The ISME Journal (2013) 7, 2169–2177

www.nature.com/ismej

ORIGINAL ARTICLE

Previously unknown and highly divergent ssDNA viruses populate the oceans

Jessica M Labonté[1,5] and Curtis A Suttle[1,2,3,4]
[1]Department of Microbiology and Immunology, University of British Columbia, Vancouver, BC, Canada; [2]Department of Earth, Ocean and Atmospheric Sciences, University of British Columbia, Vancouver, BC, Canada; [3]Department of Botany, University of British Columbia, Vancouver, BC, Canada and [4]Canadian Institute of Advanced Research, University of British Columbia, Vancouver, BC, Canada

Single-stranded DNA (ssDNA) viruses are economically important pathogens of plants and animals, and are widespread in oceans; yet, the diversity and evolutionary relationships among marine ssDNA viruses remain largely unknown. Here we present the results from a metagenomic study of composite samples from temperate (Saanich Inlet, 11 samples; Strait of Georgia, 85 samples) and subtropical (46 samples, Gulf of Mexico) seawater. Most sequences (84%) had no evident similarity to sequenced viruses. In total, 608 putative complete genomes of ssDNA viruses were assembled, almost doubling the number of ssDNA viral genomes in databases. These comprised 129 genetically distinct groups, each represented by at least one complete genome that had no recognizable similarity to each other or to other virus sequences. Given that the seven recognized families of ssDNA viruses have considerable sequence homology within them, this suggests that many of these genetic groups may represent new viral families. Moreover, nearly 70% of the sequences were similar to one of these genomes, indicating that most of the sequences could be assigned to a genetically distinct group. Most sequences fell within 11 well-defined gene groups, each sharing a common gene. Some of these encoded putative replication and coat proteins that had similarity to sequences from viruses infecting eukaryotes, suggesting that these were likely from viruses infecting eukaryotic phytoplankton and zooplankton.
The ISME Journal (2013) **7,** 2169–2177; doi:10.1038/ismej.2013.110; published online 11 July 2013
Subject Category: Evolutionary genetics
Keywords: ssDNA viruses; microbial diversity; viral diversity

Introduction

Single-stranded DNA (ssDNA) viruses are major pathogens of plants and animals. There are seven families of ssDNA viruses that are recognized by the International Committee on Virus Taxonomy (King *et al.*, 2012) based on the host range and the type of ssDNA (segmented or not-segmented, positive-sense or negative-sense, circular or linear) composing the genome. Thus, there are two families of bacteriophages (*Inoviridae* and *Microviridae*) and five families of viruses infecting eukaryotes (*Nanoviridae* and *Geminiviridae* infecting plants; *Circoviridae*, *Parvoviridae* and *Anelloviridae* infecting animals). The genomes are small (between 1.4 and 8.5 kb), and can encode as few as two genes, a capsid and a replication initiator.

Viruses are the most abundant (Suttle, 2005) and genetically diverse (Breitbart *et al.*, 2002; Angly *et al.*, 2006) life forms in the biosphere; yet, little is known about the diversity of ssDNA viruses in natural systems, the evolutionary relationships among them and with characterized viruses, and the role they have in ecosystems. Sequences with similarity to ssDNA viruses have been found in metagenomic data from multiple environments (reviewed in Rosario and Breitbart, 2011; Rosario *et al.*, 2012). For example, sequences from the *Microviridae* and *Circoviridae* have been observed in marine environments (Angly *et al.*, 2006; Rosario *et al.*, 2009), freshwater (López-Bueno *et al.*, 2009) and modern stromatolites (Desnues *et al.*, 2008), and similar to those from the *Circoviridae*, *Geminiviridae*, *Nanoviridae* and *Parvoviridae* were observed in corals (Thurber *et al.*, 2008). However, the identification of ssDNA viruses relies on comparative analysis with sequences in databases that do not

Correspondence: CA Suttle, Department of Earth, Ocean and Atmospheric Sciences, University of British Columbia, 2207 Main Mall, Vancouver, BC, Canada V6T 1Z4.
E-mail: suttle@science.ubc.ca
[5]Current address: Single Cell Genomic Center, Bigelow Laboratory for Ocean Sciences, 60 Bigelow Drive, East Boothbay, ME 04544, USA.
Received 30 October 2012; revised 30 May 2013; accepted 4 June 2013; published online 11 July 2013

adequately reflect the diversity of ssDNA in nature; hence, the diversity of ssDNA viruses remains poorly characterized.

Here we present a comprehensive metagenomic study of marine ssDNA viruses and reveal their genetic diversity in samples from temperate and subtropical waters. Our results greatly extend the existing view of diversity in ssDNA viruses by uncovering new groups of ssDNA viruses that are divergent enough at the sequence level that they could represent new families. Our study demonstrates that the oceans harbor hundreds of previously unknown genetically distinct groups of ssDNA viruses that are likely significant pathogens of the phytoplankton and microzooplankton underlying marine food webs.

Materials and methods

Collection and preparation of samples

Samples were collected from five distinct geographic regions (Supplementary Table 1) as follows: the coastal waters of British Columbia Strait of Georgia (SOG, 82 samples), the Gulf of Mexico (GOM, 41 samples) and Saanich Inlet (SI, 11 samples). Water samples ($\sim$20 l for SI; $\sim$200 l for the others) were collected using GO-FLO or Niskin bottles mounted either on a rosette (SOG and GOM) or directly on a hydrographic wire (SI). For each sample, the viruses were concentrated $\sim$10 to 100-fold ($\sim$200 ml final volume) using ultrafiltration (Suttle *et al.*, 1991). Briefly, particulate matter was removed by pressure filtering (<17 kPa) the samples through 142-mm-diameter glass fiber (MFS GC50, nominal pore size 1.2 μm) and polyvinylidene difluoride (Millipore (Billerica, MA, USA) GVWP, pore size 0.22 μm) filters connected in series. The viral size fraction in the filtrate was then concentrated by ultrafiltration through a 30-kDa-molecular-weight cutoff cartridge (Amicon S1Y30, Millipore), and stored at 4°C in the dark until processed.

In order to integrate variation within a region, virus concentrates (VCs) collected from different locations and at different times within a geographic region were combined into a single mix (Supplementary Table 1). Two of these mixes (GOM, SOG) correspond to GOM and BBC, respectively, used in the study by Angly *et al.* (2006), in which marine viral ssDNA sequences were first observed. Two ml from each VC collected from SOG and neighboring inlets and bays were pooled into three mixes based on the year of collection (BC1—1999, 23 samples; BC3—2000, 26 samples; BC4—2004, 16 samples) and one mix based on salinity (BC2—low salinity, 19 samples). Similarly, we made four mixes from the GOM samples: Eastern GOM (8 samples), Northern GOM (6 samples), Western GOM (6 samples) and Texas Coast (13 samples). For SI, we used surface samples from the months of April 2007 and January, March, May, July, August and November 2008.

ssDNA preparation

Ten ml of each pooled mix (4 mixes from SOG, 4 from GOM and 7 mixes from SI) was filtered through a 0.22-μm pore-size syringe filter (polyvinylidene difluoride; Millipore) to remove any bacteria, and ssDNA was extracted using QIAprep Spin M13 kits (Qiagen, Missisauga, MA, USA), according to the manufacturer's protocol. Given the very low concentration of ssDNA (<50 ng per sample), we took advantage of the bias of multiple displacement amplification (MDA) for short segments of ssDNA (Lizardi *et al.*, 1998; Dean *et al.*, 2001), and used Repli-g Mini kits (Qiagen) to amplify DNA from 5 μl of each ssDNA preparation. MDA enhances chimera formation, creates random overamplification, and can be biased towards GC-rich regions (Rodrigue *et al.*, 2009), but for low concentrations of ssDNA it produces the greatest amplification with the lowest associated bias (Pinard *et al.*, 2006). The purified DNA was resuspended in 100 μl of RNAse- and DNAse-free water (Invitrogen, Carlsbad, CA, USA) and denaturation of dsDNA was reduced during MDA by adding the stop solution N1 immediately after the denaturation solution D1. As MDA creates high concentrations of ssDNA, a renaturation step was added by warming the purified DNA to 94 °C followed by slow cooling to 4 °C in steps of 1 °C every 30 s. The DNA was kept at 4 °C until further used. For SI, the samples were processed as described above, except that 10 ml of individual VCs was used instead of a VC mix.

Metagenome analysis, binning and assembly

Metagenomic libraries were constructed from ssDNA MDA products from SI, SOG and GOM. The purified MDA DNA was concentrated using a Millipore YM-30 Microcon centrifugal filter to a final volume of $\sim$50 μl; Sequencing of 3-5 μg of dsDNA was performed at Génome Québec, McGill University (SOG metagenome) and the Broad Institute at the Massachussetts Institute of Technology (GOM and SI metagenomes) following the Roche 454 GS FLX Titanium (454 Life Sciences, Branford, CT, USA) technology according to the manufacturer's instructions.

The sequences were quality and linker trimmed, and assembled into contiguous sequences (contigs) using the Newbler Assembler (Roche). tBLASTx with an e-value cutoff of 10^{-5} was used to compare the individual reads and assembled sequences with the NCBI database, as well as a subset of the database containing all ssDNA viral genomes. The scaffolds were examined with Consed (Gordon, 2001), while BLAST, genome circularization, annotations, MUSCLE alignments and phylogeny were done within Genious Pro v6.0 by Biomatters (http://www.geneious.com/). The metagenomic reads from Angly *et al.* (2006) were downloaded from CAMERA (Seshadri *et al.*, 2007). The composite genomes were assembled from contigs that had

identical sequences at the beginning and the end. Only circular genomes with an average of at least threefold coverage were kept for further analyses.

Feature frequency profiles and network representation
BLAST was used to compare the complete circular genomes with those from the NCBI database, including sequenced isolates infecting plants, animals and bacteria, as well as environmental genomes from other metagenomic libraries (10 circovirus-like genomes from an Antarctic lake (López-Bueno *et al.*, 2009), 9 from marine environments (Rosario *et al.*, 2009) and 11 cycloviruses from chimpanzee stools (Li *et al.*, 2010)). Only the NCBI genomes that were similar to at least one environmental genome were kept for further analysis. The feature frequency profile (FFP) analyses were performed as described by Sims *et al.* (2009) using the author's scripts. To avoid a bias introduced by sequences varying more than fourfold in length (longer sequences contain more polynucleotides), we kept only the genomes that were larger than 800 bp, and separated the longer sequences into 700–1100-bp fragments (Sims *et al.*, 2009). As the orientation of the genomes was not always known, calculations were performed on both strands. For example, the FFP of a 2400-nucleotide-long genome would be done on six fragments of 800 bp, including three forward and three reverse sequences, while the FFP of a 1000-bp genome would be done on two 1000-bp fragments, one forward and one reverse. Sequences with less than four homologs were removed to allow better visualization on a multi-dimensional-scaling plot. Sequences belonging to viruses in the *Microviridae* were removed from the FFP analysis, because these viruses have relatively larger genomes, and FPP is very sensitive to genome size. The feature frequency was calculated for 4-, 5-, 6-, 7-, 8- and 9-mers, but the 7-mers (heptamers) were better at discriminating known viral families (Supplementary Figures 1 and 2). The Jensen–Shannon divergence was then calculated and the results displayed in a PHYLIP-format matrix. Neighbor-joining analysis was performed with PHYLIP v3.6 (http://evolution.genetics.washington.edu/phylip.html) and a multidimensional-scaling analysis with R (Team RDC, 2011). The tBLASTx (e-value $>10^{-10}$) results comparing the ssDNA composite genomes with each other, with other environmental genomes and with the isolates were presented as a network using Cytoscape (Shannon *et al.*, 2003). For the network presentation, we linked up to five hits to each node.

Protein and phylogenetic analysis
For each group in the network, the open reading frames were identified and translated with GeneMark using the heuristic approach for viral sequences (Borodovsky *et al.*, 2003). The proteins were aligned using MUSCLE (Edgar, 2004). To limit sequencing errors and avoid potential chimeras from MDA amplification, only conserved full-length proteins were kept for further analysis. This conservative approach resulted in up to half of the sequences being removed. The alignments were submitted to HHpred to predict the putative function of the conserved proteins (Söding *et al.*, 2005). The replication proteins were trimmed to the conserved motifs, aligned using MAFFT with the E-INF-I algorithm (Katoh *et al.*, 2002), and the alignment was manually edited in Geneious. Maximum likelihood analyses were performed using phyML with the WAG model, a gamma distribution and bootstrapping with 100 replicates and 100 approximate likelihood ratio tests (Guindon *et al.*, 2010). Trees were viewed with FigTree (http://tree.bio.ed.ac.uk/software/figtree/).

Nucleotide sequence accession numbers
Raw reads and assembled contigs were submitted to CAMERA. The complete assembled genomes are available in GenBank (accession numbers JX904070–JX904677).

Results and discussion

The focus of this study was to reveal the diversity of ssDNA viruses in the sea and place them in an evolutionary context with extant and newly discovered groups of viruses. Remarkably, the 608 assembled genomes comprised 129 genetically distinct groups that had no recognizable similarity to each other or to other sequenced viruses, suggesting that many of these may represent new viral families. Moreover, nearly 70% of the ssDNA sequences had similarity to one of these genomes, indicating that most of the ssDNA sequences in these samples could be assigned to a genetically distinct group. The results leading to these findings are presented and discussed below.

In order to capture the diversity of three regions and allow comparisons among them, two composite samples were created from the temperate coastal Northeast Pacific Ocean (11 samples from SI and 85 samples from the SOG, respectively) and another composite sample was made from 46 samples collected from the subtropical GOM. From each composite sample, ssDNA was purified, then amplified and converted to double-stranded DNA.

Pyrosequencing produced 279 628 sequence reads (95 402 for SOG; 96 950 for SI; and 87 274 for GOM) of $\sim$500 bp in length, with 60–86% of the reads from each data set being assembled into a total of 4995 contiguous sequences (contigs) (1260 for SOG; 2399 for SI; and 1339 for GOM) ranging from 500 to 7246 bp. Comparison of the assembled sequences with ssDNA viral genomes in GenBank revealed

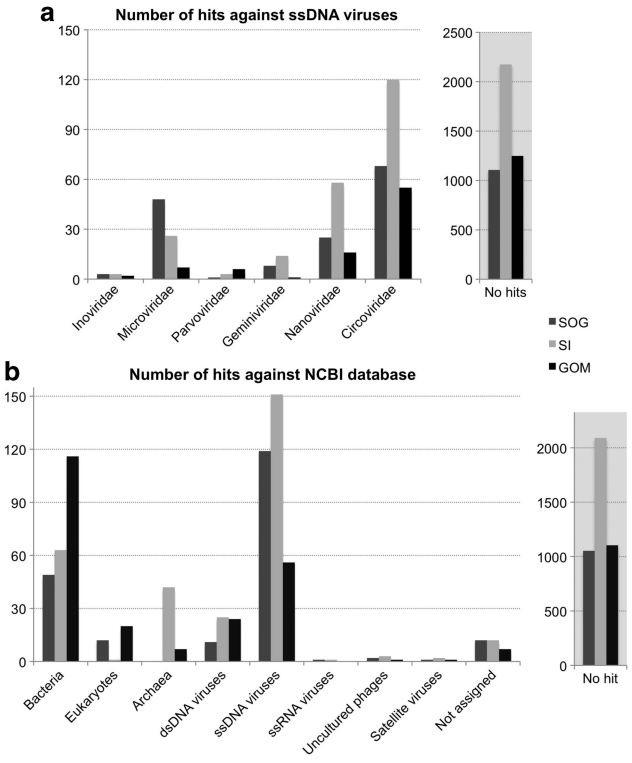

Figure 1 BLAST comparison of the contigs against (**a**) ssDNA viral families (e-value $<10^{-5}$) and (**b**) the NCBI database (e-value $<10^{-3}$) for the SOG(1260 contigs), SI(2399 contigs) and GOM (1336 contigs).

homologs to all of the ssDNA families except the *Anelloviridae* (Figure 1a). Contigs were homologous to viruses from the *Circoviridae* and *Nanoviridae* (4.9% and 2.0% of the total, respectively), with similarity being almost exclusively to the replication protein (described below). About 1.6% of the total sequences were homologous to the genus *Gokushovirus* from the *Microviridae*. Previous studies found that viruses in the *Circoviridae* and *Microviridae* (*Gokushovirus*) are widespread in marine and fresh waters (Angly *et al.*, 2006; Rosario *et al.*, 2009). As well, some aquatic circovirus-like sequences appear to encode a capsid protein known previously only in RNA viruses (Diemer and Stedman, 2012). RNA viral sequences were not observed in our data, and since the sequence coding the replication protein from this putative virus contained premature stop codons, similar sequences would have been excluded from our analyses. Sequences similar to those from the *Anelloviridae*, which occur in insects and mammals including California sea lions and Pacific harbor seals (Ng *et al.*, 2011), were not found in our samples, suggesting that they were rare or too divergent to be assigned to the family. Viruses from the *Parvoviridae* family have linear genomes and would not be enriched by MDA, and were excluded from this analysis. Only one contig had a low similarity to parvoviruses.

The contigs were then compared with the nr GenBank database to check for bacterial contamination. About 6% of contigs had homology to archaeal and bacterial genomes, with hits to hypothetical or phage-like proteins being most common. There were also a few homologs to sequences from eukaryotes (0.6%) (Figure 1b), although most were to a 'circovirus-like replication protein' in the draft genome of the anaerobic protozoan parasite *Giardia intestinalis* (Franzén *et al.*, 2009), suggesting that related species might be hosts for environmental circoviruses.

A total of 13% of the contigs could be assembled into 608 putative complete composite genomes of ssDNA viruses (128 from SOG, 307 from SI and 210 from GOM), almost doubling the number of sequenced ssDNA viruses in the NCBI database. To be a composite genome, the beginning and the end of the assembled contig had to be identical and the average coverage had to be at least threefold.

The composite genomes were compared with other ssDNA viruses and environmental sequences using the FFP, and by sequence similarity based on results from tBLASTx. The FFP uses the Jensen–Shannon divergence algorithm to compare the frequency of polynucleotides (here, heptamers) to generate a distance matrix, which is used to perform cluster (neighbor-joining) and multidimensional-scaling analyses (Figure 2). Only genomes that were similar to at least four other isolate or environmental genomes were used in the analysis to avoid cluttering the diagram with data corresponding to rare genomes. The genomes were divided into five marine FFP clusters. Except for Cluster 3, which overlapped with nanoviruses, the clusters were distinct from known families of ssDNA viruses. As shown by neighbor-joining and multidimensional-scaling analyses, the FFP discriminated established families of viruses, including nanoviruses, providing evidence that the FPP clusters are robust (Figure 2 and Supplementary Figure 2). The polynucleotide frequency does not necessarily represent gene conservation and evolution, but is indicative of host–virus co-evolution (Pride *et al.*, 2006), suggesting that viruses within a FFP cluster infect related hosts. Therefore, the overlap of Cluster 3 with nanoviruses may indicate that viruses in this cluster infect photosynthetic organisms.

In the second approach, sequence homology based on results from tBLASTx was shown as a network, where each node is a complete genome (composite or isolate) and each link represents a hit with an e-value $<10^{-10}$ (Figure 3); therefore, each cloud represents a group of sequences sharing a gene homolog. This resolved 129 genetically distinct groups represented by at least one complete genome (Figure 3), with related sequences being grouped based on sequence homology. With the exception of sequences that might be from multipartite genomes, these genetically distinct groups likely represent distant evolutionary lineages. While 94 of the composite genomes had no similarity to other

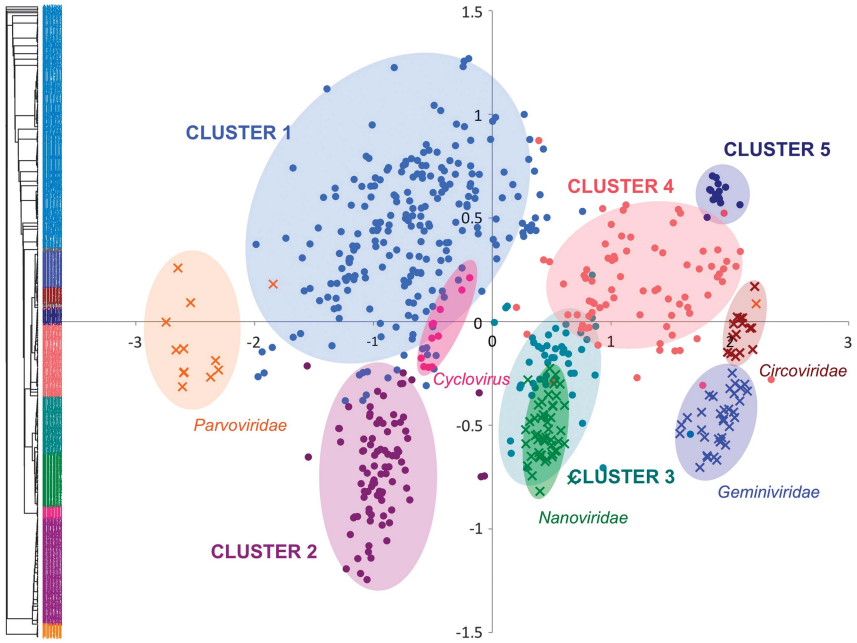

Figure 2 FFP analyses of ssDNA virus isolates from the NCBI database and genomes from this study. Neighbor-joining tree (left) and multidimensional scaling (right) (goodness of fit = 0.6495) of viral isolates (crosses) and composite genomes (dots) demonstrates that FPP of heptamers is able to resolve evolutionary relationships among ssDNA viruses. The shaded areas emphasize the established families of ssDNA viruses and the new evolutionary clusters identified in this study.

genomes, most sequences fell within 11 major coding DNA sequence (CDS) groups.

Each node in the network (Figure 3) was assigned to its FFP cluster (from Figure 2). Most of the complete genomes (84%) fell into 11 major CDS groups (Figure 3) comprising more than six genomes, with each member of a group sharing a conserved gene (Supplementary Table 2). Although the genome organization was usually conserved within a CDS group (Supplementary Table 2), the diversity was much larger than previously known. Only two of these clusters contain previously sequenced viruses. CDS group 6 contains viruses from the *Nanoviridae* and *Circoviridae* while CDS group 7 comprises viruses from the *Microviridae*. FFP Clusters 1 and 4 resolved into more than one CDS group. This may be because ssDNA genomes can be multipartite. For example, nanoviruses can have 6–11 circular ssDNA molecules of ~1 kb, each encoding a gene (Gronenborn, 2004), while begomoviruses (*Geminiviridae*) can be bipartite (King *et al.*, 2012). Consequently, viruses falling in the same FFP cluster but in different CDS groups may belong to the same viral family, or infect similar hosts. The similar number of sequences in CDS groups 1A (86 genomes) and 1B (73 genomes) suggests that these sequences come from multipartite viruses or that one is a satellite virus of the other. Furthermore, CDS groups 4A (13 genomes) and 4B (17 genomes) also have a similar number of sequences and come from the same FFP cluster.

Most groups were distributed similarly in temperate and subtropical waters, although sequences in CDS group 2 occurred more frequently in the subtropical GOM (1.8% of the contigs) than in the temperate SOG (0.7%) and SI (0.1%), whereas the opposite was true for *Microviridae* (3.7%, 1.2% and 0.6% for SOG, SI and GOM, respectively). This is consistent with some groups of viruses being widely distributed in nature (Short and Suttle, 2002; Breitbart and Rohwer, 2005; Short and Suttle, 2005; Labonté *et al.*, 2009), while others are more restricted in distribution (Short and Suttle, 2005; Tucker *et al.*, 2011).

Finally, each contig was assigned to a CDS group with a tBLASTx e-value $<10^{-10}$ (Figure 4). In each metagenomic data set the most frequently occurring sequences fell into CDS groups 1A and 1B, followed by the Circovirus-like group (Figure 4). Moreover, nearly 70% of the 4995 contigs had similarity to at least one composite genome (Figure 4); thereby, our analysis allowed for most contigs to be placed in a genomic context.

CDS group 6 was intriguing because it contained genomes from multiple FFP clusters (Figure 3). Genomes within CDS group 6 share a rolling-circle replication protein commonly found in circoviruses and nanoviruses. A similar replication protein is also found in geminiviruses (Gronenborn, 2004) and some plasmids (Gibbs *et al.*, 2006). The translated proteins contained all five conserved replicase motifs involved in rolling-circle replication (Supplementary Figure 3), including the motif involved in the initiation and termination of rolling-circle DNA replication (motif 2), a DNA-linking tyrosine (motif 3) and the Walker A motif, which is a

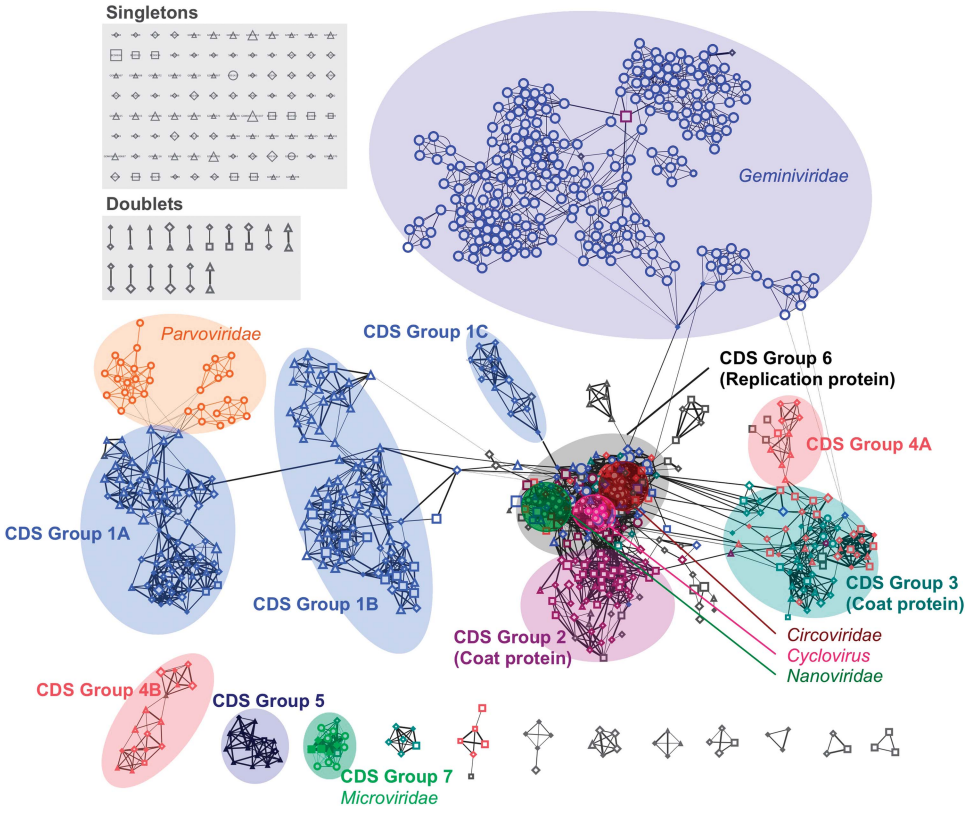

Figure 3 Network representation of the BLAST comparisons of the environmental genomes (i.e. genomes assembled from metagenomic data) with previously known ssDNA viruses (e-value $<10^{-5}$) and with other environmental metagenomes (e-value $<10^{-10}$). Each node represents a complete metagenome or genome (circle: isolate; triangle: GOM; diamond: SI; square: SOG) and each link represents a BLAST hit. The color of the outline of the node represents the viral family (blue: *Geminiviridae*; green: *Nanoviridae*; dark red: *Circoviridae*; orange: *Parvoviridae*; olive green: *Microviridae*) or the color of the cluster assigned in the FFP analysis from Figure 2 (blue: Cluster 1, dark blue: Cluster 3, aqua: Cluster 4 and purple: Cluster 5). The shaded areas highlight those genomes that have a conserved CDS in common. A solid colored node means that the genome contains the full-length conserved protein. The shaded gray boxes (upper left) encompass whole genomes with either no recognizable similarity to other genomes (singletons) or with similarity to one other genome (doublets).

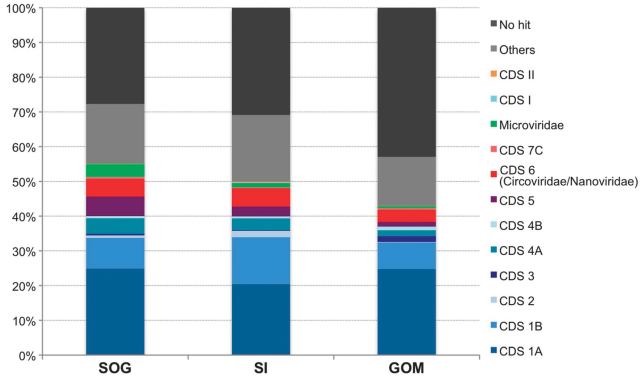

Figure 4 Relative percentage of contigs from each of the viral groups identified in this study for the SOG, SSI and GOM as determined by BLAST comparison (e-value $<10^{-10}$).

putative NTP-binding site (motif 4) (Ilyina and Koonin, 1992; Mankertz *et al.*, 1998). Replication is initiated by the recognition of a stem-loop structure at the origin of replication. In circoviruses, the loop contains a conserved motif of 9 bp (i.e. nonanucleotide) located between the 5′-ends of the two main inversely encoded open-reading frames (King *et al.*, 2012; Rosario *et al.*, 2012). Phylogenetic analysis of the replication protein sequences revealed at least 10 new clades of environmental ssDNA viruses that are distinct from terrestrial viruses (Figure 5 and Supplementary Figure 4). Interestingly, these new clades are also congruent with respect to genome organization, as similar replication proteins are found in genomes with the same genomic features (number of open-reading frames, orientation of open-reading frames and presence or absence of the typical nonanucleotide sequence (NANTATTAC) in the stem loop) (Supplementary Figure 5). Cycloviruses are usually found within the gut of animals (Li *et al.*, 2010), which is a different environment than marine. Therefore, they may infect very different hosts, which may provide an explanation as to why they are so different than the marine ssDNA viruses.

BLAST comparisons to the GenBank database of the conserved proteins in the other 10 remaining CDS groups did not reveal significant similarity. To identify the proteins, we used homology detection and structure prediction by HMM-HMM (hidden

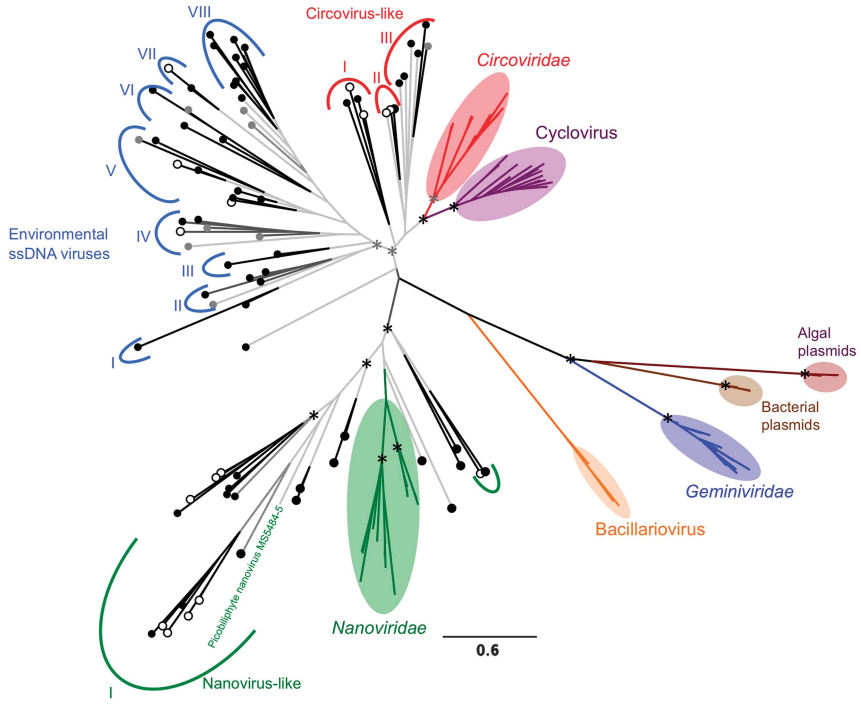

Figure 5 Unrooted phylogenetic analysis (maximum likelihood; model WAG; 100 bootstrap replicates) representing the genetic relatedness of the rolling-circle replication protein of nanoviruses (green), geminiviruses (blue), circoviruses (red), cycloviruses (purple) and the environmental sequences (black dots: this study, gray dots: other studies). The black, dark gray and light gray branches represent >90%, 75–89% and <75% bootstrap support, respectively. Black and gray asterisks at internal nodes represent at least 90% and 75% aLTR bootstrap support, respectively. Roman numerals represent new deeply branched phylogenetic groups of the rolling-circle replication protein from viruses that likely infect phytoplankton (green), zooplankton (red) or other protists (blue).

Markov model) comparison (HHpred). Homology to known protein structure was found for two more proteins. The conserved protein for CDS group 4A is similar to the coat protein of tobacco necrosis satellite virus 1 (e-value = 0.00041), while CDS group 3 has weak similarity to ryegrass mottle virus (e-value = 15). As these viruses infect plants, it suggests that viruses within these CDS groups infect phytoplankton.

One reason why BLAST analysis may have revealed so few similar sequences to those in our data set is because extant databases are overrepresented with data from ssDNA viruses infecting terrestrial plants and animals. As well, the very high mutation rates in ssDNA viruses can result in rapid sequence divergence. For example, mutation rates of begomoviruses have been reported to be as high as 10^{-3}–10^{-4} substitutions per site per year (Duffy *et al.*, 2008). Given the use of MDA, pyrosequencing and sometimes low coverage, it was not possible to evaluate the mutation rate and genetic variability within consensus genomes. High mutation rates can cause multiple substitutions that lead to proteins with similar functions, but extremely diverse sequences (Duffy and Holmes, 2008). An example of this is the conserved jelly-roll motif in capsid proteins in which there is no recognizable amino-acid homology, but it is argued that the proteins share a common evolutionary history (Bamford, 2002). Nonetheless, marine and terrestrial ssDNA viruses are distinct.

In order to identify ssDNA viruses from other marine metagenomic libraries, we compared our data with four DNA-virus metagenomic libraries constructed from the Arctic Ocean, Sargasso Sea, SOG and GOM, where the SOG and GOM samples were from the same composite samples used in this study (Angly *et al.*, 2006). About 5–15% of the sequences from Angly *et al.* (2006) that originated from British Columbia, the GOM and the Sargasso Sea were identified as ssDNA (Supplementary Figure 6), while ssDNA was not found in the Arctic data set. Finally, to compare ssDNA viruses from temperate with those from subtropical waters, BLAST comparisons were made among our ssDNA metagenomic data sets. About 50% of the ssDNA sequences from temperate waters and the GOM were homologous to each other, whereas data sets from the SOG and SI were 72–82% similar to each other (Supplementary Figure 7). This result, as well as the observation that ~70% of the sequences in these data sets had homologs within the assembled genomes (Figure 4), indicates that we have representative genomes for most of the circular ssDNA viruses in these samples. Moreover, as these data are from hundreds of pooled samples encompassing temperate and subtropical environments, it suggests that we may have sequenced representatives of most

of the ssDNA viruses in the surface waters of the SOG and the GOM.

The direct purification and sequencing of ssDNA from temperate and subtropical marine biomes yielded 608 complete genomes comprising 129 genetically distinct new groups that have little or no recognizable similarity at the sequence level. Given that viruses within extant families have significant genetic similarity, it suggests that many of these new sequence groups belong to previously unknown families of viruses. The high evolutionary divergence of these sequences also suggests that they belong to viruses that infect a wide diversity of organisms. Although some of these sequence groups likely stem from viruses that infect bacteria, the few sequence groups that could be associated with extant viruses belonged to families of viruses that infect eukaryotes, suggesting that some of these new groups comprise viruses that are infecting the eukaryotic phytoplankton and zooplankton that underlie marine food webs.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgements

We thank SJ Hallam and members from his laboratory for providing filtered water from Saanich Inlet, and many members of the Suttle laboratory for their collecting and processing of the samples that made this study possible. Special thanks to CG Howes for bioinformatics help, and GE Sims for the FFP perl script. This research was supported by the Natural Science and Engineering Research Council of Canada (NSERC) through a postgraduate scholarship (JML) and Discovery grants (CAS). Sample collection was facilitated through Ship-time grants from NSERC that supported sample collections from theStrait of Georgia (CAS) and the Saanich Inlet time series (PD Tortell and SJ Hallam), the US National Science Foundation (Gulf of Mexico), and through the Canadian Arctic Shelf Exchange Study (NSERC) and the Japan/Canada Western Arctic Climate Study. Access to sequencing was funded by the Gordon and Betty Moore Foundation through a grant to the Broad Institute, and by NSERC and the Tula Foundation using facilities at the McGill University and Génome Québec Innovation Centre.

References

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

Bamford DH. (2002). Evolution of viral structure. *Theor Popul Biol* **61**: 461–470.

Borodovsky M, Mills R, Besemer J. (2003). Prokaryotic gene prediction using GeneMark and GeneMark.hmm. *Curr Protoc Bioinformatics* **Chapter 4**: 4.5.1–4.5.16.

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

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

Dean F, Nelson J, Giesler T, Lasken R. (2001). Rapid amplification of plasmid and phage DNA using phi29 DNA polymerase and multiply-primed rolling circle amplification. *Genome Res* **11**: 1095–1099.

Desnues C, Rodriguez-Brito B, Rayhawk S, Kelley S, Tran T, Haynes M *et al.* (2008). Biodiversity and biogeography of phages in modern stromatolites and thrombolites. *Nature* **452**: 340–343.

Diemer GS, Stedman KM. (2012). A novel virus genome discovered in an extreme environment suggests recombination between unrelated groups of RNA and DNA viruses. *Biol Direct* **7**: 13.

Duffy S, Holmes EC. (2008). Phylogenetic evidence for rapid rates of molecular evolution in the single-stranded DNA begomovirus tomato yellow leaf curl virus. *J Virol* **82**: 957–965.

Duffy S, Shackelton LA, Holmes EC. (2008). Rates of evolutionary change in viruses: patterns and determinants. *Nat Rev Genet* **9**: 267–276.

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

Franzén O, Jerlström-Hultqvist J, Castro E, Sherwood E, Ankarklev J, Reiner DS *et al.* (2009). Draft genome sequencing of *Giardia intestinalis* assemblage B isolate GS: is human giardiasis caused by two different species? *PLoS Pathog* **5**: e1000560.

Gibbs MJ, Smeianov VV, Steele JL, Upcroft P, Efimov Ba. (2006). Two families of Rep-like genes that probably originated by interspecies recombination are represented in viral, plasmid, bacterial, and parasitic protozoan genomes. *Mol Biol Evol* **23**: 1097–1100.

Gordon D, Desmarais C, Green P. (2001). Automated finishing with autofinish. *Genome Res* **11**: 614–625.

Gronenborn B. (2004). Nanoviruses: genome organisation and protein function. *Vet Microbiol* **98**: 103–109.

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

Ilyina TV, Koonin EV. (1992). Conserved sequence motifs in the initiator proteins for rolling circle DNA replication encoded by diverse replicons from eubacteria, eucaryotes and archaebacteria. *Nucleic Acids Res* **20**: 3279–3285.

Katoh K, Misawa K, Kuma K, Miyata T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* **30**: 3059–3066.

King A, Adams M, Carstens E, Lefkowitz E. (2012). *Virus Taxonomy: Ninth Report of the International Committee on Taxonomy of Viruses*, 2nd edn. Elsevier Academic Press: San Diego, CAa, USA.

Labonté JM, Reid KE, Suttle CA. (2009). Phylogenetic analysis indicates evolutionary diversity and environmental segregation of marine podovirus DNA polymerase gene sequences. *Appl Environ Microb* **75**: 3634–3640.

Li L, Kapoor A, Slikas B, Bamidele OS, Wang C, Shaukat S *et al.* (2010). Multiple diverse circoviruses infect farm

animals and are commonly found in human and chimpanzee feces. *J Virol* **84**: 1674–1682.
Lizardi PM, Huang X, Zhu Z, Bray-Ward P, Thomas DC, Ward DC. (1998). Mutation detection and single-molecule counting using isothermal rolling-circle amplification. *Nat Genet* **19**: 225–232.
López-Bueno A, Tamames J, Velázquez D, Moya A, Quesada A, Alcamí A. (2009). High diversity of the viral community from an Antarctic lake. *Science* **326**: 858–861.
Mankertz A, Mankertz J, Wolf K, Buhk HJ. (1998). Identification of a protein essential for replication of porcine circovirus. *J Gen Virol* **79**: 381–384.
Ng TFF, Wheeler E, Greig D, Waltzek TB, Gulland F, Breitbart M. (2011). Metagenomic identification of a novel anellovirus in Pacific harbor seal (*Phoca vitulina richardsii*) lung samples and its detection in samples from multiple years. *J Gen Virol* **92**: 1318–1323.
Pinard R, De Winter A, Sarkis GJ, Gerstein MB, Tartaro KR, Plant RN *et al.* (2006). Assessment of whole genome amplification-induced bias through high-throughput, massively parallel whole genome sequencing. *BMC Genomics* **7**: 216.
Pride DT, Wassenaar TM, Ghose C, Blaser MJ. (2006). Evidence of host-virus co-evolution in tetranucleotide usage patterns of bacteriophages and eukaryotic viruses. *BMC Genomics* **7**: 8.
Rodrigue S, Malmstrom RR, Berlin AM, Birren BW, Henn MR, Chisholm SW. (2009). Whole genome amplification and de novo assembly of single bacterial cells. *PloS One* **4**: e6864.
Rosario K, Breitbart M. (2011). Exploring the viral world through metagenomics. *Curr Opin Virol* **1**: 289–297.
Rosario K, Duffy S, Breitbart M. (2012). A field guide to eukaryotic circular single-stranded DNA viruses: insights gained from metagenomics. *Arch Virol* **157**: 1851–1871.
Rosario K, Duffy S, Breitbart M. (2009). Diverse circovirus-like genome architectures revealed by environmental metagenomics. *J Gen Virol* **90**: 2418–2424.
Seshadri R, Kravitz SA, Smarr L, Gilna P, Frazier M. (2007). CAMERA: a community resource for metagenomics. *PLoS Biol* **5**: e75.
Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D *et al.* (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**: 2498–2504.
Short CM, Suttle CA. (2005). Nearly identical bacteriophage structural gene sequences are widely distributed in both marine and freshwater environments. *Appl Environ Microb* **71**: 480–486.
Short SM, Suttle CA. (2002). Sequence analysis of marine virus communities reveals that groups of related algal viruses are widely distributed in nature. *Appl Environ Microb* **68**: 1290–1296.
Sims GE, Jun S-R, Wu GA, Kim S-H. (2009). Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. *Proc Natl Acad Sci USA* **106**: 2677–2682.
Söding J, Biegert A, Lupas AN. (2005). The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res* **33**: W244–W248.
Suttle CA. (2005). Viruses in the sea. *Nature* **437**: 356–361.
Suttle CA, Chan AM, Cottrell MT. (1991). Use of ultrafiltration to isolate viruses from seawater which are pathogens of marine phytoplankton. *Appl Environ Microb* **57**: 721–726.
Team RDC (2011). *R: A Language and Environment for Statistical Computing*. Team RDC: Vienna, Austria.
Thurber RLV, Barott KL, Hall D, Liu H, Rodriguez-Mueller B, Desnues C *et al.* (2008). Metagenomic analysis indicates that stressors induce production of herpes-like viruses in the coral Porites compressa. *Proc Natl Acad Sci USA* **105**: 18413–18418.
Tucker KP, Parsons R, Symonds EM, Breitbart M. (2011). Diversity and distribution of single-stranded DNA phages in the North Atlantic Ocean. *ISME J* **5**: 822–830.

Supplementary Information accompanies this paper on The ISME Journal website (http://www.nature.com/ismej)