

2006

The Marine Viromes of Four Oceanic Regions

Florent Angly
San Diego State University

Ben Felts
San Diego State University

Mya Breitbart
University of South Florida, mya@usf.edu

Peter Salamon
San Diego State University

Robert Edwards
San Diego State University

See next page for additional authors

Follow this and additional works at: https://scholarcommons.usf.edu/msc_facpub

 Part of the [Life Sciences Commons](#)

Scholar Commons Citation

Angly, Florent; Felts, Ben; Breitbart, Mya; Salamon, Peter; Edwards, Robert; Carlson, Craig; Chan, Amy; Haynes, Matthew; Kelley, Scott; Liu, Hong; Mahaffy, Joseph; Mueller, Jennifer; Nulton, James; Olson, Robert; Parsons, Rachel; Rayhawk, Steve; Suttle, Curtis; and Rohwer, Forest, "The Marine Viromes of Four Oceanic Regions" (2006). *Marine Science Faculty Publications*. 777.

https://scholarcommons.usf.edu/msc_facpub/777

This Article is brought to you for free and open access by the College of Marine Science at Scholar Commons. It has been accepted for inclusion in Marine Science Faculty Publications by an authorized administrator of Scholar Commons. For more information, please contact scholarcommons@usf.edu.

Authors

Florent Angly, Ben Felts, Mya Breitbart, Peter Salamon, Robert Edwards, Craig Carlson, Amy Chan, Matthew Haynes, Scott Kelley, Hong Liu, Joseph Mahaffy, Jennifer Mueller, James Nulton, Robert Olson, Rachel Parsons, Steve Rayhawk, Curtis Suttle, and Forest Rohwer

The Marine Viromes of Four Oceanic Regions

Florent E. Angly^{1,2}, Ben Felts^{2,3}, Mya Breitbart^{1*}, Peter Salamon^{2,3}, Robert A. Edwards^{1,2,4,5}, Craig Carlson⁶, Amy M. Chan⁷, Matthew Haynes¹, Scott Kelley^{1,4}, Hong Liu¹, Joseph M. Mahaffy^{2,3}, Jennifer E. Mueller¹, Jim Nulton^{2,3}, Robert Olson⁸, Rachel Parsons⁹, Steve Rayhawk^{1,2}, Curtis A. Suttle^{7,10,11}, Forest Rohwer^{1,4*}

1 Department of Biology, San Diego State University, San Diego, California, United States of America, **2** Computational Science Research Center, San Diego State University, San Diego, California, United States of America, **3** Department of Mathematics, San Diego State University, San Diego, California, United States of America, **4** Center for Microbial Sciences, San Diego State University, San Diego, California, United States of America, **5** Fellowship for Interpretation of Genomes, Burr Ridge, Illinois, United States of America, **6** University of California Santa Barbara, Santa Barbara, California, United States of America, **7** Department of Earth and Ocean Sciences, University of British Columbia, Vancouver, British Columbia, Canada, **8** Argonne National Laboratory, Argonne, Illinois, United States of America, **9** Bermuda Biological Station for Research, St. George's, Bermuda, **10** Department of Microbiology and Immunology, University of British Columbia, Vancouver, British Columbia, Canada, **11** Department of Botany, University of British Columbia, Vancouver, British Columbia, Canada

Viruses are the most common biological entities in the marine environment. There has not been a global survey of these viruses, and consequently, it is not known what types of viruses are in Earth's oceans or how they are distributed. Metagenomic analyses of 184 viral assemblages collected over a decade and representing 68 sites in four major oceanic regions showed that most of the viral sequences were not similar to those in the current databases. There was a distinct "marine-ness" quality to the viral assemblages. Global diversity was very high, presumably several hundred thousand of species, and regional richness varied on a North-South latitudinal gradient. The marine regions had different assemblages of viruses. Cyanophages and a newly discovered clade of single-stranded DNA phages dominated the Sargasso Sea sample, whereas prophage-like sequences were most common in the Arctic. However most viral species were found to be widespread. With a majority of shared species between oceanic regions, most of the differences between viral assemblages seemed to be explained by variation in the occurrence of the most common viral species and not by exclusion of different viral genomes. These results support the idea that viruses are widely dispersed and that local environmental conditions enrich for certain viral types through selective pressure.

Citation: Angly F, Felts B, Breitbart M, Salamon P, Edwards R, et al. (2006) The marine viromes of four oceanic regions. *PLoS Biol* 4(11): e368. DOI: 10.1371/journal.pbio.0040368

Introduction

Most marine viruses are phages (bacteriophages) that kill the heterotrophic and autotrophic microbes (both Bacteria and presumably Archaea) that dominate the world's oceans [1]. Phages and the other major microbial predator guild, nanoflagellates, control the numbers of marine microbes to a concentration of about $\sim 5 \times 10^5$ cells per ml of surface seawater [2,3].

Phages affect microbial evolution by inserting themselves into genomes as prophages. Prophages often account for most of the difference between strains of the same microbial species [4], and they can dramatically change the phenotype of the hosts via lysogenic conversion. For example, many nonpathogens and pathogens only differ by prophages that encode exotoxin genes [5]. Phages also affect microbial evolution by moving genes from host to host. It has been hypothesized that most of the orphan open reading frames (ORFans) in microbial genomes are actually of phage origin [6]. Phages may also affect microbial evolution by killing specific microbes. Various Lotka-Volterra models, called "kill-the-winner," predict that as one microbial strain becomes dominant, its viral predator kills it and leaves open a niche that can be used by a related strain that is resistant to the phage [7,8]. This model may explain the enormous microdiversity observed in microbial communities [9].

The advent of whole-community genome sequencing (i.e., metagenomics) is rapidly changing the way viral and microbial diversity are assayed. Using this approach, it is possible to rapidly characterize the metabolic diversity and community

structure of any microbial ecosystem [10–19]. We studied the marine viral metagenome (virome) of four oceanic regions. The viromes were obtained by pyrosequencing uncultured viral assemblages that were integrated over 4,600 km in distance, 3,000 m in depth, and over a decade in time in order to characterize them and identify patterns of viral distribution and diversity.

Materials and Methods

Samples and Sequencing

Samples were collected from four oceanic regions (Figure 1). Briefly, the viral samples were concentrated on tangential flow filters (30–100-kD cutoff), distributed into 50-ml tubes and stored at 4 °C in the dark. A single sample was collected from the Sargasso Sea (labeled SAR) on 30 June 2005.

Academic Editor: Nancy A. Moran, University of Arizona, United States of America

Received April 8, 2006; **Accepted** September 5, 2006; **Published** November 7, 2006

DOI: 10.1371/journal.pbio.0040368

Copyright: © 2006 Angly et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abbreviations: bp, base pair; PTP, permutation tail probability; ssDNA, single-stranded DNA

* To whom correspondence should be addressed. E-mail: forest@sunstroke.sdsu.edu

† Current address: University of South Florida, St. Petersburg, Florida, United States of America

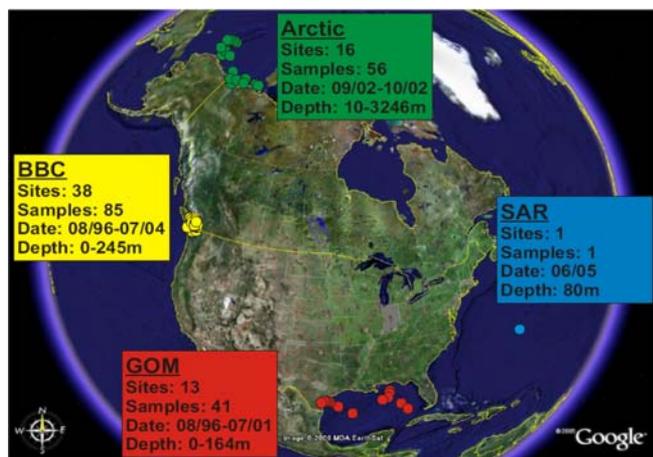


Figure 1. Sampling Sites

The circles represent the sampling locations in the Sargasso Sea (SAR), Gulf of Mexico (GOM), British Columbia (BBC), and the Arctic Ocean. The number of samples taken at each location and combined for sequencing, as well as the date and depth range, are shown in the boxes. DOI: 10.1371/journal.pbio.0040368.g001

Chloroform was added to this sample to stop microbial growth. Integrative samples, representing multiple sites and times, were assembled from the Gulf of Mexico (labeled GOM; 13 sites; 42 individual samples), the British Columbia coastal waters (labeled BBC; 38 sites; 85 individual samples), and the Arctic Ocean (labeled Arctic; 16 sites; 56 individual samples). These samples represent the combined viral assemblages of four oceanic regions over approximately one decade (sample details are described in Protocol S1).

Viral particles were purified using a combination of filtration and density-dependent centrifugation ([10]; <http://scums.sdsu.edu/isolation.html>, accessed 15 September 2006). The cesium chloride gradient was designed to recover virions with densities from 1.35 g ml⁻¹ to 1.5 g ml⁻¹. Viral DNA was isolated by a formamide/CTAB extraction [20], and the resulting DNA was amplified with Genomiphi and sequenced using pyrophosphate sequencing (454 Life Sciences, Branford, Connecticut, United States) [21] (see Protocol S1 for details on the technology). Each Genomiphi reaction started with 100–150 ng of DNA, above the 10 ng recommended by the manufacturer. A total of 181,044,179 base pairs (bp) of DNA sequence data was generated from the four libraries (SAR, 42 Mbp; GOM, 27 Mbp; BBC, 43 Mbp; and Arctic, 69 Mbp). The difference in library size was due to differences in number of successful reads during the pyrosequencing. The 1,768,297 sequences had an average length of 102 bp. The GOM, BBC, Arctic, and SAR metagenomes are deposited on the SDSU Center for Universal Microbe Sequencing website at (<http://scums.sdsu.edu/phage/Oceans>, accessed 15 September 2006).

Bioinformatics

The metagenome sequences from each of the libraries were compared to the SEED nonredundant database and environmental database using BLASTX [22]. The SEED includes the GenBank database supplemented with other complete and draft genome sequences. The environmental database consists of the microbial assemblages from the Iron Mountain acid mine drainage [16], Sargasso Sea [17], whale fall [18], and

Minnesota farm soil [18]. All large-scale computational analyses were performed on the Terraport and National Microbial Pathogen Data Resource cluster at Argonne National Laboratory. Individual analyses were performed on a 12-node Orion desktop cluster (Orion, Santa Clara, California, United States).

These comparisons were supplemented with more extensive TBLASTN and with TBLASTX comparisons [22] of either selected portions of the data against the complete non-redundant database or the whole library compared to boutique databases. The same cutoff *E* value was always used for the same database and BLAST search method. In addition, the sequences were compared to the phage and prophage sequences from 510 genomes of the phage genome database (RA Edwards, unpublished data). A FASTA file of these genomes is at <http://scums.sdsu.edu/phage/Oceans>.

Taxonomic Composition of the Metagenomes

In an approach similar to previous work [10–12], the best similarity for each metagenomic sequence was automatically parsed and assigned as “known” if there was a significant similarity ($E \leq 10^{-5}$) to a sequence from the nonredundant nucleotide database, else “environmental” for a significant similarity to any environmental database sequence, and else “unknown” (if there was no significant similarity to any database). The number of similarities in each group was then counted (Figure 2A). These numbers were also averaged for the four samples. In a second step, the sequences from the “known” group were classified as viral, bacterial, archaeal, or eukaryotic based on their highest similarity (Figure 2B). To assess the contribution of the prophages (often similar to bacterial sequences), TBLASTX was used to compare the sequences against the complete phage genome sequences. Any significant similarity in the previous four taxonomic groups that was also similar to a prophage sequence was assigned to the prophage group instead. The prophage sequences for these analyses were extracted from complete microbial genomes. A complete list is available at the supporting website (<http://scums.sdsu.edu/phage/Oceans>). The average of these numbers for the four samples was also calculated.

Assembly and Verification of Single-Stranded DNA, the chp1-Like Microphage from the Sargasso Sea

The single-stranded DNA (ssDNA) chp1-like microphage was partially assembled from all of the sequences that had significant TBLASTX similarities ($E \leq 10^{-5}$). The assembly parameters were a minimal match percentage of 85% and a 20-bp minimum overlap using Sequencher 4.0 (Gene Codes, Ann Arbor, Michigan, United States). These sequences alone did not result in the assembly of a complete genome due to areas with low similarity to known chp1-like microphage. To complete the assembly, batches of sequences from the Sargasso Sea sample were added to these assemblies until complete coverage was obtained (the consensus sequence is in Protocol S1). The PCR primers SARssDNAF (5′ TGC GGA GAA TAT GGT GAT GA 3′), SARssDNAR1 (5′ CGG TTA TTA CGC CTG TCG TT 3′), and SARssDNAR2 (5′ CCA TGG TAG GGC AGA GGT AA 3′) were designed based on the consensus sequence. A PCR was run against the original Sargasso Sea sample DNA. The reaction mixture (50 μl total volume) contained target DNA, 1 mM of each primer, and 1X FidelityTaq master mix (USB, Cleveland,

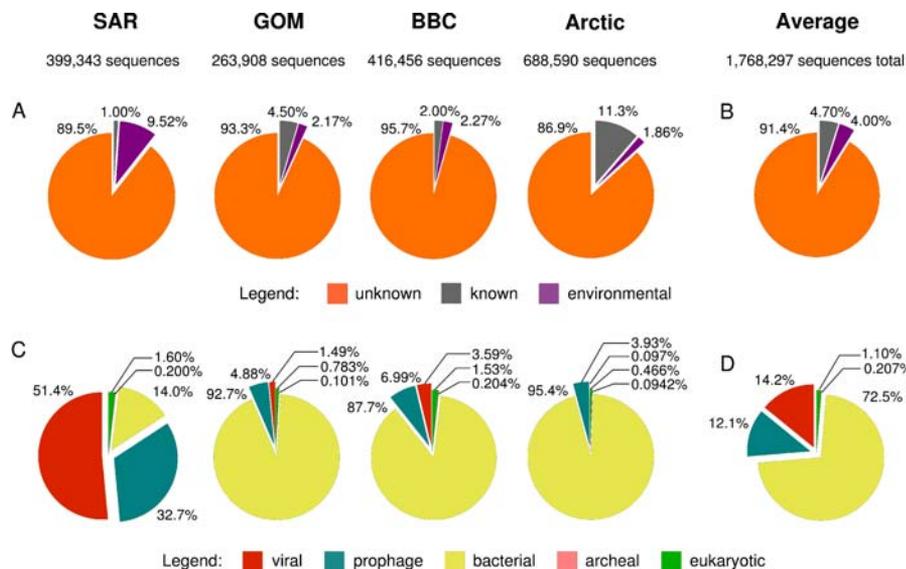


Figure 2. Composition of the Assemblage Genome Sequences as Determined by Similarity to Known DNA and Protein Sequences (A) The percent of “known” sequences compared to the SEED and environmental databases. A sequence was considered “known” if it had a significant similarity ($E < 10^{-5}$) to the SEED, else “environmental” if it had a similarity to any environmental database, and else “unknown”. (B) Breakdown of the “known” sequences into viral (both eukaryotic and bacteriophages), prophage, Bacteria, Archaea, or Eukarya. DOI: 10.1371/journal.pbio.0040368.g002

Ohio, United States). The thermocycler conditions were: 5 min at 94 °C; 30 cycles of 1 min at 94 °C, 1 min at 65 °C – 0.5 °C per cycle, 3 min at 72 °C; and 10 min 72 °C. Positive PCR products were sequenced for verification of sequence length, and identity was confirmed using TBLASTX.

Permutation Tail Probability Tests of Phylogenetic Similarity between Phage Assemblages

Because the sequences did not originate from a single genetic locus, the evolutionary relationships could not be determined by using standard alignment-based phylogenetic analyses. To determine phylogeny, the sequences were first mapped to the Phage Proteomic Tree based on their best TBLASTX similarity. The version of the Phage Proteomic Tree used here contained 510 complete phage genomes (<http://scums.sdsu.edu/phage/Oceans>) and was constructed as described previously [23]. Permutation tail probability (PTP) was then used to infer phylogenetic similarity among the phage assemblages. The PTP test uses phylogenetic parsimony to determine whether a given characteristic correlates with phylogeny [24]. Briefly, if a sequence had a best similarity to a phage genome on the Phage Proteomic Tree, it was scored on a tree using Phylogenetic Analysis Using Parsimony software (PAUP) [25]. The number of steps that would be required to produce a tree from one sample to another was then determined. To assign significance, this value was compared to a distribution produced by randomizing the input tree 10,000 times.

Genetic Isolation by Distance of the Phage Assemblages

Isolation by Distance Web Service (IBDWS) ([26]; <http://biome.sdsu.edu/ibdws>) was used to test for a correlation between the geographic distance between two samples and the genetic divergence between viral assemblages. This online software uses Mantel tests to determine whether marine phages in closer physical proximity have greater genetic similarity (as measured by Φ_{ST}) than those separated by large

geographic distances. For these tests, the current datasets were combined with data from the California coast [10]. The Arlequin program [27] was used to calculate Φ_{ST} . The Φ_{ST} statistic compares the phylogenetic diversity within each assemblage to the total phylogenetic diversity of the combined assemblages using the equation:

$$\Phi_{ST} = (\theta_T - \theta_W) / \theta_T \quad (1)$$

where θ_T is the total phylogenetic diversity of two assemblages and θ_W is the phylogenetic diversity within each assemblage or population. A Φ_{ST} value close to zero means there is complete overlap in the phylogenetic diversity, whereas values greater than zero indicate increasing levels of phylogenetic differentiation up to a value of 1, indicating complete differentiation.

Assembly and Mathematical Modeling of Viral Assemblage Diversity

To estimate viral diversity, sets of 10,000 random sequences from each oceanic region were assembled using TIGR Assembler [28] with a minimum overlap length of 35 bp, a minimal match percentage of 98% and no alignment error in 32 bp to identify overlapping sequences (contigs) [10]. The Perl script used to automate this task is available at <http://scums.sdsu.edu/phage/Oceans>. Average contig spectra were calculated (Figure S3) over ten repetitions, and the maximum likelihood assemblage structure of the marine viral assemblages was determined using mathematical rank-abundance models in PHAge Communities from Contig Spectra (PHACCS) ([29]; <http://biome.sdsu.edu/phaccs>). Random subsamples of the metagenomes were used instead of the totality of the whole metagenomes, because PHACCS analyses are more robust at low coverage [10,11,29]. The diversity estimates for the best-fitting assemblage model were used for each oceanic region. Detailed graphical explanations of these procedures are given in Protocol S1.

Table 1. Number of Similarities to Phage Genomes and Groups of Interest in the Four Metagenomes

Group of Interest	Phage Species	Marine Region			
		SAR	GOM	BBC	Arctic
Cyanophage	<i>Prochloro. marinus</i> ϕ P-SSM2	4661 ^a	589 ^a	1190 ^a	148 ^a
	<i>Prochloro. marinus</i> ϕ P-SSP7	4493 ^a	81 ^a	86	16
	<i>Prochloro. marinus</i> ϕ P-SSM4	1759 ^a	263 ^a	587 ^a	51
	<i>Synechococcus</i> ϕ S-PM2	1107 ^a	196 ^a	474 ^a	54
Prophage	<i>Br. melitensis</i> 16M ϕ Bruc1 pro- ϕ	12	115 ^a	92	700 ^a
	<i>Yersinia pestis</i> ϕ Yers2 pro- ϕ	12	60 ^a	34	386 ^a
	<i>Escherichia coli</i> ϕ CP4-6 pro- ϕ	4	52	24	364 ^a
	<i>Agro. tumefaciens</i> ϕ Tum2 pro- ϕ	14	43	55	281 ^a
	<i>Escherichia coli</i> ϕ CP037-7 pro- ϕ	6	40	11	240 ^a
	<i>Xy. fastidiosa</i> ϕ Xpd5 pro- ϕ	34	36	23	187 ^a
	<i>Escherichia coli</i> ϕ CP037-4 pro- ϕ	3	29	11	146 ^a
	<i>Mesorhizobium loti</i> ϕ Meso1 pro- ϕ	32	35	176 ^a	56
	<i>Pseudo. putida</i> ϕ PP03 pro- ϕ	397	57 ^a	96	1
	<i>Bd. bacteriovorus</i> ϕ MH2K	1835 ^a	20	115	0
chp1-like microphage (ssDNA)	<i>Chlamydia</i> ϕ 4	1757 ^a	5	119	0
	<i>Chlamydia</i> ϕ 3	1572 ^a	9	119	0
	<i>Chlamydia psittaci</i> ϕ 2	568 ^a	2	29	0
	<i>Chlamydia psittaci</i> ϕ chp1	519 ^a	16	60	0
	<i>Chlamydia</i> ϕ CPAR39 pro- ϕ	1548 ^a	14	112	0
	Miscellaneous	<i>Salmonella</i> ϕ epsilon15	56	41	172 ^a
<i>Burkholderia thailandensis</i> ϕ E125		7	29	29	111 ^a
<i>Roseobacteria</i> SIO67 ϕ SIO1		360	409 ^a	465 ^a	36
<i>Rhodothermus marinus</i> ϕ RM 378		301	93 ^a	206 ^a	20
α -proteobacteria ϕ JL001		333	45	197 ^a	55
<i>Bordetella</i> ϕ BIP-1		128	62 ^a	167 ^a	63
<i>Pseudo. aeruginosa</i> ϕ PaP3		123	55	161 ^a	10

^aThe ten most abundant similarities are noted for each sample.

Prochloro., *Prochlorococcus*; Br., *Brucella*; Agro., *Agrobacterium*; Pseudo., *Pseudomonas*; Bd., *Bdellovibrio*.

DOI: 10.1371/journal.pbio.0040368.t001

To analyze the degree of similarity between the viral assemblages, the amount of overlap between the assemblages was determined by assembling a mixed sample of 10,000 fragments obtained by pooling 2,500 fragments from each region. The fact that fragments from one region assembled with fragments from another region indicates overlap between the metagenomes of the two regions, and the extent of this overlap quantifies the similarity. The contig spectrum obtained from the mixed sample was modified in two respects to give what is called the cross-contig spectrum (Figure S4). First, any contig that contained fragments exclusively from a single region was removed (i.e., only contigs that included fragments from more than one region were counted). Thus for the contigs of size $q > 1$, \hat{C}_q , the number of q -contigs from the pooled sample that included fragments from more than one region, was calculated. Second, the number of 1-contigs from each region that assembled with any fragments from other regions was used as the number of 1-cross-contigs, \hat{C}_1 . The resulting cross-contig spectrum $[\hat{C}_1, \hat{C}_2, \hat{C}_3, \dots]$ was then compared to the mean cross-contig spectrum from simulated mixtures of the four assemblages. To simulate such mixtures requires a model of which genomes with a certain rank and abundance in one assemblage correspond to which genomes in another.

There are many ways to envision morphing one assemblage of genotypes (species defined on the genomic level by assembly of sequences) into another. For these analyses, two morphing modes were considered (Figure S5): (i) varying the proportion of genotypes that were shared between assemblages and (ii) varying the proportion of the genotypes whose abundance ranks were shuffled (i.e., subjected to a random permutation).

Using these two degrees of freedom, s (percent shared) and p (percent permuted), Monte Carlo analyses were performed to estimate the degree of morphing as measured by these two parameters to find maximum-likelihood values for s and p based on the closeness of the match to the cross-contig spectrum found for the pooled sample.

The Monte Carlo simulations were all performed using the best-fit models for each region. The cross-contig spectrum based on the mixed sample was used to perform these simulations (Figure S6). Each simulation included 861 pairs of s and p values spanning a 21×41 grid between 0% and 100% for each parameter. Each simulation randomly permuted the abundance rank of p of the most abundant genotypes, randomly assigned s of the genotypes to be shared, and determined the resulting predicted cross-contig spectrum. This was repeated 100 times for each combination of s and p values. The entire simulation, including the selection of the 2,500 fragments from each region, was repeated eight times resulting in 800 predicted cross-contig spectra for each combination of parameter values. The mean \hat{c}_q and variance $\hat{\sigma}_q^2$ of these 800 values were then used to construct a quasi-likelihood $\mathcal{L}(s, p)$

$$\ln \mathcal{L}(s, p) = - \sum_q \frac{(\hat{C}_q - \hat{c}_q)^2}{2\hat{\sigma}_q^2} \quad (2)$$

of matching the observed cross-contig spectrum, thereby generating a contour map of \mathcal{L} as in [11]. This log likelihood

would be expected if each cross-contig value were normally distributed. The contour map of the quasi-likelihood landscape was produced from this grid of 861 quasi-likelihood values. As a control, the whole procedure was repeated for all regions with nonoverlapping subsets of sequences all taken from the same geographical region (rather than from four different regions).

Results/Discussion

“Community” is commonly defined several ways, including “the species that occur together in space and time” [30] and “an association of interacting populations” [31]. Assemblage is probably the most proper term to describe viral groups, and most instances of “community” in the literature, both by ourselves and others, is not correct. See [32] for a disambiguation of some important ecological terms.

General Characteristics of the Marine Viral Metagenomes

On average, >91% of the sequences were not significantly similar to those in the extant databases (Figure 2A). A partial explanation for the high percentage of unknowns is almost certainly due to the shorter sequences (~100 bp on average) that are generated by pyrosequencing at 454 Life Sciences. Previous viral metagenomic studies that used Sanger sequencing (~650 bp fragments) found that >60% of the sequences were unknowns [33]. The Arctic Ocean sample had the highest percentage of known similarities (11%) to the SEED database, mostly because of the large number of prophage-like sequences (Table 1). Comparison of the marine viral sequences to the environmental database did not yield a significant number of new similarities compared to the SEED database (~2% to the environmental database), with the notable exception of the Sargasso Sea sample, where >9% of the similarities were to the environmental database, presumably because the major sources of sequences for the environmental database were the Sargasso Sea microbial metagenomes, originally collected in 2003 [17]. The overlap between the viral metagenome and the microbial metagenomes raises several important points. First, a significant number of viral sequences are retained on the larger-pore filters, either as free viruses, proviruses, or in cells undergoing a burst. The latter explanation was hypothesized by DeLong et al. [19], who observed a large number of viral similarities at one depth at the Hawaii Oceanic Time-series (HOT) station. Second, the microbial assemblages in the Sargasso Sea appear to be relatively stable over prolonged periods (~2 y). Finally, the small amount of sampling and sequencing represented by these two studies (~10¹² bp) is already constricting the unknown sequence space of the Sargasso Sea. With the continual decline in Sanger sequencing costs and introduction of large-scale pyrosequencing, metagenomic approaches should be able to characterize global sequence diversity in a relatively short period of time.

Among the fraction of sequences with similarity to the SEED database, most of the “knowns” were similarities to bacterial sequences in the Arctic, British Columbia, and Gulf of Mexico samples (Figure 2B). This can be accounted for by the following: (i) the larger number of microbial rather than viral genomes in the database, (ii) unidentified prophages within microbial genomes, (iii) the large amount of horizontal gene transfer between phages and their hosts, (iv) phages

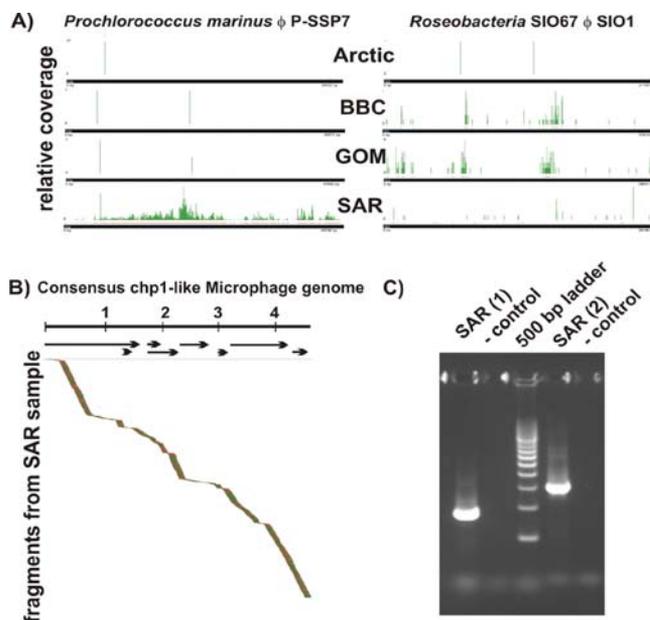


Figure 3. Distribution of Similarities and Assembly Controls

(A) Distribution of similarities between the four metagenome samples to the *P. marinus* phi P-SSP7 and *Roseobacteria* SIO67 phi SIO1 genomes (as determined by BLASTN analysis). The green bars represent the average number of sequences averaged over 100 bp windows.

(B) Comparison of fragments from the Sargasso Sea metagenome against the consensus ssDNA chp1-like microphage genome. The consensus from this assembly is in the Protocol S1.

(C) PCR verification of chp1-like microphages in original SAR sample. PCR primers were designed based on a consensus sequence from the assembly shown in (B). SAR1 is a ~900-bp fragment and SAR2 is a ~1,500-bp fragment.

DOI: 10.1371/journal.pbio.0040368.g003

carrying full genes from their host, as observed in sequenced phage genomes [34,35], and (v) the overall larger size of bacterial genes relative to viral genes, statistically increasing the probability of sequencing and hitting them.

The sample from the Sargasso Sea was exceptional in that the majority of “known” sequences were most similar to three *Prochlorococcus* phage genomes (Table 1) originally isolated from the same area of the ocean [34]. This finding suggests that just a few phage genomes from novel environments will greatly increase our understanding of viral diversity in these environments. The distribution of BLASTN similarities along the *Prochlorococcus marinus* phi P-SSP7 genome [34] is shown in Figure 3A. There is almost complete coverage of the genome within the Sargasso Sea sample. In contrast, the similarly sized *Roseobacteria* SIO67 phi SIO1 genome [36], which was isolated from near-shore waters in California, is only sparsely covered in the Sargasso Sea sample, but has higher coverage in the Gulf of Mexico and British Columbia samples. This supports the idea that certain phage groups are more prevalent in certain biogeographic regions. This general pattern was reinforced by the observation of a number of phage genomes and groups prevalent in different oceanic regions (Table 1).

The five most abundant putative viral-encoded enzymes (Table 2) appear to be involved in scavenging host nucleotides (e.g., riboreductases) and supporting host metabolism through the infection cycle (e.g., carboxylases and transferases). The viral fraction also contained *psbA* genes, which encode the D1 protein of photosystem II in the cyanobac-

Table 2. The Most Abundant Enzyme-Coding Genes in the Four Oceanic Viral Metagenomes

Marine Region	Enzyme Name	EC number	Gene Occurrences
Sargasso Sea (SAR)	Ribonucleotide reductase of class Ia (aerobic), alpha subunit	1.17.4.1	89
	Ribonucleoside-diphosphate reductase	1.17.4.1	75
	Ribonucleotide reductase of class II (coenzyme B12-dependent)	1.17.4.1	50
	GTP cyclohydrolase I, type 2	3.5.4.16	37
	Adenine-specific methyltransferase	2.1.1.72	22
Gulf of Mexico (GOM)	Formate dehydrogenase-O, major subunit	1.2.1.2	27
	Carbamoyl-phosphate synthase large chain	6.3.5.5	25
	Cytochrome c oxidase polypeptide I	1.9.3.1	24
	Ribonucleotide reductase of class II (coenzyme B12-dependent)	1.17.4.1	23
	DNA polymerase III alpha subunit	2.7.7.7	23
British Columbia coast (BBC)	Ribonucleotide reductase of class II (coenzyme B12-dependent)	1.17.4.1	34
	DNA polymerase III alpha subunit	2.7.7.7	22
	3-polypropenyl-4-hydroxybenzoate carboxylase	4.1.1.-	18
	Cytochrome c oxidase polypeptide I	1.9.3.1	18
	Ribonucleotide reductase of class Ia (aerobic), alpha subunit	1.17.4.1	18
Arctic Ocean	3-polypropenyl-4-hydroxybenzoate carboxylase	4.1.1.-	205
	DNA polymerase III alpha subunit	2.7.7.7	185
	Cytochrome c oxidase polypeptide I	1.9.3.1	175
	Isoleucyl-tRNA synthetase	6.1.1.5	157
	Methylcrotonyl-CoA carboxylase carboxyl transferase subunit	6.4.1.4	155

EC number, Enzyme Commission number.

DOI: 10.1371/journal.pbio.0040368.t002

teria. The majority of sequenced cyanophages carry this gene, and evidence is mounting that the cyanophages need the D1 protein for successful infection and replication [34,37,38]. The occurrence of *psbA* was lowest in the Arctic sample, probably reflecting a decrease in the host and cyanophage numbers in the colder environments.

Discovery of an Abundant Marine ssDNA Phage Group

The Sargasso Sea sample had a large number of sequences (6% of the total; Table 1) with significant similarities to chp1-like Chlamydia microvirus (Microviridae family). These viruses are small ssDNA phages. Assemblies from these sequences resulted in the near-complete genomes of several marine Microviridae phages from the Sargasso Sea sequences (Figure 3B). To our knowledge, this is the first report describing the presence of this phage group in the marine environment, which was previously overlooked because the amplification and cloning methods excluded ssDNA viruses. The only other report of ssDNA viruses in the marine environment was a Circovirus that infected diatoms [39]. However, the marine sequences in this study did not show any similarity to that virus. Sequences with significant similarity to the chp1-like phages were observed less frequently in the British Columbia (~10-fold less common than in SAR) and Gulf of Mexico samples (~100-fold less common than in SAR). No sequences from this group were found in the Arctic sample (Table 1 and Figure 4). Primers were designed against these genomes and appropriately sized DNA fragments were amplified from the Sargasso Sea sample (Figure 3C). No amplicons were detected in the Gulf of Mexico or British Columbia samples, suggesting that they were present at numbers below the level of detection in this PCR or had a divergent sequence. A geographical constraint that limits the distribution of these viruses would be most consistent with these results. However concerns about sample amplification and storage bias make it

impossible to accurately access the relative abundances of these viruses at this point.

Every Phage Everywhere?

The distribution of similarities to the chp1-like Microphage, *P. marinus* ϕ P-SSP7, *Roseobacteria* SIO67 ϕ SIO1, and others in the viral-fraction suggests that viruses have restricted geographical distributions similar to those observed in micro- and macro-organisms [40,41]. This is in contrast to studies that have shown that identical phage genes are distributed throughout the biosphere and that phages from soils and sediments can replicate in marine microbial populations [3,42,43]. To determine whether all marine phages are spread everywhere or if there is a strong regionalization, three different approaches were used.

A new version of the Phage Proteomic Tree was constructed, and similarities from the samples were mapped onto this tree (Figure 4). Eighty-four phage species were specific to one marine region, whereas 45 were common to all four. From the remaining phage species, 102 were found in several oceanic regions. The phylogenetic parsimony of phages from each sample was compared to the Phage Proteomic Tree using the PTP tests, because viruses do not have a single genetic locus conserved across all genomes. The PTP test showed that the distribution of phages in the marine samples is not random. First, marine phages are phylogenetically distinct from the available genomes, suggesting a “mariness” to the group as a whole ($p < 0.0001$; 10,000 randomizations). Second, there was a significant difference between phages from the different oceanic regions ($p < 0.0001$; 10,000 randomizations), supporting a geographical specificity for viruses despite the wide prevalence of some phage species.

An Isolation By Distance (IBD) approach demonstrated that there was a significant positive correlation between geographic distance (km) and genetic distance (as measured by Φ_{ST}) (Mantel test; $Z = -78.9$; $r = 0.585$; $p < 0.017$) (Figure 5),

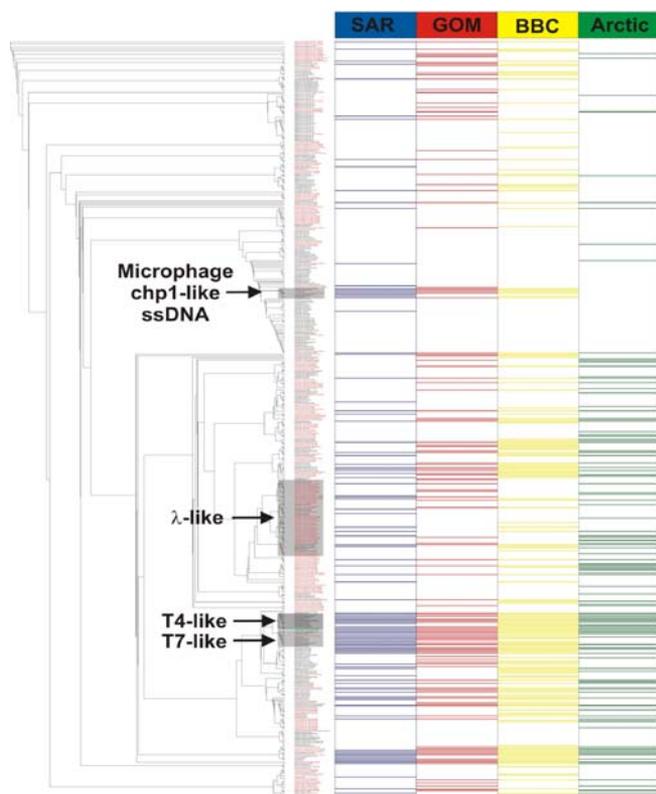


Figure 4. Types of Phages in the Four Metagenomes

A new version of the Phage Proteomic Tree (left panel) was constructed from 510 complete phage and prophage genomes using the previously described method [23]. The metagenomic sequences were compared to the phage on the Phage Proteomic Tree using TBLASTX, and the colored bars on the right represent significant similarities (E -value < 0.0001). Names of prophages are in red and the *Prochlorococcus* phage genomes are in green. An electronic version of the tree and a FASTA list of phage and prophage genomes used to make the tree are available at the SDSU Center for Universal Microbe Sequencing website (<http://scums.sdsu.edu/phage/Oceans>).

DOI: 10.1371/journal.pbio.0040368.g004

indicating that the further two sites are from each other, the more differences there are between the viral assemblages. The magnitude of the slope was very small with only $3.28 \times 10^{-5} \Phi_{ST}/\text{km}$.

Considering that any two locations on Earth can be separated by a maximum of 20,000 km (half the circumference of the globe), by extrapolation, any two viral assemblages could have a phylogenetic diversity of at most $0.656 \Phi_{ST}$. Although these data suggest a limit to the distribution of viruses among marine environments (e.g., due to limited viral movement or geographical selective pressure) ($\Phi_{ST} \gg 0$), it also indicates that no two marine viral assemblages could be totally different ($\Phi_{ST} \ll 1$). Rather, they would exhibit a relatively large phylogenetic overlap.

Together the PTP and IBD test support that the marine virome is composed of specific viral groups. These viral assemblages undergo a regionalization, although a large fraction is vastly widespread. It is possible that some viruses are distributed ubiquitously, but their relative contribution to overall assemblage structure differs between oceanic regions. If this were true, then cross-contigs—i.e., contigs

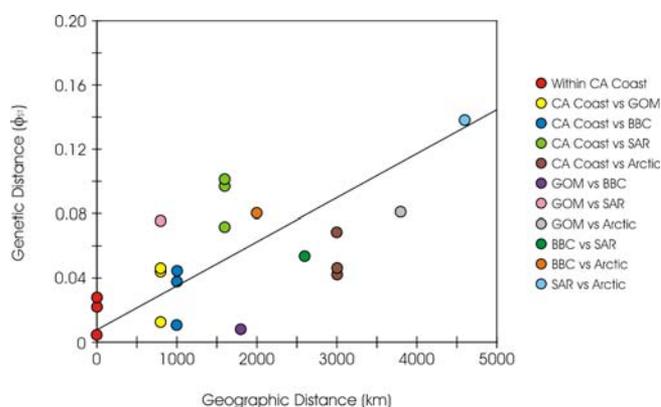


Figure 5. Relationship between Geographic and Genetic Distances of Marine Viral Assemblages

In addition to the four metagenomes sequenced for this study, the previous viral metagenomes from the San Diego area (California coast) were also included in this analysis [10]. There was a significant correlation of $3.28 \times 10^{-5} \Phi_{ST} / \text{km}$ (Mantel test, $Z = -78.9$, $p < 0.017$, $r = 0.585$). DOI: 10.1371/journal.pbio.0040368.g005

made of sequences from different metagenomes—would reflect this composition.

In the computer model of cross-contig analysis, all four viral assemblages were considered at the same time. Assemblies were performed and cross-contigs were identified. A Monte Carlo simulation was used to explain the average cross-contig spectrum. A full description of the assemblies and Monte-Carlo simulations are in the Protocol S1.

A number of genotypes (varied between 0% and 100%) were arbitrarily and randomly defined as shared between samples; at the same time, the occurrence of individuals in the viral assemblage was also varied (Figure 6). As an illustration, imagine two assemblages sharing 100 viruses, but with the relative rank on a rank-abundance curve being shuffled for the top viruses in the assemblage (see Protocol S1). The best explanation of the observed cross-contigs is shown in Figure 6 and estimates that 35% of the most abundant genomes in any sample would have to be permuted in their relative abundance rank and that 100% of the viruses would have to be shared between samples. The intrasample controls showed that 85%–95% of the most abundant genomes were shared and 0%–0.5% were permuted (although 100% and 0% were expected, respectively). This discrepancy is probably due to limitations in the methodology used.

This cross-contig analysis suggests that any two viral assemblages could have a vast majority of species in common and the order of the ranks in the rank-abundance curve could be determined by shuffling about a third of the most abundant species. These results confirm that geographical and changing environmental conditions allow different viral genotypes to become more or less prevalent within different assemblages while sharing essentially the same types of viruses. The less abundant viruses are not lost altogether, merely reduced in occurrence.

Local Versus Global Diversity

Using the PHACCS analysis system [29], the genotype richness, diversity, and evenness of the different metagenomes were estimated (Table 3). The British Columbia viral metagenome was the most genotype-rich (129,000 predicted genotypes) and diverse (H' of 10.8 nats), whereas the Arctic

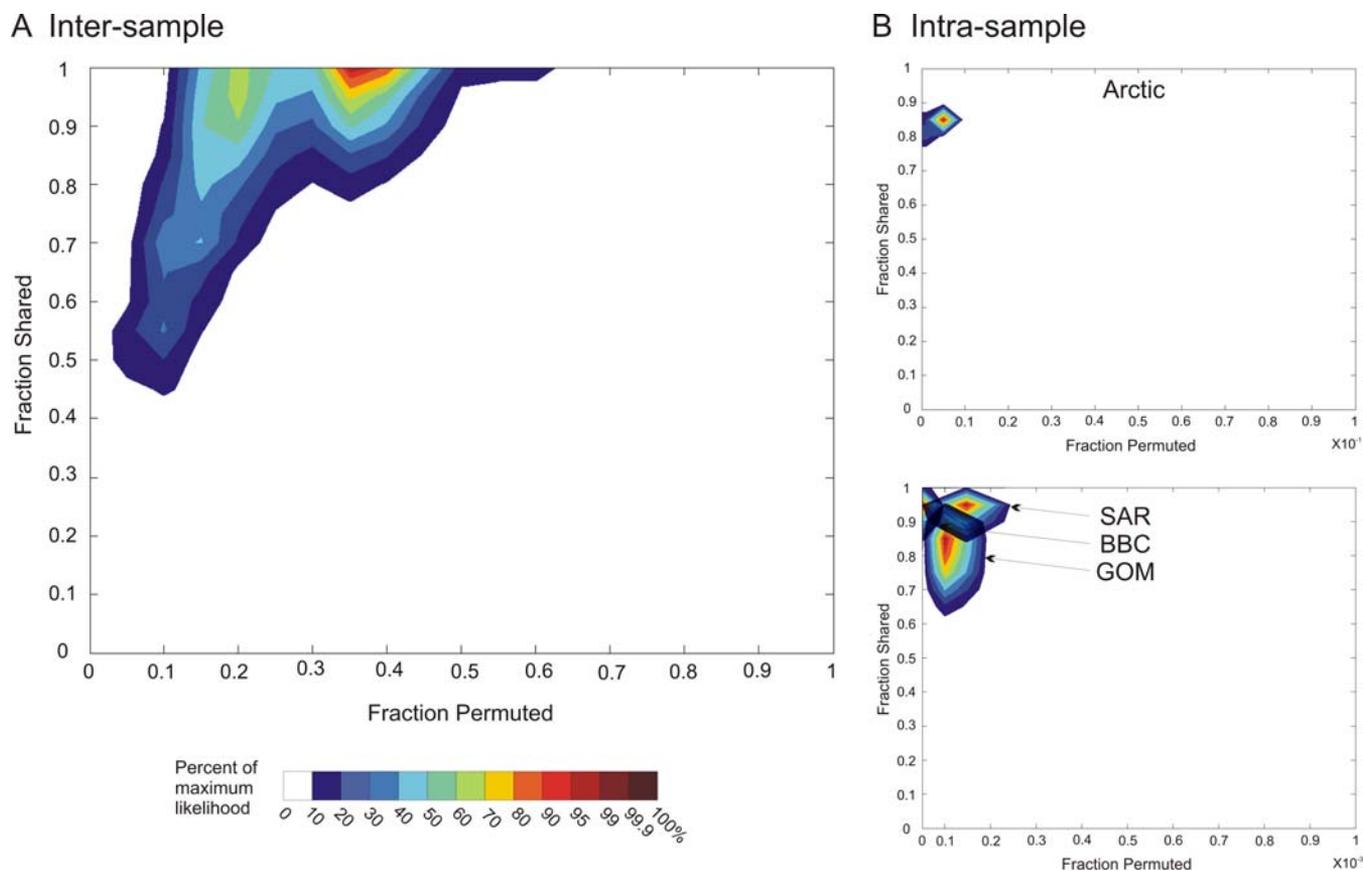


Figure 6. Monte Carlo Simulation of Cross-Contigs between Metagenomic Samples

(A) For the intersample analysis, the maximum likelihood occurred at 35% fraction permuted and 100% fraction shared. (B) The maximum likelihood was between 0% and 0.5% fraction permuted and 85% and 95 % fraction shared for the intrasample controls.

DOI: 10.1371/journal.pbio.0040368.g006

metagenome was the least genotype-rich (532 predicted genotypes) and diverse (H' of 6.05 nats).

Being located on the west coast of the North American continent, the coast of British Columbia is in an upwelling area. It is also enclosed and fed by many rivers. These conditions might importantly increase the diversity of microbial communities and thus provide an explanation for the very high viral assemblage diversity estimated in this oceanic region. Omitting the BBC, the viral diversity of the other regions (the Gulf of Mexico, Sargasso Sea, and Arctic Ocean) correlate with the well-established North-South latitudinal diversity gradient [44], with a larger diversity at lower latitudes. Planktonic diversity patterns of near-shore versus off-shore (more diverse plankton assemblages off-shore) [45] were not observed here; the large spatial scale of the sampling probably masked this effect if present.

Assemblies of the mixed sample were used to predict global viral diversity using PHACCS. A total of 57,600 different viral genotypes in all four regions (H' of 9.8 nats) was estimated. This number is smaller than the number of genotypes predicted in the BBC sample, which may indicate an undersampling for the mixed metagenome or be due to some of the assumptions of the model. Taken together, these data indicate that the global marine viral richness could be as high as a few hundred thousand species, with a regional

richness sometimes almost as high, likely because of migration processes.

Integrative Versus Single Samples

It was expected that the integrated samples would be more even because it is assumed the viruses that were most abundant at one spatial-temporal time point would be rarer at another (“kill-the-winner” hypothesis). As summarized in Table 3, the evenness of the single time point sample (SAR 0.905) fell in between that of the three integrated samples (Arctic 0.964; BBC 0.918; GOM 0.851). Similarly, the predicted richness (5140 genotypes) and diversity (H' 7.74 nats) at the single point represented by the Sargasso Sea sample fell in between that of the integrated samples (richness 532–129,000; H' 6.05–10.8 nats). Because of factors with a supposedly greater impact, like latitude, it is not clear that integrating individual samples gave a greater depth of coverage.

Without a doubt, many interesting trends based on depth and a wide variety of other spatial, biological, and temporal parameters were missed by the integrative sampling used here. However, this sampling does provide a useful overview of the marine virome on a global and regional scale. Currently, there are no real criteria as to what constitutes a useful size or time scale for sampling natural viral assemblages, so there is no particular advantage or disadvantage to keeping samples separate or analyzing them as a metadataset.

Rather the sampling scheme should be driven by the question being addressed. Viral assemblages are interesting in their own right, not just in context of their host communities. However, future studies should also start cross-correlating the viruses with their hosts. Of particular interest will be determining if the “islands” and ORFans observed in microbial genomes are represented in the virome [6,46].

Potential Sampling and Processing Biases

Sampling bias in the current datasets was primarily due to loss of large viruses during filtering. Currently, there is no experimental method to avoid this problem. The cesium chloride gradients used here recover all known phage groups, and essentially all the viral-like particles in the starting samples migrate to the proper density in these preparations (as observed by epifluorescence microscopy; unpublished data). Unfortunately, the cloning methods used here will not recover RNA viruses. Suttle et al. [47,48] have shown that RNA viruses are present in the marine environment. Whereas most electron microscopy [49,50] and nucleic acid-based studies [51] have not found RNA viruses in large numbers, RNA viruses are still believed to be important components of the marine virome that need additional study.

Another potential source of bias is the different times that the samples were stored before processing. Phage particles are very stable and often stored for decades at 4 °C. This is a commonly known lab phenomenon and is supported by the observation that the oldest viral concentrates (~12 y old) in this study had very high concentration of viruses (>10⁹ viral-like particles per ml). Different phages, however, may have different decay rates under these conditions. This does not seem to be especially problematic, because there is no correlation between the types of viruses observed and the storage time. For example, the Arctic and SAR samples are the most recently harvested samples, yet they have the biggest differences in terms of types of phages (Table 1). Nonetheless, there may be effects of storage on the composition of the viral assemblages. For this reason, analyses based on absolute abundances of one specific virus to another were avoided in this study. Instead, the presence of a sequence in the metagenome was simply assumed to mean that the virus was in the original sample (i.e., an occurrence).

Whole-genome amplification techniques introduce biases in the relative concentrations of different genomes. Tests of Genomphi by the manufacturer and others [52,53] have not

reported a significant bias in the amplification of circular double-stranded DNA (dsDNA), with the exception of very small dsDNA targets (<1 kb), which are much smaller than the vast majority of marine viruses, and of ssDNA, which will probably be a preferred target for the DNA polymerase. Although not bias-free, Genomphi is the most accurate amplification method available [54]. Interesting trends associated with viral assemblage structure may have been missed because of our choice of using presence/absence data for the analyses presented here, but by being conservative there should not be any effects of storage, amplification, and sampling biases on our interpretations.

Conclusion

The metagenomic analysis of viral assemblages from the Arctic Ocean, the coast of British Columbia, the Gulf of Mexico, and the Sargasso Sea presented here has changed our perception on the composition of viral assemblages in the sea. First, there is clear evidence that the composition of viral assemblages varies in different geographic regions probably reflecting selective pressure. Previously overlooked viral groups, such as ssDNA viruses and prophages, can be major constituents of marine viral assemblages (Sargasso Sea and Arctic Ocean, respectively). Second, global viral diversity is high (possibly a few hundred thousand viral species), but regional diversity can be almost as high due to viral migration. This migration provides opportunities for global exchange of DNA among viral genomes, as predicted by the mosaic model [55]. Viral diversity also varied according to latitude, with a higher richness at low latitudes. Finally, it seems that although some viral species are endemic and others are ubiquitous, the vast majority are widespread and shared between several oceanic regions. Invasion and replacement by new phages does not appear to be an important structuring factor for these viral assemblages. What sets different assemblages apart is likely the change in abundance of its most abundant members, supporting to some extent the old tenet “everything is everywhere, but, the environment selects” [56] for marine viruses.

Supporting Information

Figure S1. Frequency of Homopolymeric Tracts in the Four Marine Viromes, the Complete Phage Genomes, and Twenty, Randomly Chosen Microbial Genomes

The tracts from 3 nucleotides (nt) to 15 nt were counted and normalized to the number of bases in each sequence. One 3-nt tract is

Table 3. Viral Assemblage Structure Predicted from Assembly of Metagenomic Sequences

Sample	Richness	Evenness	Most Abundant Genotype (%)	Shannon-Wiener Index
Arctic	532 genotypes	0.964	2.27	6.05 nats
BBC	129,000 genotypes	0.918	7.28	10.8 nats
GOM	15,400 genotypes	0.851	13.3	8.21 nats
SAR	5140 genotypes	0.905	8.45	7.74 nats
Mixed	57,600 genotypes	0.895	9.34	9.81 nats

Ten separate assemblies of 10,000 sequences chosen at random from each library were performed for each sample. For the mixed sample, 2,500 randomly chosen fragments were used from each library. The average contig spectrum was used to predict assemblage structure using PHACCS.

DOI: 10.1371/journal.pbio.0040368.t003

found approximately every 30 bp, whereas one 15-nt tract is found approximately every 10 million bp. The 510 complete phage genomes totaled 18,909,173 bp in length, and the microbial genomes totaled 22,110,123 bp in length. The lengths of the pyrosequenced libraries are given in the text.

Found at DOI: 10.1371/journal.pbio.0040368.sg001 (1.3 MB TIF)

Figure S2. Relative Abundance of Phages in the Four Metagenomes
Because of way the samples were stored and the long storage time, the distribution shown may not accurately reflect the reality.

Found at DOI: 10.1371/journal.pbio.0040368.sg002 (104 KB TIF)

Figure S3. Determining a Normal Contig Spectrum

Found at DOI: 10.1371/journal.pbio.0040368.sg003 (135 KB TIF)

Figure S4. Getting a Cross-Contig Spectrum.

Found at DOI: 10.1371/journal.pbio.0040368.sg004 (3.1 MB TIF)

Figure S5. The Possible Scenarios Considered in the Monte Carlo Simulation to Explain the Observed Cross-Contigs

Found at DOI: 10.1371/journal.pbio.0040368.sg005 (143 KB TIF)

Figure S6. Analyzing a Cross-Contig Spectrum

Found at DOI: 10.1371/journal.pbio.0040368.sg006 (589 KB TIF)

Protocol S1. Details on Materials and Methods.

Found at DOI: 10.1371/journal.pbio.0040368.sd001 (39 KB PDF)

Accession Numbers

The Genome Projects Database (<http://www.ncbi.nlm.nih.gov/Genomes>) accession numbers for the sequences are 17765 (GOM),

17767 (BBC), 17769 (Arctic), and 17771 (SAR); the Genome Catalogue (<http://gensc.sf.net>) accession numbers are 000002_GCAT (GOM), 000003_GCAT (BBC), 000004_GCAT (Arctic), and 000005_GCAT (SAR); and the GOLD database (<http://www.genomesonline.org>) GOLDstamps are GM00060 (GOM), GM00061 (BBC), GM00062 (Arctic), and GM00063 (SAR).

Acknowledgments

The GOM, BBC, and Arctic samples were collected with the generous help of the crew and scientists aboard the research vessels *Longhorn*, *Mirai*, *Raddisson*, *Walton Smith*, and *Vector*. We are grateful to A.M. Comeau, A.C. Ortmann, C.M. Short, S.M. Short, M.G. Weinbauer, and S.W. Wilhelm for sample collection and processing; as well as K. Shimada, E.C. Carmack, and J. Paul for providing the opportunity to participate in the *Mirai* and *Walton Smith* expeditions. The Sargasso Sea samples were collected with the assistance of R. Morris and the Captain and Crew or the R.V. *Weatherbird II*.

Author contributions. FR conceived and designed the experiments. CC, AC, MH, and RP performed the experiments. FA, BF, MB, PS, RE, SK, HL, JMM, JEM, JN, SR, CS, and FR analyzed the data. RO contributed reagents/materials/analysis tools. FA, BF, MB, RE, JEM, CS, and FR wrote the paper.

Funding. These collections were supported by NSERC grants to CAS (Discovery, Shiptime and Research Network [CASES]) and by ONR and NSF grants to CAS. Ship time and collection was supported by an NSF microbial observatory grant to CAC. The Marine Microbial Initiative by the Gordon and Betty Moore Foundation (FR) sponsored the sequencing, bioinformatics, and mathematical analyses.

Competing interests. The authors have declared that no competing interests exist.

References

- Weinbauer MG (2004) Ecology of prokaryotic viruses. *FEMS Microbiol Rev* 28: 127–181.
- Wilcox RM, Fuhrman JA (1994) Bacterial viruses in coastal seawater: Lytic rather than lysogenic production. *Mar Ecol Prog Ser* 114: 35–45.
- Sano E, Carlson S, Wegley L, Rohwer F (2004) Movement of viruses between biomes. *Appl Environ Microbiol* 70: 5842–5846.
- Canchaya C, Fournous G, Chibani-Chennoufi S, Dillmann ML, Brussow H (2003) Phage as agents of lateral gene transfer. *Curr Opin Microbiol* 6: 417–424.
- Davis BM, Waldor MK (2002) Mobile genetic elements and bacterial pathogenesis. In: Craig NL, Gragie R, Gellert M, Lambowitz AM, editors. *Mobile DNA II*. Washington (DC): ASM Press. pp. 1040–1055.
- Daubin V, Ochman H (2004) Bacterial genomes as new gene homes: The genealogy of ORFans in *E. coli*. *Genome Res* 14: 1036–1042.
- Thingstad TF, Lignell R (1997) Theoretical models for the control of bacterial growth rate, abundance, diversity and carbon demand. *Aquat Microb Ecol* 13: 19–27.
- Hoffmann K, Rodriguez-Brito B, Breitbart M, Bangor D, Angly FE, et al. (2005) The structure of marine phage populations. In: Kjelstrup S, Hustad J, Gundersen T, Røsjorde A, Tsatsaronis G, editors. *Trondheim (Norway): Tapir Academic Press*. 5 p.
- Thompson JR, Pacocha S, Pharino C, Klepac-Ceraj V, Hunt DE, et al. (2005) Genotypic diversity within a natural coastal bacterioplankton population. *Science* 307: 1311–1313.
- Breitbart M, Salamon P, Andresen B, Mahaffy JM, Segall AM, et al. (2002) Genomic analysis of uncultured marine viral communities. *Proc Natl Acad Sci U S A* 99: 14250–14255.
- Breitbart M, Felts B, Kelley S, Mahaffy J, Nulton J, et al. (2004) Diversity and population structure of a nearshore marine sediment viral community. *Proc R Soc Lond Ser B Biol Sci* 271: 565–574.
- Breitbart M, Hewson I, Felts B, Mahaffy JM, Nulton J, et al. (2003) Metagenomic analyses of an uncultured viral community from human feces. *J Bacteriol* 185: 6220–6223.
- Breitbart M, Rohwer F (2005) Method for discovering novel DNA viruses in blood using viral particle selection and shotgun sequencing. *Biotechniques* 39: 729–736.
- Edwards R, Rodriguez-Brito B, Wegley L, Haynes M, Breitbart M, et al. (2006) Using pyrosequencing to shed light on deep mine microbial ecology. *BMC Genomics* 7.
- Zhang T, Breitbart M, Lee W, Run J-Q, Wei C, et al. (2006) RNA viral community in human feces: Prevalence of plant pathogenic viruses. *PLoS Biol* 4: e3. DOI: 10.1371/journal.pbio.0040003.
- Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, et al. (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* 428: 37–43.
- Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, et al. (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science* 304: 66–74.
- Tringe SG, von Mering C, Kobayashi A, Salamov AA, Chen K, et al. (2005) Comparative metagenomics of microbial communities. *Science* 308: 554–557.
- DeLong EF, Preston CM, Mincer T, Rich V, Hallam SJ, et al. (2006) Community genomics among stratified microbial assemblages in the ocean's interior. *Science* 311: 496–503.
- Sambrook J, Fritsch EF, Maniatis T (1989) *Molecular cloning: A laboratory manual*. New York: Cold Spring Harbor Laboratory Press. 1659 p.
- Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437: 376–480.
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215: 403–410.
- Rohwer F, Edwards R (2002) The phage proteomic tree: A genome based taxonomy for phage. *J Bacteriol* 184: 4529–4535.
- Martin AP (2002) Phylogenetic approaches for describing and comparing the diversity of microbial communities. *Appl Environ Microbiol* 68: 3673–3682.
- Swofford DL (2000) PAUP*. Phylogenetic analysis using parsimony (*and other methods). 4.0 ed. Sunderland (Massachusetts): Sinauer Associates.
- Jensen JL, Bohonak AJ, Kelley ST (2005) Isolation by distance, web service. *BMC Genetics* 6: 13.
- Schneider S, Kueffer JM, Roessli D, Excoffier L (1997) Arlequin: A software for population genetic data analysis. 1.1 ed. Geneva: Genetics and Biometry Lab, Department of Anthropology.
- Sutton GG, White O, Adams MD, Kerlavage AR (1995) TIGR assembler: A new tool for assembling large shotgun sequencing projects. *Genome Sci Technol* 1: 9–19.
- Angly F, Rodriguez-Brito B, Bangor D, McNairn P, Breitbart M, et al. (2005) PHACCS, an online tool for estimating the structure and diversity of uncultured viral communities using metagenomic information. *BMC Bioinformatics* 6: 41.
- Begon M, Harper JL, Townsend CR (1990) *Ecology: Individuals, populations, and communities*. Boston: Blackwell Scientific Publications. 945 p.
- Ricklefs RE (1990) *Ecology*. New York: WH Freeman. 896 p.
- Fauth JE, Bernardo J, Camara M, Resetarits WJ, Buskirk JV, et al. (1996) Simplifying the jargon of community ecology: A conceptual approach. *Am Nat* 147: 5.
- Edwards R, Rohwer F (2005) Viral metagenomics. *Nature Rev Microbiol* 3: 504–510.
- Sullivan MB, Coleman ML, Weigle P, Rohwer F, Chisholm SW (2005) Three *Prochlorococcus* cyanophage genomes: Signature features and ecological interpretations. *PLoS Biol* 3: e144. DOI: 10.1371/journal.pbio.0030144.
- Seguritan V, Feng IW, Rohwer F, Swift M, Segall AM (2003) Genome sequences of two closely related *Vibrio parahaemolyticus* phages, VP16T and VP16C, that coexist in the same host. *J Bacteriol* 185: 6434–6447.

36. Rohwer F, Segall AM, Steward G, Seguritan V, Breitbart M, et al. (2000) The complete genomic sequence of the marine phage Roseophage SIO1 shares homology with non-marine phages. *Limnol Oceanogr* 42: 408–418.
37. Zeidner G, Bielawski JP, Shmoish M, Scanlan DJ, Sabehi G, et al. (2005) Potential photosynthesis gene recombination between *Prochlorococcus* and *Synechococcus* via viral intermediates. *Environmental Microbiology* 7: 1505–1513.
38. Sullivan MB, Lindell D, Lee JA, Thompson LR, Bielawski JP, et al. (2006) Prevalence and evolution of core photosystem II genes in marine cyanobacterial viruses and their hosts. *PLoS Biol* 4: e234. DOI: 10.1371/journal.pbio.0040234.
39. Nagasaki K, Tomaru Y, Takao Y, Nishida K, Shirai Y, et al. (2005) Previously unknown virus infects marine diatom. *Appl Environ Microbiol* 71: 3528–3535.
40. Martiny JBH, Bohannan BJM, Brown JH, Colwell RK, Fuhrman JA, et al. (2006) Microbial biogeography: Putting microorganisms on the map. *Nat Rev Microbiol* 4: 102–112.
41. Dolan JR (2005) An introduction to the biogeography of aquatic microbes. *Aquat Microb Ecol* 41: 39–48.
42. Breitbart M, Rohwer F (2004) Global distribution of nearly identical phage-encoded DNA sequences. *FEMS Microbiol Lett* 236: 245–252.
43. Short CM, Suttle CA (2005) Nearly identical bacteriophage structural gene sequences are widely distributed in marine and freshwater environments. *Appl Environ Microbiol* 71: 480–486.
44. Hillebrand H (2004) Strength, slope and variability of marine latitudinal gradients. *Marine Ecol Prog Ser* 273: 251–267.
45. Hutchinson GE (1961) The paradox of the plankton. *Am Nat* 95: 137–145.
46. Coleman ML, Sullivan MB, Martiny AC, Steglich C, Barry K, et al. (2006) Genomic islands and the ecology and evolution of *Prochlorococcus*. *Science* 311: 1768–1770.
47. Culley AI, Lang AS, Suttle CA (2003) High diversity of unknown picorna-like viruses in the sea. *Nature* 424: 1054–1057.
48. Culley AI, Lang AS, Suttle CA (2006) Metagenomic analysis of coastal RNA virus communities. *Science* 312: 1795–1798.
49. Nomizu T, Mizuike A (1986) Electron microscopy of submicron particles in natural waters: Specimen preparation by centrifugation. *Mikrochim Acta* 1: 65–72.
50. Moebus K (1991) Preliminary observations on the concentration of marine bacteriophages in the water around Helgoland. *Helgo Meeresunters* 45: 411–422.
51. Steward GF, Montiel JL, Azam F (2000) Genome size distributions indicate variability and similarities among marine viral assemblages from diverse environments. *Limnol Oceanogr* 45: 1697–1706.
52. Hutchison III CA, Smith HO, Pfannkoch C, Venter JC (2005) Cell-free cloning using phi29 DNA polymerase. *Proc Natl Acad Sci U S A* 102: 17332–17336.
53. Yokouchi H, Fukuoka Y, Mukoyama D, Calugay R, Takeyama H, et al. (2006) Whole-metagenome amplification of a microbial community associated with scleractinian coral by multiple displacement amplification using phi-29 polymerase. *Environ Microbiol* 8: 1155–1163.
54. Pinard R, de Winter A, Sarkis GJ, Gerstein MB, Tartaro KR, et al. (2006) Assessment of whole genome amplification-induced bias through high-throughput, massively parallel whole genome sequencing. *BMC Genomics* 7: 216.
55. Hendrix RW, Lawrence JG, Hatfull GF, Casjens S (2000) The origins and ongoing evolution of viruses. *Trends Microbiol* 8: 499–500.
56. de Wit R, Bouvier T (2006) “Everything is everywhere, but, the environment selects”; what did Baas Becking and Beijerinck really say? *Environ Microb* 8: 755–758.