classification of assemblies for each library suggests a similar composition throughout the season, with a persistent dominance of +ssRNA viruses in the order Picornavirales.

Genome reconstruction

Reconstruction of RNA viral genomes from the metagenomic sequences of each individual library and from the pooled libraries revealed novel phylotypes with similar genomic organization (Fig. 2), three of which appear to be most closely related to diatom-infecting viral isolates based on RNA-dependent RNA polymerase (RdRp) phylogeny. The contigs PAL128, PAL156, PAL438, PAL473, and PAL_E4 ranged in size from 7,047 to 8,660 bp with a GC content of 39.2–46.6% (Fig. 2, Table 4). Two of the assembled contigs, PAL156 and PAL_E4 appear to represent a large portion of the RNA virus community at different times throughout the season. Of the pooled reads, 12.1% assembled to form PAL156. From the individual libraries, 46% of reads from C_20110107 and 13% from D_20110127 mapped to PAL156. PAL_E4 was assembled with reads from library E_20110217, of which 60.7% formed the contig, and over half of the reads in each individual library mapped to PAL_E4, except for in C_20110107 (only 23%). When all sequences were pooled together, 53% of the total reads from all libraries can be mapped back to PAL_E4. PAL_E4 shared 100% amino acid (aa) identity with another contig from the total reads assembly that covered only the length of open reading frame (ORF) 1 of the PAL_E4 genome and is represented in the same cluster as PAL_E4 in the RdRp tree (below). Additional contigs with at least 98% aa identity to portions of the PAL_E4 genome were assembled from the other individual libraries.

Each of the five assembled genomes contained two ORFs ranging in size from 1,872 to 5,574 bp (Table 4; Fig. 2). The first ORF encodes for non-structural proteins (e.g., the RdRp) in all contigs, while the second encodes for structural proteins (i.e., capsid proteins) in contigs PAL128, PAL156, PAL473 and PAL_E4. All of the genomes contained a region with significant BLASTx hits...
A helicase gene was identified in ORF1 for all genomes except for PAL473, but none contained an identifiable protease gene. Three to four structural proteins were identified in all of the genomes, except for PAL438, that were most similar to the capsid binding site of picornaviruses and the VP4 and capsid protein of dicistroviruses. The second ORF of PAL438 represents a polyprotein with no significant similarity to other known structural proteins, but a portion of this ORF did have a top hit to a helicase gene (E value: 4.2E−11). A polyadenylated [poly(A)] tail was detected for both PAL438 and PAL473.

<table>
<thead>
<tr>
<th>Sample date (YYYYMMDD)</th>
<th>Length (nt)</th>
<th>G+C (%</th>
<th>Assembled (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Count</td>
<td>Mean</td>
<td>Min</td>
</tr>
<tr>
<td>20101113</td>
<td>Reads</td>
<td>57,065</td>
<td>429</td>
</tr>
<tr>
<td></td>
<td>Contigs</td>
<td>1,020</td>
<td>765</td>
</tr>
<tr>
<td>20101218</td>
<td>Reads</td>
<td>74,580</td>
<td>412</td>
</tr>
<tr>
<td></td>
<td>Contigs</td>
<td>248</td>
<td>663</td>
</tr>
<tr>
<td>20110107</td>
<td>Reads</td>
<td>76,980</td>
<td>431</td>
</tr>
<tr>
<td></td>
<td>Contigs</td>
<td>500</td>
<td>675</td>
</tr>
<tr>
<td>20110127</td>
<td>Reads</td>
<td>78,321</td>
<td>416</td>
</tr>
<tr>
<td></td>
<td>Contigs</td>
<td>430</td>
<td>697</td>
</tr>
<tr>
<td>20110217</td>
<td>Reads</td>
<td>19,150</td>
<td>428</td>
</tr>
<tr>
<td></td>
<td>Contigs</td>
<td>316</td>
<td>723</td>
</tr>
<tr>
<td>20110303</td>
<td>Reads</td>
<td>59,806</td>
<td>415</td>
</tr>
<tr>
<td></td>
<td>Contigs</td>
<td>138</td>
<td>668</td>
</tr>
<tr>
<td>Total</td>
<td>Reads</td>
<td>365,902</td>
<td>417</td>
</tr>
<tr>
<td></td>
<td>Contigs</td>
<td>1,794</td>
<td>698</td>
</tr>
</tbody>
</table>

Listed are the number and length (nucleotides, nt) of initial quality controlled reads and of contigs after assembly, percent GC content, and the percentage of reads that were assembled to form a contig.

Table 3. Distribution of hits to viral sequences for the individual sample libraries or a library representing all sequences combined (Total).

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>ssRNA viruses</td>
<td>Picornavirales</td>
<td>Dicistroviridae</td>
<td>1.01</td>
<td>0</td>
<td>0.04</td>
<td>0.04</td>
<td>0</td>
<td>0</td>
<td>0.30</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Marnaviridae</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Picornaviridae</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.04</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unassigned</td>
<td>0.41</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Labymaviridae</td>
<td>13.8</td>
<td>93.9</td>
<td>97.5</td>
<td>52.3</td>
<td>90.9</td>
<td>55.4</td>
<td>43.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bacillarnavirida</td>
<td>47.7</td>
<td>3.96</td>
<td>0.15</td>
<td>41.3</td>
<td>0.60</td>
<td>43.4</td>
<td>10.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unclassified</td>
<td>22.9</td>
<td>0.39</td>
<td>1.42</td>
<td>0.49</td>
<td>2.83</td>
<td>0.83</td>
<td>43.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Uncultured</td>
<td>14.1</td>
<td>1.75</td>
<td>0.78</td>
<td>5.75</td>
<td>5.65</td>
<td>0.35</td>
<td>2.06</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dsRNA viruses</td>
<td>0</td>
<td>0</td>
<td>0.01</td>
<td>0.05</td>
<td>0.02</td>
<td>0</td>
<td>0</td>
<td>0.04</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dsDNA viruses</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.01</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All other viruses</td>
<td>0.11</td>
<td>0.06</td>
<td>0.08</td>
<td>0.07</td>
<td>0.01</td>
<td>0.01</td>
<td>0.15</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(RdRp phylogenetics)

The RdRp gene is a useful molecular marker for investigating the diversity of picornavirids and is currently used for their taxonomic classification by the International Committee on Taxonomy of Viruses (ICTV) (Koonin and Dolja 1993). A maximum likelihood tree of RdRp gene sequences shows the identified environmental sequences from this study relative to other environmental sequences and known viruses within the order Picornavirales (Fig. 3). A search for RdRp sequences in the assemblies and unassembled reads for each metagenome, based on four of the

(E value ≤ 10−5) to + ssRNA RdRps in the first ORF. A helicase gene was identified in ORF1 for all genomes except for PAL473, but none contained an identifiable protease gene. Three to four structural proteins were identified in all of the genomes, except for PAL438, that were most similar to the capsid binding site of picornaviruses and the VP4 and capsid protein of dicistroviruses. The second ORF of PAL438 represents a polyprotein with no significant similarity to other known structural proteins, but a portion of this ORF did have a top hit to a helicase gene (E value: 4.2E−11). A polyadenylated [poly(A)] tail was detected for both PAL438 and PAL473.
seven conserved motifs (mpl-like region), returned 22 unique picornavirad-like sequences from A_20101113, 3 from B_20101218, 29 from C_20110107, 17 from D_20110127, 5 from E_20110217, and 8 from F_20110303. Clustering at a 75.7% aa identity threshold to approximate conservative species-level diversity, as described in the Experimental Procedures section, resulted in 13 clusters in A_20101113, 2 in B_20101218, 2 in C_20110107, 3 in D_20110127, 4 in E_20110217, and 3 in F_20110303. Of those clusters, four were found in at least one other sample at 100% aa identity, and seven were found at ≥ 94% aa identity.

Phylogenetic analysis of the translated RdRp genes from individual reads and assembled contigs revealed that most of the identified picornavirad-like sequences were most closely related to other environmental RdRp sequences and diatom-infecting virus isolates and were only distantly related to other cultivated isolates (Fig. 3). These phylotypes were clustered based on a 98% aa identity, and the normalized proportion of reads from each library that comprise each sequence cluster is color-coded in the pies (Fig. 3). Many of the identified phylotypes consist of sequences from library A_20101113, and most clusters contain sequences from two or more libraries. None of the RdRp sequences were closely related to other known viruses within the families Dicistroviridae, Secoviridae, Picornaviridae, Iflaviridae or Labyrnaviridae. The RdRp from contig PAL473 was most closely related to sequences from two pennate diatom-infecting viruses (CcloRNAV01 and 02, bootstrap value 96%). The RdRp from PAL156

![Genome maps of contigs PAL128, PAL156, PAL438, PAL473, and PAL_E4. The first four genomes were assembled from the pooled reads of all six viromes, and the last was assembled from reads in library E_20110217. The conserved gene regions were identified by the CDD NCBI BLAST searches. The helicase genes are shown in purple, the RNA-dependent RNA polymerase genes in green, and the structural genes in red, pink, light orange and yellow. Untranslated regions are shown in light grey and the ORFs are shown in orange. The light blue A indicates where a polyadenylated tail was identified. The pie charts on the right show the normalized percent of reads that mapped to each assembled genome from each metagenome sample. Libraries are designated by the color indicated in the legend that corresponds to each collection date (YYYYMMDD).]

**Fig. 2.** Genome maps of contigs PAL128, PAL156, PAL438, PAL473, and PAL_E4. The first four genomes were assembled from the pooled reads of all six viromes, and the last was assembled from reads in library E_20110217. The conserved gene regions were identified by the CDD NCBI BLAST searches. The helicase genes are shown in purple, the RNA-dependent RNA polymerase genes in green, and the structural genes in red, pink, light orange and yellow. Untranslated regions are shown in light grey and the ORFs are shown in orange. The light blue A indicates where a polyadenylated tail was identified. The pie charts on the right show the normalized percent of reads that mapped to each assembled genome from each metagenome sample. Libraries are designated by the color indicated in the legend that corresponds to each collection date (YYYYMMDD).

**Table 4. Summary of assembled genomes.**

<table>
<thead>
<tr>
<th>Genome</th>
<th>Size (bp)</th>
<th>% GC</th>
<th>Average coverage of reads</th>
<th>% Reads mapped</th>
<th>ORF 1 size (bp)</th>
<th>ORF 2 size (bp)</th>
<th>5' UTR</th>
<th>IGR</th>
<th>3' UTR</th>
<th>% UTR</th>
<th>Poly (A) tail</th>
</tr>
</thead>
<tbody>
<tr>
<td>PAL128</td>
<td>8660</td>
<td>42.2</td>
<td>81</td>
<td>0.4</td>
<td>5574</td>
<td>2397</td>
<td>195</td>
<td>67</td>
<td>427</td>
<td>8.0</td>
<td>ND</td>
</tr>
<tr>
<td>PAL156</td>
<td>7897</td>
<td>41.9</td>
<td>2,258</td>
<td>12.1</td>
<td>4764</td>
<td>2640</td>
<td>69</td>
<td>315</td>
<td>109</td>
<td>6.2</td>
<td>ND</td>
</tr>
<tr>
<td>PAL438</td>
<td>7109</td>
<td>39.2</td>
<td>29</td>
<td>0.1</td>
<td>4623</td>
<td>1872</td>
<td>5</td>
<td>223</td>
<td>324</td>
<td>7.8</td>
<td>Y</td>
</tr>
<tr>
<td>PAL473</td>
<td>6360</td>
<td>42.7</td>
<td>25</td>
<td>0.1</td>
<td>3153</td>
<td>2700</td>
<td>104</td>
<td>200</td>
<td>203</td>
<td>8.0</td>
<td>Y</td>
</tr>
<tr>
<td>PAL_E4</td>
<td>8187</td>
<td>46.6</td>
<td>604*</td>
<td>52.9</td>
<td>4980</td>
<td>2748</td>
<td>65</td>
<td>133</td>
<td>261</td>
<td>5.6</td>
<td>ND</td>
</tr>
</tbody>
</table>

Listed are the estimated minimum genome size, percent GC content, average coverage, percent of reads mapped from the pooled metagenomes, sizes of open reading frames (ORFs), untranslated regions (UTRs), and intergenic regions (IGR), and whether or not a poly(A) tail was identified. The asterisk (*) indicates numbers relative to only library E_20110217, as this genome was assembled with only those reads.
was most closely related to the RdRp of a centric diatom-infecting virus in the same group with one of the bacillarna-viruses (Csp03RNAV), while PAL128 and PAL438 were only closely related to other environmental sequences from both temperate and tropical coastal seawater. The PAL_E4 RdRp formed a deeply branching cluster with other RdRp contigs from the total library assembly that was not similar to any known isolates.

Discussion

RNA virus abundance

Our estimates of the numerical contribution of RNA viruses to the viroplankton in the polar waters of the WAP were similar to those previously reported for tropical waters (Steward et al., 2013). The two study sites are geographically distant, have large differences in temperature and seasonality of light flux, and differ in the variable limiting primary production (nutrients vs. light). Despite these differences, both study sites are coastal habitats prone to blooms of eukaryotic plankton, including a large contribution from diatoms. It therefore remains unclear whether a high relative abundance of RNA viruses is a general phenomenon in the ocean or is restricted to such coastal habitats.

It is also possible that the relative abundance of RNA viruses we calculated in both cases derives from methodological artifacts. Our approach has been indirect and relies on concentrated viral assemblages and fluorescence-based measures of nucleic acid concentrations, which could lead to biases in apparent abundances. Differences in the recovery efficiencies of DNA- vs. RNA-containing viruses, for example, could skew our relative abundance estimates in either direction. We note however, that two very different virus concentration procedures (flocculation vs. ultrafiltration) have now been used with similar results. Although we have tested the fluorescence-based nucleic acid assays and the selective enzymatic degradation of

Fig. 3. RdRp phylotypes. This unrooted phylogenetic analysis (maximum likelihood; model WAG; 1000 bootstraps) is based on the alignment of the conserved domains 4 to 7 of the RdRp gene (Koonin and Dolja, 1993; Culley et al., 2014). The tree shows the position of environmental sequences from this study (marked with pies) and previous studies (Culley and Steward, 2007, Culley et al., 2003, 2007, 2014), relative to known viruses in the order Picornavirales. The branches of cultured isolates are shown with a dashed line and the solid lines indicate environmental sequences. The pies represent phylotype clusters (98% aa identity) from this study. The area of each pie is proportional to the log of the total number of reads in the cluster, and the wedges represent the normalized percent of all reads (contigs or singletons) contributed from each metagenome. Libraries are designated by the color indicated in the legend that corresponds to each collection date (YYYYMMDD). The asterisk (*) indicates unpublished diatom-infecting viruses isolated from Kaneohe Bay. PAL128, PAL156, PAL438, PAL473, and PAL_E4 indicated on the tree represent the five nearly complete genomes assembled in this study. The scale bar is equal to 0.3 substitutions per site. The specific sequences from the alignment are listed in Supporting Information Table S1.
DNA vs. RNA used to distinguish between them (Steward et al., 2013), a caveat is that these tests were conducted with purified nucleic acids of known composition (dsDNA and ssRNA). Incomplete digestion of nucleic acids extracted from the purified viral fraction of natural samples, or variations in fluorescence yields for single- vs. double-stranded DNA or RNA, could lead to inaccuracies in our estimates of nucleic acid concentrations. Specifically, the contributions of ssDNA and dsRNA are poorly constrained at present.

Like the ssRNA viruses, ssDNA viruses are small and difficult to detect (Tomaru and Nagasaki, 2007; Holmfeldt et al., 2012). Metagenomic studies reveal their presence (Angly et al., 2006; Rosario et al., 2009; Labonte and Suttle, 2013), but have not yet shed light on their numerical contribution relative to dsDNA viruses, because the two most common amplification methods used prior to generating libraries bias either for (rolling circle amplification) or against (linker ligation) ssDNA (Kim and Bae, 2011). Of direct relevance to our studies is that ssDNA is not efficiently detected by the DNA assay we used. As a consequence, we may have underestimated the abundance of DNA viruses if those containing ssDNA comprise a large fraction of the total DNA-containing viruses. Our estimates of RNA virus abundance should therefore be interpreted as relative to dsDNA viruses only.

The evidence available so far suggests that marine planktonic viruses containing RNA are predominantly viruses of eukaryotes, while those containing dsDNA are predominantly viruses of prokaryotes (Edwards and Rohwer, 2005; Lang et al., 2009; Kristensen et al., 2010). Analyses of metagenomes prepared from virus-like RNA (this study) vs. DNA (Brum et al., 2016) from Antarctic coastal waters during the same season and at the same location are consistent with this view. In the latter study, bacteriophages dominated the identifiable virus sequences and relatively few sequences resembling known algal viruses were detected. However, the library preparation method was not designed to detect ssDNA. The frequent isolation of ssDNA viruses that infect diatoms (e.g., Nagasaki et al., 2005; Tomaru et al., 2013b; Kimura and Tomaru, 2015), suggests that these types of viruses may also play an important role in the diatom-dominated Antarctic summer phytoplankton bloom, but this has not yet been investigated.

Given the strong seasonality in the community composition in polar waters, with blooms of eukaryotic phytoplankton in the summer transitioning to bacterial heterotrophy through the long dark winter (Ducklow et al., 2007; Ducklow et al., 2012; Steinberg et al., 2012), we hypothesized that there would be a seasonal cycle in the relative abundance of RNA vs. dsDNA viruses. Our data are consistent with this hypothesis, with the lowest relative abundance of RNA viruses occurring in the earliest spring sampling and the highest in mid-summer, but there is considerable variability among the samples. Analysis of additional samples over multiple seasons and at least one complete annual cycle will be necessary to properly test this idea.

### Diversity of RNA viruses

A phylogenetic comparison of RdRp genes found in the Antarctic metagenomes to other known RdRps revealed that most of the Antarctic RdRps (including those from three assembled genomes) fall within a rather broad cluster of putative protistan-infecting viruses similar to the results from tropical waters (Culley and Steward, 2007). The only cultivated representatives in this portion of the tree so far, are all viruses infecting diatoms. It was noted that picornavirads tend to cluster by the type of host that they infect (Culley and Steward, 2007), but most of the data informing this observation were from viruses of plants and animals. If this holds true for the diverse, uncultivated marine picornavirads, it suggests that many of the environmental sequences we identified may also infect diatoms (or perhaps other stramenopiles). This is consistent with the predominance of diatoms in the austral summer phytoplankton bloom in the coastal waters of the WAP (Ducklow et al., 2007) and suggests that viral infection may contribute significantly to the seasonal phytoplankton dynamics in this region. Some of the RdRp phylogenotypes were more divergent and did not cluster with any other known viruses, yet they appear to be persistent throughout the season and represent a significant portion in the RNA virus community. A concerted effort to isolate more virus-protist systems from diverse ocean habitats may be helpful in placing novel viruses like these into an ecological context.

The diversity of RNA viruses (as suggested by the number of protein clusters in the rarefaction curves and the RdRp-based phylogeny) appeared to be highest at the beginning of the season (A_20101113), when the abundance of RNA viruses also appeared to be low. The apparent increase in abundance, and decrease in RNA virus diversity, in subsequent samplings is consistent with viruses responding to phytoplankton blooms dominated by relatively few taxa. The inferred richness appears to vary little throughout the rest of the season, but bacillamaviruses appear to be most highly represented in libraries B_20101218, C_20110107, and E_20110217. Many of the RdRp phylogenotypes are represented in multiple metagenomes (four of which were found with 100% aa identity), which suggests a persistence of some virus types throughout the season. There may also be a succession of different phylogenotypes infecting similar hosts, as seen with previous virus-host bloom dynamics in mesocosm experiments (Schroeder et al., 2003). However, the resolution of sampling and taxonomic classifications does not allow us
to draw conclusions about whether the variations in viral community composition that we observed represent meaningful ecological shifts or simply reflect sampling variability. These are questions we will address in a future report using RT-qPCR to quantify the dominant RNA virus phylogenotypes at higher temporal resolution.

Assembled genomes

The five nearly complete genomes assembled from the libraries in this study represent a significant contribution to the nine marine picornavirad genome sequences assembled from previous marine RNA viral metagenomes (Culley et al., 2007; Culley et al., 2014). However, without any host data, we can only assess potential ecological significance through phylogenetic analyses. The genomic configuration and RdRp phylogenetic placement of PAL128, PAL156 and PAL473 are most similar to that of viruses within the genus *Bacillarnavirus*, a diatom-infecting group of viruses, again suggesting the importance of these viruses in influencing the dynamics of their dominant diatom hosts. PAL438 and PAL_E4 contain dicistronic genome structures (similar to Bacillarnavirus), but contain RdRps that are highly divergent from any known picornavirads. Although the second ORF of PAL438 does not contain any genes similar to the structural genes in the NCBI database, and the syntenous regions of the genome were slightly smaller than the others, this contig was verified with two sets of primers overlapping both ORFs. Thus, this may be a very divergent type of virus with unique structural proteins since there are currently no examples of this genome configuration within the order *Picornavirales*. None of the genomes contained sequences with significant BLAST hits to any proteases, a similar finding to that reported by Culley et al., (2014) where only two of six assembled genomes contained sequences with hits to a protease. However, since this gene is required for +ssRNA viral replication and an ORF exists in the location where a protease is normally encoded, we presume these viruses contain a highly divergent protease.

The number of reads that recruit to a given contig can provide some insight to the relative abundance of the assembled contig and the composition of the RNA viral community at a given time. These data suggest that in January (the dates of collection for libraries C_20110107 and D_20110127), PAL156 may represent nearly half of the total RNA viruses. PAL_E4 on the other hand, may represent nearly half of the total RNA viruses during all other sampling periods of the season. The high similarity of PAL_E4 to other contigs assembled from all of the individual library assemblies suggests that this phylotype (or perhaps members of a group of closely related phylootypes) was likely present throughout the season. The other contigs, PAL128, PAL438 and PAL473, recruited less than 1% of the reads from each of the viromes, suggesting that they do not represent a very large proportion of the viruses at any given sample date throughout the season. These estimates based on recruitment are subject to biases from metagenomic library construction through cDNA synthesis or PCR amplification (Culley et al., 2014) and would need to be confirmed through reverse-transcription quantitative-PCR (RT-qPCR).

**RNA viral metagenomes**

An overwhelming majority of the sequences in our metagenomes prepared from Antarctic marine RNA viruses matched +ssRNA virus groups identified within the order *Picornavirales*, a finding very similar to those of studies in temperate (Culley et al., 2003) and tropical (Steward et al., 2013; Culley et al., 2014) systems. A comparison of the metagenomes in this study with the Kaneohe Bay RNA viral metagenomes (Culley et al., 2014) revealed similarities among the communities. In both cases, the sequences were either dominated by those most similar to viruses or had ‘no hits’ in the database, with less than 4% of sequences matching cellular organisms, suggesting that our multi-step purification procedure was highly selective for viruses. Of the viral sequences, the majority matched +ssRNA viruses within the order *Picornavirales*, and very few matched dsRNA viruses (0.04% at Palmer Station and 1.3% in Kaneohe Bay). Differences among the metagenomes were apparent within the ssRNA virus classifications (Fig. 4). The Kaneohe Bay samples had a much larger representation of *marnaviroids* and *dicistroviroids* of which the latter might be expected in a coastal habitat with a high abundance of copepods and more terrestrial input and runoff. The Palmer Station LTER Station B metagenome had a much higher percentage of bacillarnavirus-like sequences, perhaps reflecting the importance of diatoms to summer productivity.

Although all three of the studies mentioned above indicate that picornavirads dominate the pool of marine RNA viruses, it is possible that this simply reflects biases in library preparation. The random-priming sequence-independent single-primer amplification (RP-SISPA) reaction we employed, for example, could be biased towards these viruses. Studies of the SISPA amplification procedure have suggested biases in coverage within genomes that is driven by the non-random portions of the primer (Rosseel et al., 2013). It is conceivable that such an effect could lead to biased amplification of picornavirads relative to other types of viruses, but there is no empirical evidence for this. Djikeng and colleagues (2008) assessed genome coverage after amplification with RP-SISPA and found little difference among different RNA virus types (MS2, HRV16, TA and NDV), suggesting that among RNA viruses with ss genomes (both + and − sense), picornavirads (i.e.,
HRV16) are not amplified with any particular bias. A more relevant concern is the observed tendency of SISP A to lead to over-representation of the dominant sequences (Karlsson et al., 2013). If this effect were influencing our results, picornavirads might still be dominant, but perhaps not to the extent they appear in the metagenomic data. Even so, a bias in the SISP A method we used here seems unlikely to explain the picornavirad dominance, since this phenomenon was first observed in coastal temperate waters (Culley et al., 2003) using a different, linker-based method of amplification (Reyes and Kim 1991; Breitbart et al., 2002).

However, we cannot yet rule out all other potential sources of bias. It is possible, for example, that picornavirads are unusually robust among the RNA viruses and preferentially survive the initial virus concentration procedures. There may also be biases during the RT step against dsRNA (as a result of inefficient denaturation) or in favor of poly-A tailed RNA (as a result of a small amount of poly-T-tailed primer along with the excess random hexamer-tailed primers).

**Conclusions**

Our data on the mass ratios of virus-like nucleic acids suggest that the abundance of RNA viruses is equal to or greater than dsDNA viruses in the waters near Palmer Station for much of the summer and that these RNA viruses are dominated by members of the order Picornavirales. This is consistent with previous metagenomic studies of marine RNA viruses, but the possibility remains that these studies are influenced by methodological biases. It is likely that neither ssDNA nor dsRNA are proportionally accounted for by current protocols for quantifying viral RNA and DNA or preparing metagenomes. Whether this significantly distorts our view of the relative contributions of DNA vs. RNA viruses in the virioplankton is not yet known. Nevertheless, our data illustrate for the first time that RNA-containing viruses are a dynamic component of the plankton in Antarctic coastal waters and are likely an important source of phytoplankton mortality during the intense annual summertime bloom.

**Experimental procedures**

**Field sample collection and processing**

Hydrographic data and seawater samples were collected from the Palmer Long-Term Ecological Research (PAL-LTER) Station B (64° 46.77′S, 64° 04.35′W), located south of Anvers Island on the WAP. Temperature and salinity were recorded with a Seabird CTD nearly biweekly from November 3, 2010 through March 3, 2011. Chlorophyll a concentrations were determined by fluorometry of extracted pigments (Parsons et al., 1984) and obtained from the Palmer LTER data site (http://oceaninformatics.ucsd.edu/datazoo/data/palldata/data sets). Metagenome samples were collected on six occasions from November 13, 2010 to March 3, 2011 (Table 1). Seawater was pumped from 5 m depth with a submersible stainless steel pump (Monsoon XL DTW; Proactive Environmental Products, Bradenton, FL, USA), passing directly through an in-line polyethersulfone membrane filter capsule (Whatman 0.2 μm Polycap TC 36; GE Healthcare Bio-Sciences, Piscataway, NJ, USA) into acid-washed, polycarbonate carboys. The filtered seawater was transported to a cold room (4°C) within 1 h and stored there until it could be processed (within 2 h). High-molecular-weight dissolved and colloidal organic matter (which included viruses) was concentrated to a final volume of about one liter via tangential flow ultrafiltration with 0.3 m2 of composite regenerated cellulose membranes (three stacked Pellicon 2 mini cassettes, 30 kDa Ultrace; Millipore, Billerica, MA, USA). The retentates

---

**Fig. 4.** Taxonomic classification of assemblies from Palmer Station metagenomes compared with Kaneohe Bay metagenomes (Culley et al., 2014). Percentages are relative to the total ssRNA virus hits for the pooled library assembly from Palmer Station and for the combined data from Kaneohe Bay 2009 and 2010 metagenomes. Krona charts were made with Krona (Ondov et al., 2011).
were concentrated further with centrifugal ultrafiltration devices (30 kDa, Centricron 70, Millipore) by centrifuging at 2500 RPM for 20 min in a swinging bucket rotor (GH-3.8A; Beckman Coulter Inc., Brea, CA, USA). Each sample was concentrated on one device by repeated emptying of the filtrate and refilling of the sample reservoir between spins until the entire sample was concentrated. Samples were recovered by inverting the filter into the collection cup provided and centrifuging at 2000 RPM for 3 min, then transferring the retentate (final volume ~1.5 ml) to a 2 ml polypropylene tube (screw-cap with o-ring). Concentrates were flash-frozen in liquid N₂, stored at ~80°C, then shipped to our home laboratory on dry ice. Samples were thawed and viruses were purified through sequential step and continuous CsCl buoyant density gradients (Lawrence and Steward, 2010). Fractions of approximately 0.5 ml were collected from the final gradient from top to bottom using an Auto Densi-Flow (Labconco, Kansas City, MO, USA).

**Buoyant density fraction analysis**

The density of each fraction was measured using a positive-displacement micropipet and analytical balance (Lawrence and Steward, 2010). After buffer exchange into SM (100 mM NaCl, 8 mM MgSO₄, 50 mM Tris, pH 7.5) using centrifugal ultrafiltration units (0.5 ml Amicon Ultra 30 kDa; Millipore), nucleic acids were extracted from a portion of each fraction with the MasterPure Complete DNA and RNA Purification Kit (Epicentre, Madison, WI, USA), according to the manufacturer's instructions. Extracted fractions were split for subsequent DNA and RNA analyses. The total DNA in each fraction was measured by fluorometry using the Quant-iT DNA Assay Kit (Life Technologies, Carlsbad, CA, USA) in a multi-well plate fluorometer (2030 Multilabel Reader, PerkinElmer, Waltham, MA, USA). Samples for RNA analysis were treated with TURBO DNase (Life Technologies) to avoid non-specific signal and RNA was quantified by fluorometry using the Quant-iT RNA Assay Kit (Life Technologies) in a cuvette fluorometer (TD-700, Turner Designs, Sunnyvale, CA, USA).

**Calculation of RNA and DNA virus abundance**

To estimate the number of RNA and DNA viruses present in each sample, we used a mass apportioning approach, as previously described (Steward et al., 2013). The mass of RNA was summed for the fractions in a narrow buoyant density range (1.39–1.51 g ml⁻¹) that seems to be typical of many marine RNA viruses (Steward et al., 2013; Culley et al., 2014). We summed the DNA content from a slightly broader buoyant density range (1.34–1.51 g ml⁻¹) that encompassed the main DNA peak in the virus density range. Based on the data in the Ninth Report of the International Committee on Taxonomy of Viruses and as reported previously (Steward et al., 2013), we calculated an average RNA mass of 5.1 × 10⁻¹⁸ g per 9 kb ssRNA virus. We assumed that all unknown sequences from the viral RNA fraction (classified as ‘No hits’) were derived from RNA viruses. We assumed an average DNA mass of 5.5 × 10⁻¹⁷ g per virion, which is equivalent to 50 kb of dsDNA, an average that was found in a wide range of environments (Steward et al., 2000). After converting the RNA and DNA masses to viral abundances, we calculated the percentage of total viruses that were RNA-containing. To get a sense of what the possible range of the percentage might be, we used more extreme assumptions as previously described (Steward et al., 2013). For the high estimate, we assumed the average DNA content per dsDNA virus was twofold greater (100 kb), as calculated for marine metagenomic data (Angly et al., 2006). For the low estimate, we assumed that the average dsDNA viral genome was twofold smaller (25 kb), and that only the sequences from the putative viral RNA fraction with significant hits to viruses in GenBank (47% of the total) were viral.

**Metagenome construction**

For each sampling date, a single fraction was chosen for sequencing that represented the buoyant density at which the peak RNA was found in all but one sample (E_20110217). In this latter sample, the fraction with a buoyant density most similar to the others (1.45 g ml⁻¹) was used for consistency, even though it was not the peak (Supporting Information Fig. S1). The RNA extracted from these selected fractions was amplified with a random-priming sequence-independent single-primer amplification (RP-SISPA) approach (Djikeng et al., 2008; Culley et al., 2010), followed by a reconditioning reaction to both increase yield and minimize heteroduplex formation (Duhaime et al., 2012). Discrepancies in the volumes and concentrations for the amplification reaction have appeared in the literature, so the complete reaction conditions are provided hereof. For random-priming cDNA synthesis, 5–10 μl of extracted viral RNA was added to a 0.2 ml PCR tube with SISPA primers: 20 pmol FR26RV-N (GCCGGAGCTCTGCAGATATC(T)20), 10 nmol each dNTP, and a final extension at 68°C for 1 h to allow efficient reverse transcription, 50°C for 3 min to denature the product and allow annealing of additional primer, and 94°C for 5 min to disrupt RNA secondary structure, then placed on ice for at least 1 min. While still on ice, 1 × First Strand Buffer, 0.1 μmol DTT, 40 U RNase OUT, and 200 U Superscript III were added to the tube and mixed gently (20 μl final volume). The reactions were incubated sequentially at 25°C for 10 min to keep random hexamers annealed and initiate transcription, 50°C for 1 h to allow efficient reverse transcription, and 94°C for 3 min to denature the product and allow annealing of additional FR26RV-N primer to the first-strand cDNA product. Samples were then rapidly cooled to 4°C. A complementary second DNA strand was then synthesized by adding 2.5 U Klenow Fragment 3’-5’ exo– (New England Biolabs Inc., Ipswich, MA, USA) and incubating at 37°C for 1 hr. The reaction was terminated by incubating at 75°C for 10 min.

Amplification of the cDNA was carried out in triplicate 50 μl reactions containing 1–5 μl of template with a final concentration of 1 × Reaction Buffer, 0.2 mM dNTPs, 1 μM FR20RV (SISPA) primer, and 2.5 U Expand High Fidelity Plus Enzyme (Roche, Indianapolis, IN, USA). PCRs were incubated at 94°C for 2 min, followed by 32 cycles of denaturation at 94°C for 30 s, annealing at 65°C for 30 s, extension at 68°C for 4 min, and a final extension at 68°C for 7 min. A fraction of each reaction was loaded onto a 1% agarose gel containing 1 × SYBR Safe stain and 1 × TAE and run at 100V for 40 min to ensure that amplification resulted in visible PCR products (typically a smear 250–1000 bp). Reconditioning reactions were then
performed on 10-fold dilutions of amplified DNA in a fresh PCR mix (200 μl reactions with 20 μl original PCR mixture as template, and all other reagents in the same concentrations as the small-scale PCR). Thermal cycling conditions were the same as before, but with only three cycles of denaturation, annealing, and extension. All PCR products were pooled and concentrated in a centrifugal ultrafiltration unit (0.5 ml Amicon Ultra 100 kDa; Millipore), purified with QIAquick PCR Purification Kit (Qiagen, Valencia, CA, USA), and eluted with 150 μl TE which was warmed to 80°C immediately prior to adding to the column. Samples were loaded onto a 0.5% agarose gel containing 1 × SYBR Safe stain and 1 × TAE, and run at 100V for 60 min. Amplicons of 500–1000 bp were gel purified to obtain 100–500 ng per sample and diluted to equal concentrations prior to sequencing.

Amplicons were sequenced with the 454 GS FLX+ System (Roche) at the ASGPB sequencing facility at the University of Hawai'i at Mānoa. The yield of high-quality reads was low, so samples were resubmitted for additional sequencing with the 454 GS FLX System (Roche). For each metagenome, the sequences from both runs were pooled and passed through the same quality-control parameters to remove any low quality reads. SISPA primers were removed, and reads were quality filtered (> 21 average phred score, dereplication) using PRINSEQ v0.20.3 (Schmieder and Edwards 2011). Trimmed, high-quality sequences from each metagenome were assembled separately, as well as combined and treated as a single library, with the CLC Genomics Workbench version 5.0 (CLCBio, Cambridge, MA, USA). Assemblies were constructed with the default parameters using a non-global alignment.

**Metagenomic and genomic analyses**

Rarefaction curves were generated in Metavir (Roux et al., 2011) using the total unassembled, quality-filtered reads and a clusterization percentage of 75. Quality-filtered reads and assembled contigs from individual metagenomes and the pooled libraries were analysed with BLASTx (Altschul et al., 1990) searches in the NCBI nr (‘non-redundant’) database. Hits with an E value < 10^-5 were considered significant. This output was analysed using MEGAN v. 5.5.0 (Huson et al., 2011), which assigns each contig, and all of the reads that comprise it, to the top hit of the blast search (top 5%). Krona charts were made from the MEGAN analyses using Krona (Ondov et al., 2011). ORFs were identified using Geneious 6.1.8 and used to search the Conserved Domain Database (CDD) NCBI database (Marchler-Bauer et al., 2011). Primer pairs were designed using Primer3 (Koressaar and Remm, 2007; Untergasser et al., 2012) to target regions of both the structural and non-structural proteins of about 800–1000 bp, including the untranslated region between the two ORFs, to verify the consensus sequence of the assembled genomes. The quality-filtered reads from each metagenome were mapped back to the assembled genomes in order to determine the relative percentage of the metagenome that each assembled genome comprised. Mapping parameters were set as the default for CLC Genomics Workbench version 5.0.

In addition, primers were designed to target the 3' end of PAL128 in order to close the second ORF. Equal volumes of extracted viral RNA (2 μl) were pooled from each metagenome sample, and a 3 μl aliquot of the mixture was used for cDNA synthesis. RNA was converted to cDNA using only a poly-T tagged primer (FR40RV-T; Djikeng et al., 2008) under the same reaction conditions as described above, but without FR26RV-N primer. The resulting product was PCR amplified with a forward primer designed from the assembled contig and reverse primer FR20RV under the same PCR conditions described above, without a reconditioning step. The amplicon was concentrated and purified with a MinElute PCR Purification Kit (Qiagen) and run on a 0.5% agarose gel in 1 × TAE stained with SYBR Safe. The single band was visualized with a Safe Imager 2.0 Blue-Light Transilluminator (Life Technologies), excised and purified with a MinElute Gel Extraction Kit (Qiagen), and verified by sequencing. For that purpose, the amplicons were cloned into the pSMART HC Kan vector (Lucigen) according to the manufacturer’s instructions.

**Identification and analysis of RdRp phylotypes**

HMMER 3.0 (Eddy 1998) was used to identify RdRp-like sequences from quality-filtered assembled and unassembled reads using a profile hidden Markov model (profile HMM). A profile HMM was produced based on amino acid alignments of conserved RdRp regions 4 through 7 (Koonin et al., 2008) of representative picornavirids isolated and amplified or assembled environmental virus sequences (Culley et al., 2003; Culley and Steward, 2007; Culley et al., 2014). Sequences fitting the criteria of the model were retrieved from the libraries translated in all six frames. RdRp sequences that matched a known RdRp gene by BLAST with an E value ≤ 10^-5 were considered to be significant. Translated RdRp sequences that were more than 98% identical at the aa level were considered a single phylotype. Sequences were also clustered at a level of 75.7% aa identity to provide some sense of taxonomic diversity. This aa sequence identity threshold was chosen as an arbitrary, but probably conservative, approximation of species-level diversity based on the ICTV classifications of picornavirids. It represents the greatest distance between any two officially classified strains of picornavirids currently classified as the same species (Human Rhinovirus 2 and Human Rhinovirus 89). A maximum-likelihood tree of RdRp gene sequences was created with PHYML, WAG model, and 1000 bootstraps (Guindon et al., 2010) from protein sequences aligned with MAFFT (Katoh et al., 2005). The sequences in this tree can be found in Supporting Information Table S1.

**Accession numbers**

Metagenomic data referred to in this manuscript are available through the GenBank Sequence Read Archive (SRA). The raw reads are available under the accession numbers SRR1648122 (A_20101113), SRR1648127 (B_20101218), SRR1648128 (C_20110107), SRR1648129 (D_20110217), SRR1648130 (E_20110217), and SRR1648131 (F_20110303). The assembled genomes are also available through GenBank accession numbers KT727023 (PAL128), KT727024 (PAL156), KT727025 (PAL438), KT727026 (PAL473), and KT727027 (PAL_E4).
Acknowledgements

We thank the Palmer LTER group for assistance with sampling and logistics. In particular, Alice Alpert, Jennifer Brum, Kaycee Coleman, Carolina Funkey, Mike Garzio and Edgar Woznica provided invaluable assistance with collecting samples under often challenging conditions. We also thank Mahdi Belcaid and Guylaine Poisson at the Bioinformatics Laboratory at the University of Hawaii at Manoa for the use of their computer cluster for sequence assembly and BLAST searches. This work was supported by NSF grants to GFS and AIC (ANT 09-44851) and by NSF funding to the Center for Microbial Oceanography: Research and Education (EF 04-24599).

References


© 2016 The Authors. Environmental Microbiology published by Society for Applied Microbiology and John Wiley & Sons Ltd., Environmental Microbiology, 18, 3714–3727


**Supporting information**

Additional Supporting Information may be found in the online version of this article at the publisher’s web-site:

**Fig. S1.** Distribution of nucleic acids after separation of viral concentrates in CsCl buoyant density gradients. DNA and RNA per fraction for samples collected on November 13, 2010 (A), December 18, 2010 (B), January 7, 2011 (C), January 27, 2011 (D), February 17, 2011 (E), and March 3, 2011 (F). DNA and RNA in the grey highlighted fractions (14–18) were in the density range of 1.39–1.50 g ml$^{-1}$ and considered viral. The fractions used for metagenomic library preparation are indicated by an asterisk (*).

**Fig. S2.** A. Rarefaction curves of all six viromes collected throughout the season at Station B as determined by MetaVir (Roux et al., 2011). Libraries are designated by the color indicated in the legend that corresponds to each sample collection date (YYYYMMDD). In this analysis, sequence reads from each library were plotted as a function of the number of unique clusters (threshold = 75%) identified for the whole metagenome (left) and subsampled from each metagenome (right). Metavir requires at least 50,000 sequences to conduct a subsampled analysis, so E_20110217 was omitted from the second plot. B. Taxonomic classification of individual, quality-filtered reads (“Reads”, left) or assembled contigs (“Assembly”, right) from all six RNA virus metagenomes combined. Classifications were assigned based on comparison to the NCBI nr database using the BLASTx algorithm and a threshold $E$ value of $\leq 10^{-5}$.

**Table S1.** List of sequences and their accession numbers that were used to make the RdRp alignment and tree in Figure 4.