

# Metagenomic Exploration of Viruses throughout the Indian Ocean

Shannon J. Williamson<sup>1\*</sup>, Lisa Zeigler Allen<sup>1,4</sup>, Hernan A. Lorenzi<sup>2</sup>, Douglas W. Fadrosh<sup>1</sup>, Daniel Bami<sup>3</sup>, Mathangi Thiagarajan<sup>2</sup>, John P. McCrow<sup>1</sup>, Andrey Tovchigrechko<sup>2</sup>, Shibu Yooseph<sup>3</sup>, J. Craig Venter<sup>1</sup>

**1** Microbial and Environmental Genomics, J. Craig Venter Institute, San Diego, California, United States of America, **2** Informatics, J. Craig Venter Institute, Rockville, Maryland, United States of America, **3** Informatics, J. Craig Venter Institute, San Diego, California, United States of America, **4** University of California San Diego, Scripps Institution of Oceanography, La Jolla, California, United States of America

## Abstract

The characterization of global marine microbial taxonomic and functional diversity is a primary goal of the Global Ocean Sampling Expedition. As part of this study, 19 water samples were collected aboard the *Sorcerer II* sailing vessel from the southern Indian Ocean in an effort to more thoroughly understand the lifestyle strategies of the microbial inhabitants of this ultra-oligotrophic region. No investigations of whole viroplankton assemblages have been conducted on waters collected from the Indian Ocean or across multiple size fractions thus far. Therefore, the goals of this study were to examine the effect of size fractionation on viral consortia structure and function and understand the diversity and functional potential of the Indian Ocean virome. Five samples were selected for comprehensive metagenomic exploration; and sequencing was performed on the microbes captured on 3.0-, 0.8- and 0.1  $\mu\text{m}$  membrane filters as well as the viral fraction ( $<0.1 \mu\text{m}$ ). Phylogenetic approaches were also used to identify predicted proteins of viral origin in the larger fractions of data from all Indian Ocean samples, which were included in subsequent metagenomic analyses. Taxonomic profiling of viral sequences suggested that size fractionation of marine microbial communities enriches for specific groups of viruses within the different size classes and functional characterization further substantiated this observation. Functional analyses also revealed a relative enrichment for metabolic proteins of viral origin that potentially reflect the physiological condition of host cells in the Indian Ocean including those involved in nitrogen metabolism and oxidative phosphorylation. A novel classification method, MGTAXA, was used to assess virus-host relationships in the Indian Ocean by predicting the taxonomy of putative host genera, with *Prochlorococcus*, *Acanthochlois* and members of the SAR86 cluster comprising the most abundant predictions. This is the first study to holistically explore viroplankton dynamics across multiple size classes and provides unprecedented insight into virus diversity, metabolic potential and virus-host interactions.

**Citation:** Williamson SJ, Allen LZ, Lorenzi HA, Fadrosh DW, Bami D, et al. (2012) Metagenomic Exploration of Viruses throughout the Indian Ocean. PLoS ONE 7(10): e42047. doi:10.1371/journal.pone.0042047

**Editor:** Jack Anthony Gilbert, Argonne National Laboratory, United States of America

**Received:** August 22, 2011; **Accepted:** July 2, 2012; **Published:** October 17, 2012

**Copyright:** © 2012 Williamson 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.

**Funding:** This research was supported by the Office of Science (BER), U.S. Department of Energy, Cooperative Agreement No. De-FC02-02ER63453, the Gordon and Betty Moore Foundation, the National Science Foundation award 0850256 and TeraGrid allocation DEB100001 on the Texas Advanced Computing Center Ranger. The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing Interests:** The authors have declared that no competing interests exist.

\* E-mail: shannon.j.williamson@gmail.com

† These authors contributed equally to this work.

## Introduction

Viruses (predominantly bacteriophages) are the most abundant biological components of marine ecosystems, generally outnumbering their microbial hosts by an order of magnitude [1]. Viruses are significant agents of microbial mortality, influencing diversity, and play an integral role in marine ecosystem processes such as nutrient transformation and cycling [2–4]. Through interactions with microbes, viruses also influence the global flow of genes affecting host cell phenotype, niche adaptation and evolution [5]. Metagenomic investigations of viruses collected from varied marine ecosystems have provided insight into local and global viral diversity [6–8], genotypic distribution [9,10], functional potential [11–15] and replication strategy [16,17]. The majority of these metagenomic studies have been conducted on the viral fraction of marine samples, where a pre-filtration step (generally ranging from 0.22- to 0.45  $\mu\text{m}$ ) was used to physically separate viral particles from cellular organisms. Alternatively, metagenomic

investigations have been conducted on virus-like sequences contained within the cellular fraction of marine samples, including surface water samples collected during the first reported phase of the Global Ocean Sampling (GOS) Expedition (termed Phase I) [12], and a depth profile collected from the HOT station ALOHA [18]. In these instances, viral sequences were identified based on their homology to known viruses. Each of these strategies has its limitations. Examinations of viruses within the cellular fraction of metagenomic data alone inevitably results in an underestimation of viral sequences due to its dependency on similarity to previously sequenced viruses and are constrained to only those viruses (or their nucleic acids) that were physically captured. Similarly, evaluation of only the viral fraction can result in a less than comprehensive picture of viral diversity and functional capacity since it excludes viruses removed through pre-filtration.

To date, no investigations of whole viroplankton assemblages have been conducted on waters collected from the Indian Ocean,

however, targeted studies of cyanophage [19] and a virus infecting a heterotrophic flagellate [20] have been reported. In order to gain a more thorough understanding of the GOS Indian Ocean (GOS-IO) viromes and to better appreciate the implications of size fractionation, both the cellular and viral fractions were examined using metagenomic approaches that target dsDNA viral sequences. Virus-host associations were also evaluated using a novel classification method that predicts the taxonomy of putative microbial hosts using assembled viral metagenomic data based on polynucleotide compositional signatures described by Interpolated Context Models (ICMs) adopted from Phymm [21]. This study represents the first comparative analysis of viruses across multiple size classes from marine water samples.

## Results and Discussion

### Dataset characteristics and identification of viral sequences from filter fractions

Surface water samples (~400 L) were collected from 17 sites from the tropical Indian Ocean between August and October 2005 aboard the S/V Sorcerer II. Two additional sites were sampled off the island of Zanzibar, Tanzania using alternate vessels (**Figure 1; Table S1**). The microbial community was pre-filtered using 20  $\mu\text{m}$  mesh Nynet net and then size fractionated by serial filtration through 3.0-, 0.8-, and 0.1  $\mu\text{m}$  membrane filters. The viral fraction of water samples (i.e. <0.1  $\mu\text{m}$ ) was concentrated and purified as described previously [22] (see Materials and Methods for details). DNA was extracted from microbial cells and viral particles as previously described [22,23] and sequenced using a combination of Sanger and pyrosequencing technologies.

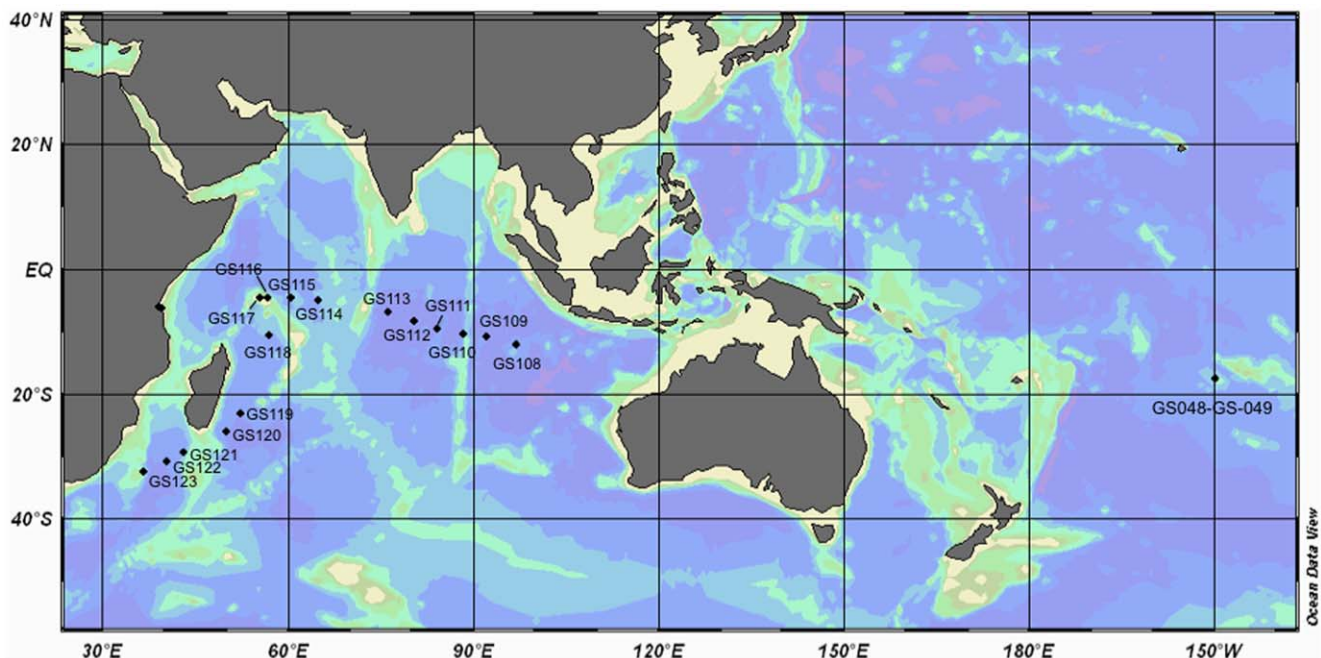
While the microbes retained within the 0.1–0.8  $\mu\text{m}$  size fraction were sequenced for all samples, five independent samples were selected for comprehensive metagenomic exploration with sequencing performed on all microbial size fractions 20–0.1  $\mu\text{m}$ , including the viral fraction (<0.1  $\mu\text{m}$ ) (**Table 1; Table S2**). The five samples that were comprehensively sequenced were selected

based on sufficient quantity and quality of DNA across all size fractions. A total of approximately 228K Sanger and 2.2M 454 Titanium sequence reads were produced from the five viral libraries, resulting in a combined 2.7M predicted ORFs (see Materials and Methods for details on the ORF prediction pipeline). The Sanger and 454 datasets had average read lengths of 989 bp and 360 bp, respectively. Despite rigorous purification of the viral concentrates (VCs) prior to metagenomic library construction, it became apparent during sequence analysis that one library, GSIOVIR110, contained non-trivial amounts of cellular contamination similar to sequences found in the larger filter data (0.1–20  $\mu\text{m}$ ). Therefore, this sample was not included in subsequent analyses with the exception of co-assembly with data from all size fractions.

The identification of viral sequences within the metagenomic data from larger size fractions was accomplished using a phylogenomic approach, the Automated Phylogenetic Inference System (APIS), that was designed for annotation of genomic [24] and metagenomic [25] datasets. APIS taxonomically and functionally classifies each sequence inferred from neighbor-joining phylogenetic trees (see Materials and Methods for more details); unlike previous reports using homology-based methods [12,18]. This analysis resulted in the identification of 102,790 predicted proteins of viral origin within the larger size fractions (0.1–20  $\mu\text{m}$ ) (**Table S2**), representing 2.8% of the total predicted proteins from the Indian Ocean microbial dataset (~3.6M total predicted proteins). This estimate is almost identical to a previous study which found that viral sequences accounted for ~3% of GOS Phase I microbial data collected between Nova Scotia and French Polynesia [12].

### Viral genotypic diversity

Viral genotypic diversity was estimated at the Indian Ocean sampling locations using the Phage Communities from Contig Spectrum (PHACCS) tool [26]. Richness, evenness and diversity estimates were obtained for each viral library using the complete



**Figure 1. Map of Indian Ocean indicating where samples for metagenomic analysis were collected.**

doi:10.1371/journal.pone.0042047.g001

**Table 1.** Sanger and 454 pyrosequencing statistics for Indian Ocean viral metagenomes.

Sample	# Sequences	Avg. Read Length (bp)	%GC	# Predicted Proteins	% Phylogenetic Trees	% Unknown
<i>Sanger</i>						
GSIOVIR108	44,504	990	39.1	70,563	14.42	12.5
GSIOVIR110*	47,572	977	35.8	73,947	n/a	8.8
GSIOVIR112	46,443	1,003	38.2	74,721	26.70	10.7
GSIOVIR117	43,530	992	38.8	74,679	21.56	12.4
GSIOVIR122	46,584	983	39.3	71,007	23.69	11.3
<b>Total</b>	<b>228,633</b>	<b>989</b>	<b>38.2</b>	<b>364,917</b>	<b>Avg. = 21.59</b>	<b>Avg. = 11.1</b>
<i>454</i>						
GSIOVIR108	318,094	389	42	254,792	14.12	15.2
GSIOVIR110*	331,693	329	43	231,953	n/a	12.7
GSIOVIR112	492,396	358	42	344,886	17.91	11.3
GSIOVIR117	710,101	378	41	574,756	14.12	12
GSIOVIR122	337,266	347	45	226,097	22.41	12.5
<b>Total</b>	<b>2,189,550</b>	<b>360</b>	<b>40.4</b>	<b>2,362,318</b>	<b>Avg. = 17.14</b>	<b>Avg. = 11.9</b>

\*Sample not used in taxonomic and functional analyses due to cellular contamination. n/a denote where the data was not used in study. Percent of phylogenetic trees denotes the amount of sequences (ORFs) that were taxonomically classified via APIS.  
doi:10.1371/journal.pone.0042047.t001

set of reads generated upon sequencing. Comparisons were performed across the following categories based on the contig spectrum produced from Newbler-based assemblies of: 1) fragmented Sanger data (i.e. unmated), 2) paired-end Sanger data, 3) all 454 data, and 4) sub-sampled 454 data (**Table 2**, **Table 3**, and **Table S3**). Estimates of genotype richness varied widely between the Sanger and most of the 454 categories, but the relative ranking was conserved. On average, the most diverse site among all assemblies tested was GSIOVIR117, which also had the largest number of genotypes. The least diverse site, on average, was GSIOVIR122, which also had the lowest number of genotypes (**Table 2** and **Table 3**). Evenness estimates were similar across categories with a slight increase in assemblies that utilized all of the 454 data per library rather than sub-samples (**Table 2**). Richness and evenness estimates, along with the most abundant genotype percentages, suggest that all sites harbored diverse viral populations with most of the diversity residing in the “long tail” of the genotype distribution.

To account for the possibility of varying genome size, diversity estimates were also obtained on the sub-sampled 454 datasets using a range of average viral genome sizes (**Table 3**). For these analyses, sample richness varied greatly depending on the average genome size used. The greater richness estimates (e.g. >100k genotypes) could reflect a limit in the PHACCS algorithm and likely do not reflect a relevant genome size estimation. It has been reported that the average genome size of marine viruses is typically between 50–80 kb [27,28]. Our data suggests the same based on estimates of richness, with an interesting change in GSIOVIR117, where the richness estimates were more in line with the Sanger estimates when using an estimated genome size of 80 kb (**Table 3**). Lastly, as mentioned above, the relative ranking of samples with respect to diversity estimates was conserved and corresponded to the total number of predicted peptides per site. However, previous reports have suggested that there is no link between sequencing depth and the number of genotypes [26]. Tools for genome size estimations of viruses are available; however depend heavily on reference databases. Based on phylogenetic classification of ORFs from sequence reads, ~20% of the GOS-IO data had significant

similarity to a known sequence and therefore was classified (**Table 1**). Thus, we feel current approaches are not directly applicable as yet to these data.

To the best of our knowledge, this is the first study to compare PHACCS-based estimations of viral diversity using sequence information produced from two different platforms (Sanger and 454), yet assembled using the same methods. The number of predicted viral genotypes (richness) in a sample was the most variable with respect to the type of data produced (Sanger mated and unmated; 454 all and sub-samples). Sub-sampling of datasets is recommended to reduce coverage and improve the quality of PHACCS estimations.

This comparison demonstrated that estimations of viral diversity using PHACCS, particularly the number of predicted genotypes, may be influenced by the nature of viral data and their subsequent assembly; yet most estimates fell within the range of previously published reports [6–8]. Furthermore, measurements of viral consortia diversity were underestimated since viruses retained in the larger size classes were excluded from the analysis, as they were not included in the assemblies. Subsequent taxonomic and functional assessments (discussed below) suggest that viral sequences found in the microbial (cellular) fraction are distinct compared to those generated from the viral-size fractions.

### Taxonomic binning of Indian Ocean virome

Sequences originating from discrete viral fractions (VFs) and the larger fractions (LFs) were taxonomically characterized using two approaches. One approach was based on a BLASTP comparison against the NCBI non-redundant (nr) protein database, which does not include predicted protein sequences from viral metagenomes. The other was a phylogenomic approach using APIS. Examination of metagenomic data using APIS resulted in ~20% of sequences from the VFs producing phylogenetic trees, and thus taxonomic classification (**Table 1**). The low level of APIS classification was due to the stringency of the method (see Materials and Methods for details). The majority of predicted protein sequences from the VF were characterized as cellular using

**Table 2.** Estimated diversity of Indian Ocean viroplankton assemblages using PHACCS.

Sample	Richness (# genotypes)	Evenness	SW Index (nats)	Most abundant genotype (%)
<i>Sanger_Frag</i>				
GSIOVIR108	3,327	0.85	6.86	5.62
GSIOVIR112	10,592	0.85	7.84	4.32
GSIOVIR117	22,040	0.88	8.83	2.47
GSIOVIR122	554	0.89	5.59	6.63
<i>Sanger_PE</i>				
GSIOVIR108	1,309	0.86	6.17	6.36
GSIOVIR112	10,560	0.85	7.83	4.34
GSIOVIR117	17,069	0.9	8.73	2.23
GSIOVIR122	388	0.89	5.32	6.96
<i>454_All</i>				
GSIOVIR108	1,237	0.93	6.6	3.14
GSIOVIR112	4,337	0.91	7.59	2.81
GSIOVIR117	4,850	0.92	7.84	2.09
GSIOVIR122	1,387	0.91	6.61	3.65
<i>454_Sub</i>				
GSIOVIR108	405±79	0.89±0.005	5.34±0.15	6.95±0.22
GSIOVIR112	26,703±844	0.84±0.003	8.55±0.05	3.79±0.13
GSIOVIR117	211,276±8,728	0.80±0.003	8.06±0.40	5.37±0.38
GSIOVIR122	559±98	0.88±0.005	5.55±0.14	6.97±0.17

Frag: fragment; PE: paired end; Sub: random subsamples; SW: Shannon-Weiner.  
doi:10.1371/journal.pone.0042047.t002

**Table 3.** Estimated diversity using varied genome sizes (kb) via PHACCS.

Richness	100 kb		80 kb		50 kb <sup>(§)</sup>		25 kb		15 kb	
<b>GS108viral</b>	212	+/- 43	261	+/- 52	405	+/- 79	784	+/- 146	1,283	+/- 233
<b>GS112viral</b>	340,000	+/- 113,137	38,437	+/- 2,966	26,703	+/- 844	300,000	+/- 282,843	73,000	+/- 10,850
<b>GS117viral</b>	109,914	+/- 312,933	15,521	+/- 23,108	211,276	+/- 8,728	119,072	+/- 309,835	26,434	+/- 19,148
<b>GS122viral</b>	120,226	+/- 315,429	367	+/- 67	559	+/- 98	1,062	+/- 178	1,720	+/- 281
<b>Evenness</b>										
<b>GS108viral</b>	0.88	+/- 0.01	0.89	+/- 0.01	0.89	+/- 0.01	0.9	+/- 0.01	0.91	+/- 0.01
<b>GS112viral</b>	0.72	+/- 0.01	0.79	+/- 0	0.84	+/- 0	0.78	+/- 0.02	0.86	+/- 0
<b>GS117viral</b>	0.79	+/- 0.04	0.82	+/- 0.02	0.8	+/- 0	0.83	+/- 0.03	0.86	+/- 0.01
<b>GS122viral</b>	0.8	+/- 0.15	0.87	+/- 0.01	0.88	+/- 0.01	0.89	+/- 0	0.9	+/- 0
<b>SW Index (nats)</b>										
<b>GS108viral</b>	4.72	+/- 0.14	4.92	+/- 0.14	5.34	+/- 0.15	6	+/- 0.13	6.48	+/- 0.13
<b>GS112viral</b>	9.07	+/- 0.21	8.36	+/- 0.04	8.55	+/- 0.01	9.68	+/- 0.29	9.57	+/- 0.11
<b>GS117viral</b>	7.42	+/- 0.91	7.41	+/- 0.64	8.06	+/- 0.4	8.36	+/- 0.85	8.58	+/- 0.53
<b>GS122viral</b>	5.26	+/- 0.79	5.13	+/- 0.15	5.55	+/- 0.14	6.18	+/- 0.14	6.66	+/- 0.14
<b>Most abundant genotype (%)</b>										
<b>GS108viral</b>	9.32	+/- 0.15	4.05	+/- 0.03	5.05	+/- 0.04	6.95	+/- 0.22	8.44	+/- 0.12
<b>GS112viral</b>	6.24	+/- 0.01	2.64	+/- 0.01	4.22	+/- 0.28	3.79	+/- 0.13	5.19	+/- 0.01
<b>GS117viral</b>	6.3	+/- 0.36	3.27	+/- 0.38	4.22	+/- 0.21	5.37	+/- 0.38	5.82	+/- 0.49
<b>GS122viral</b>	10.54	+/- 2.34	4.23	+/- 0.08	5.22	+/- 0.1	6.97	+/- 0.17	8.51	+/- 0.24

<sup>(§)</sup>indicates the genome size that was used in Table 2.  
doi:10.1371/journal.pone.0042047.t003

**Table 4.** Taxonomic characterization of Indian Ocean viroplankton assemblages.

Taxa	% LF APIS	% VF APIS	% VF BLAST
<b>“Kingdom<sup>a</sup>”</b>			
Cellular	NA	42.6	68
Virus	2.8	55.6	32
Other	NA	NA	0.1
Mixed	NA	1.8	NA
<b>“Virus<sup>b</sup>”</b>			
Metagenomic dsDNA	29.2	80.3	NA
dsDNA viruses	39.9	17.2	97.34
ssDNA viruses	0	0	0.08
ssRNA viruses	0	0	0.07
dsRNA viruses	0	0	0.01
Retro-transcribing viruses	0	0	0.02
Unclassified phage	NA	NA	2.45
Unclassified archaeal phage	0	0	0.05
Unclassified virus	0	0	0
<b>“dsDNA Virus<sup>c</sup>”</b>			
Marine Planktonic (virome)	36.1	94.5	NA
<b>Caudoviridae</b>	50.3	0.8	95.3
<i>Phycodnaviridae</i>	2.8	0.1	3.9
Marine Sediment (virome)	1.5	3.8	NA
Hot Spring (virome)	1.9	0.5	NA
<i>Mimiviridae</i>	1.4	0	0.3
<i>Iridoviridae</i>	0.4	0	0.4
Human Feces (virome)	0.1	0.2	NA
<i>Poxviridae</i>	0.1	0	0
<b>“Caudoviruses<sup>d</sup>”</b>			
<i>Myoviridae</i>	93.1	56.9	54.3
<i>Podoviridae</i>	5.1	30.6	27.6
<i>Siphoviridae</i>	1.5	11.1	17
Unclassified	0.05	0.3	1.2
Mixed	0.3	1.1	NA
<b>“NCLDVs<sup>e</sup>”</b>			
<i>Phycodnaviridae</i>	58.6	82.8	83.9
<i>Iridoviridae</i>	8	11.6	8.5
<i>Mimiviridae</i>	29.9	4.9	7.3
<i>Poxviridae</i>	2.3	0.5	0.3
<i>Ascoviridae</i>	0.3	0.1	0
<i>Asfarviridae</i>	1	0.1	0
<b>“Phycoviruses<sup>f</sup>”</b>			
Chlorovirus	46.1	80.4	80
Ostreococcus (OsV5)	43.3	17.5	18.7
Phaeovirus	1.9	0.4	0.8
Coccolithovirus	7	0.9	0.3
Prymnesiovirus ( <i>P. globosa</i> virus)	NA	NA	0.05
Raphidovirus ( <i>H. akashiwo</i> virus 01)	NA	NA	0.04
Mixed	1.3	0.8	NA

<sup>a</sup>% total protein sequences classified at highest (kingdom) taxonomic level per category (LF or VF).

<sup>b</sup>% total protein sequences characterized as viral per category.

**Table 4.** Cont.

<sup>c</sup>% of total of protein sequences characterized as dsDNA viruses comprising >0.1% in one category.

<sup>d</sup>% of total protein sequences characterized as Caudoviruses per category.

<sup>e</sup>% of total protein sequences characterized as NCLDVs per category.

<sup>f</sup>% of total protein sequences characterized as Phycoviruses per category.

n/a = not available.

doi:10.1371/journal.pone.0042047.t004

both the homology (68%) and phylogenetic-based approaches (43%) (Table 4), despite evidence that cellular contamination was not present in the viral samples (i.e. no amplification of the 16S rRNA gene during library construction; see Materials and Methods for details). A larger proportion of sequences were classified as Virus at the kingdom level through APIS (56%) than by homology comparison to the non-redundant (nr) database (32%), due to the inclusion of viral metagenomic data of diverse origin within the APIS database. As expected due to the sample processing methodology employed, viral sequences originating from the VF as well as the LF were most similar to dsDNA viruses; however, the most abundant taxonomic group varied.

According to phylogenetic characterization, the largest proportion of sequences from VF in the dsDNA virus category were most similar to sequences derived from environmental samples, including marine planktonic (~95%) and marine sediment viral sequences (~4%) (Table 4). Previous taxonomic characterization of marine viral metagenomes [6–8] suggests that the majority of viruses within these datasets are most similar to caudoviruses (i.e. tailed phages) and contain a large proportion of cyanophage-like sequences. Similarly, BLASTP-based comparisons of VF sequences against nr (NCBI) placed almost the entire dataset (~95%) within the known *Caudoviridae* family, demonstrating the utility of including viral metagenomes in search databases (Table 4). Alternatively, phylogenetic characterization of LF viral sequences indicated that these were more similar to known caudoviruses (~50%) rather than marine planktonic viruses (~36%) suggesting inherent differences in population structure between the VF and LF (Table 4). These differences were more apparent at the family level where myovirus sequences were significantly more abundant in LF than VF, while podo- and siphovirus sequences were more abundant in the VF than LF. Overall, cyanophage-like sequences comprised a significant proportion of the Indian Ocean VF and LF caudovirus data based on BLAST and phylogenetic analyses; potentially reflective of the surplus of these sequences in reference databases. Sequences were most similar to phages infecting *Prochlorococcus* (P-SSM2, P-SSP7 and P-SSM4) and *Synechococcus* (S-PM2 and the Syn group of phages) (Table S4).

The VF and LF also differed with respect to the relative distribution of nucleocytoplasmic large DNA viruses (NCLDV), a group that includes the phycodna- and Mimi-viridae families that appear to infect eukaryotic organisms (Table 4). Although the sequenced strain of Mimivirus is known to infect amoebae [29], it has been proposed that marine Mimivirus-like sequences originate from viruses that infect a variety of marine protists including eukaryotic phytoplankton, specifically hapto- and prasinophytes [30–32]. Homologs to all 6 of the NCLDV families were detected in the LF and VF with different levels of relative abundance (Table 4). Overall, the LF had a larger proportion of sequences similar to the phyco- and Mimi-like viruses than VF. When restricting taxonomic evaluation to only NCLDV-related sequences, phycovirus homologs were more prevalent in the VF than the LF. The majority of VF phycovirus homologs were further classified as Chlorovirus while LF homologs were equally

distributed between Chlorovirus and *Ostreococcus* virus OsV5 (**Table 4**). Mimivirus-like sequences were also more abundant in the LF than VF. The varying taxonomic distributions of viruses that likely infect bacteria (tailed phage) and eukaryotic phytoplankton (phyco- and Mimi-like viruses) between the VF and LF strongly suggests that size fractionation of marine microbial communities enriches for specific groups of viruses and significantly influences the distribution of viral families. This emphasizes the utility of sequencing across multiple size classes rather than just those designated as viral or cellular. The differential partitioning of viruses could be attributed to the morphology of the virus particle (for extracellular viruses), virus adsorption to host cells, active viral replication within infected host cells, and latent viral infection.

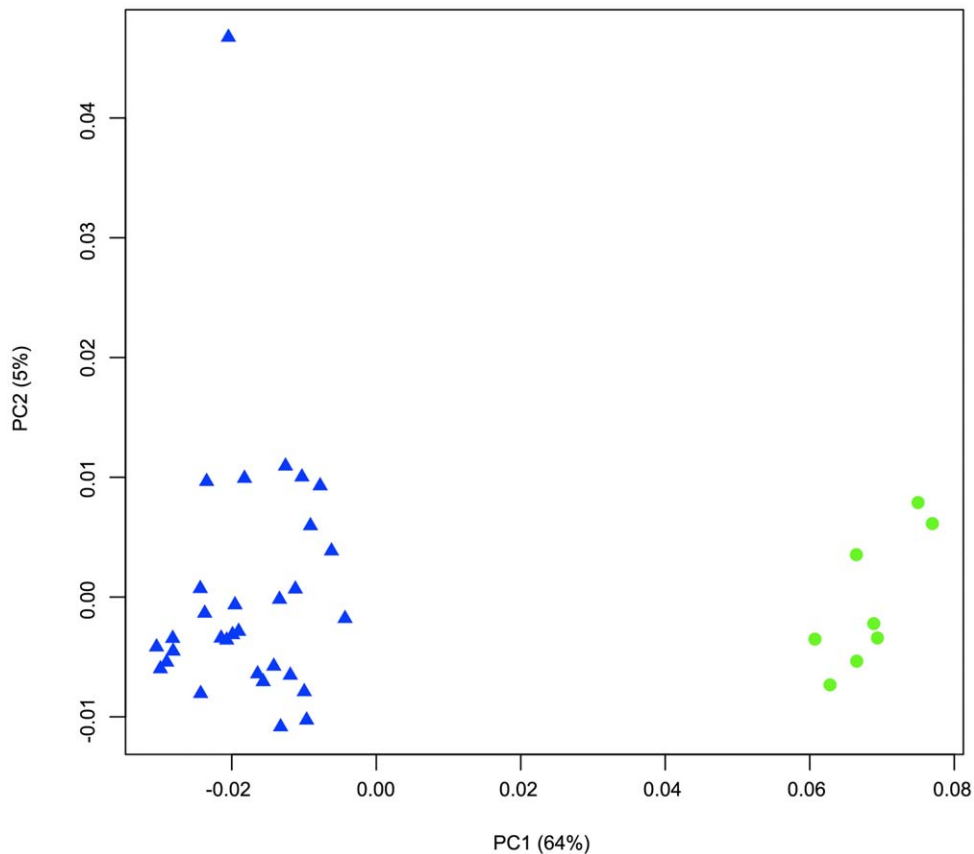
### Functional potential of Indian Ocean viruses

**Comparison of VF and LF sequences.** To examine the gene function of viruses captured in this study, ORFs derived from sequencing reads were assigned to existing protein clusters that included the published GOS data as well as reference data [33]. The GOS-IO protein dataset consisted of all ORFs from the viromes (VF and LF). Statistical analyses were used to determine if significant relationships existed between the functional repertoire of viruses within the VF and LF (Principal Component Analysis, PCA), as well as the measured oceanographic environmental parameters (Canonical Correlation Analysis, CCA). Proportional abundances for all protein clusters were calculated by library and used in each analysis. PCA indicated that viruses within these two groups were functionally divergent (**Figure 2**). However, no significant relationship existed between viral function and environmental factors as measured by CCA (data not shown). A significant amount of the variance was accounted for by the first two components (69%), demonstrating that size fractionation partitions viruses with different functional potentials, likely related to the similar observation found by taxonomic evaluation. A PCA biplot was also used to identify protein clusters containing viral sequences that were driving the separation of the VF and LF groups. Of particular note, two different clusters containing the same viral protein, large subunit terminase, were over-represented; cluster A predominantly contained LF sequences (LF = 2,683; VF = 1,014), and cluster B mostly consisted of VF sequences (VF = 74,812; LF = 329). Terminases are viral enzymes consisting of a small and large subunit that enable the packaging of DNA into viral proheads [34]. The small subunit is responsible for DNA recognition while the large subunit performs several functions including DNA cutting, portal vertex docking and ATPase-mediated translocation of DNA [34]. A phylogenetic tree was created from representative sequences contained in these two clusters including Indian Ocean VF and LF, GOS Phase I and publicly available reference sequences. The tree was characterized by two phylogenetically distinct groups that effectively partitioned the two terminase clusters (**Figure 3**). The LF group, corresponding to cluster A, contained well supported clades with *Prochlorococcus* and *Synechococcus* cyanophage reference sequences in addition to environmental viral sequences, which could represent uncultivated cyanophage. The VF group, corresponding to cluster B, was more phylogenetically diverse, characterized by multiple well-defined clades containing mixtures of reference phage and environmental viral sequences. The reference phage in the cluster B group belonged to all three of the tailed phage families rather than just described T4-like myoviruses present in the LF group, similar to the taxonomic distribution of Indian Ocean VF sequences (**Table 4**). Terminase-mediated packaging of DNA is a conserved mechanism among diverse linear, dsDNA containing viruses [34]. Several studies indicate that the terminase large

subunits share a common ancestry with other translocating ATPases including helicases and type I and III restriction endonucleases [35,36] and that terminase phylogeny may be predictive of the nature of the ends of virus DNA (e.g. cohesive ends) [37]. Clading of the environmental virus sequences with T4-like reference phages in the cluster A group suggests that these viruses may possess terminally redundant, circularly permuted genomes that are packaged using a T4-like headful packaging mechanism. However, it is difficult to speculate as to the packaging mechanisms of viruses in cluster B since the nature of the DNA ends for the reference phage is unclear.

**Pathway analyses.** Protein clusters containing viral sequences originating from the VF and LF were also categorized in the context of KEGG pathways, with the addition of a virus structure category, in order to assess the functional potential of Indian Ocean viroplankton. The majority of protein sequences from VF and LF were not mapped to a pathway and remained uncharacterized (VF = ~80%; LF = ~37%). This level of functional novelty was not unexpected due to high abundance of hypothetical proteins in each category (VF = ~68%; LF = ~50%). Smaller proportions of data were considered poorly characterized (VF = ~3%; LF = ~5%) or were not specific to a particular pathway (VF = ~3%; LF = ~8%). The remaining sequences were mapped primarily to the Virus Structure, Metabolism and Genetic Information Processing categories (**Figure 4, Table S5**). A heatmap of functional categories (**Figure S1**) further demonstrated the partitioning of VF and LF sequences as observed through PCA with differential clustering of the VF and LF. The vast majority of VF (~79%) and LF (~80%) sequences within the genetic information processing pathway were categorized as putative DNA replication, recombination and repair proteins (**Table S5**).

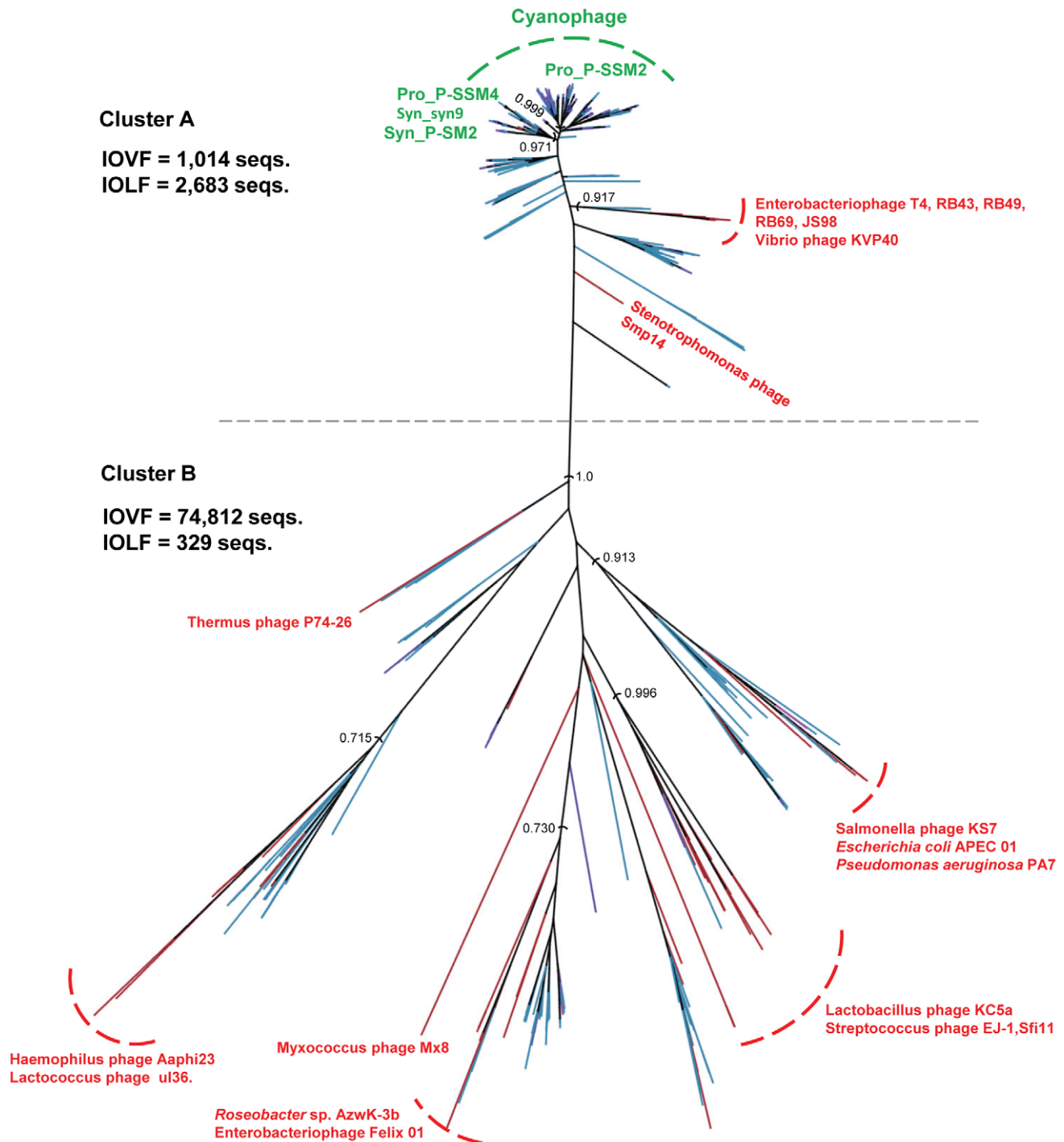
The largest proportion of viral sequences within the Metabolism pathway was attributed to energy metabolism, with a slight enrichment in the VF (**Figure 4, Table S5**). A relatively small proportion of these sequences from the VF and LF were classified into lower categories including nitrogen metabolism and oxidative phosphorylation. Although nitrogen metabolism genes have been previously documented in viral metagenomes created from a variety of environmental settings [11], the exact nature of the genes contributing to this pathway was unclear. Several studies have demonstrated that viruses infecting eukaryotic phytoplankton and zooplankton (likely to be retained in the LF) carry genes of either host or bacterial (i.e prey) origin [38–45]. However, the majority of these genes are involved with lipid, carbohydrate and protein metabolism and polyamine biosynthesis rather than nitrogen metabolism or oxidative phosphorylation [46–48]. Only VF sequences possessed enzyme commission (EC) numbers that could be linked directly to the KEGG nitrogen reduction and fixation pathway and these were examined in more detail. The majority of VF sequences within the nitrogen metabolism category were annotated as glutamate synthase (n = 98), which together with glutamine synthetase, comprise the GS-GOGAT pathway. This pathway facilitates the process of ammonium assimilation in phytoplankton [49] and is dependent on the availability of nitrogen compounds in the environment. The Indian Ocean is considered an oligotrophic water mass with very low concentrations of available nitrogen [50], and nitrogen concentrations measured in our samples were indeed indicative of a nitrogen-limited environment (**Table S1**). The presence of glutamate synthase genes suggest that viruses may play a role in nitrogen modulation and assimilation during the infection of host cells. Proteins involved in oxidative phosphorylation (OP) pathway were much more abundant than photosynthesis-related proteins,



**Figure 2. Principal component analysis (PCA) of the relative abundance of VF and LF viral sequences within protein clusters.** Viral libraries are represented by the green circles and larger fraction libraries are represented by the blue triangles.  
doi:10.1371/journal.pone.0042047.g002

comprising ~30% of VF and ~11% of LF sequences within the energy metabolism category (**Figure 3, Table S5**); with 466 VF and 25 LF sequences possessing EC numbers. NADH dehydrogenase I subunit and inorganic diphosphatase were represented in both the VF ( $n = 255$  and  $53$  respectively) and LF ( $n = 7$  and  $18$  respectively) while the *cbb3*-type cytochrome C oxidase subunit I was only detected in the VF ( $n = 158$ ). To the best of our knowledge, this is the first report of viral cytochrome C oxidase and inorganic diphosphatase genes in the marine environment. NADH dehydrogenase and cytochrome C oxidase are both components of the electron transport chain in bacteria, which is ultimately used to produce ATP. Viral type I NADH dehydrogenase genes were first reported by Alperovitch-Lavy and colleagues [51] and were detected through a combined analysis of GOS microbial scaffolds and long PCR amplification of viral fractions collected from the Pacific Line Islands [52]. Interestingly, the viral NADH dehydrogenase genes were co-localized on viral scaffolds (and amplicons) containing photosystem I and II genes suggesting that cyanophage encode this complex. A subsequent search of GOS scaffolds by Sharon and coworkers (2011) for viral auxiliary metabolic genes also revealed the presence of viral Type I NADH dehydrogenase subunits putatively involved in cyclic electron flow around PSI and respiration during viral infection. Again, these genes were attributed to cyanophage since the majority of scaffolds containing viral auxiliary genes that were examined in this study appeared to be related to known cyanophages [15]. It's possible that the viral NADH dehydrogenase genes observed in this study are of cyanophage origin due to

the abundance of cyanophage-like sequences in the Indian Ocean data. However, the abundance of virus-SAR86 host predictions (discussed later) coupled with the presence of viral *cbb3*-type cytochrome C oxidases, which are only found in proteobacteria, suggests that viruses that infect heterotrophic bacteria may also be the source of these genes. The enzyme inorganic diphosphatase catalyzes the conversion of diphosphate (Ppi) to phosphate (Pi), which is needed for the production of ATP. Out of the three OP enzymes, inorganic diphosphatase was more evenly distributed between the VF and LF suggesting that a diverse group of viruses may carry this gene. If the viral version of inorganic diphosphatase is expressed and functional during infection, viruses could potentially contribute to host ATP production. This process could temporarily prolong the lifespan of the host and increase replication efficiency, analogous to viral NADH dehydrogenase and PS genes. An alternative hypothesis is that viral inorganic diphosphatase is used to produce Pi for incorporation into viral nucleic acids. Phosphate concentration in the marine environment is thought to influence virus production due to their inherently high nucleic acid to protein ratio [52]. The ability to influence the availability of phosphate during infection could maximize nucleic acid biosynthesis. Furthermore, a variety of phosphorous metabolism genes have been detected in the genomes of cultivated viruses that infect heterotrophic bacteria, cyanophage genomes [53,54], as well as numerous viral metagenomes [11,12,55], suggesting that viruses have developed multiple strategies to address phosphate-limiting conditions. It is now well known that cyanophages carry photosynthesis (PS) related genes, including

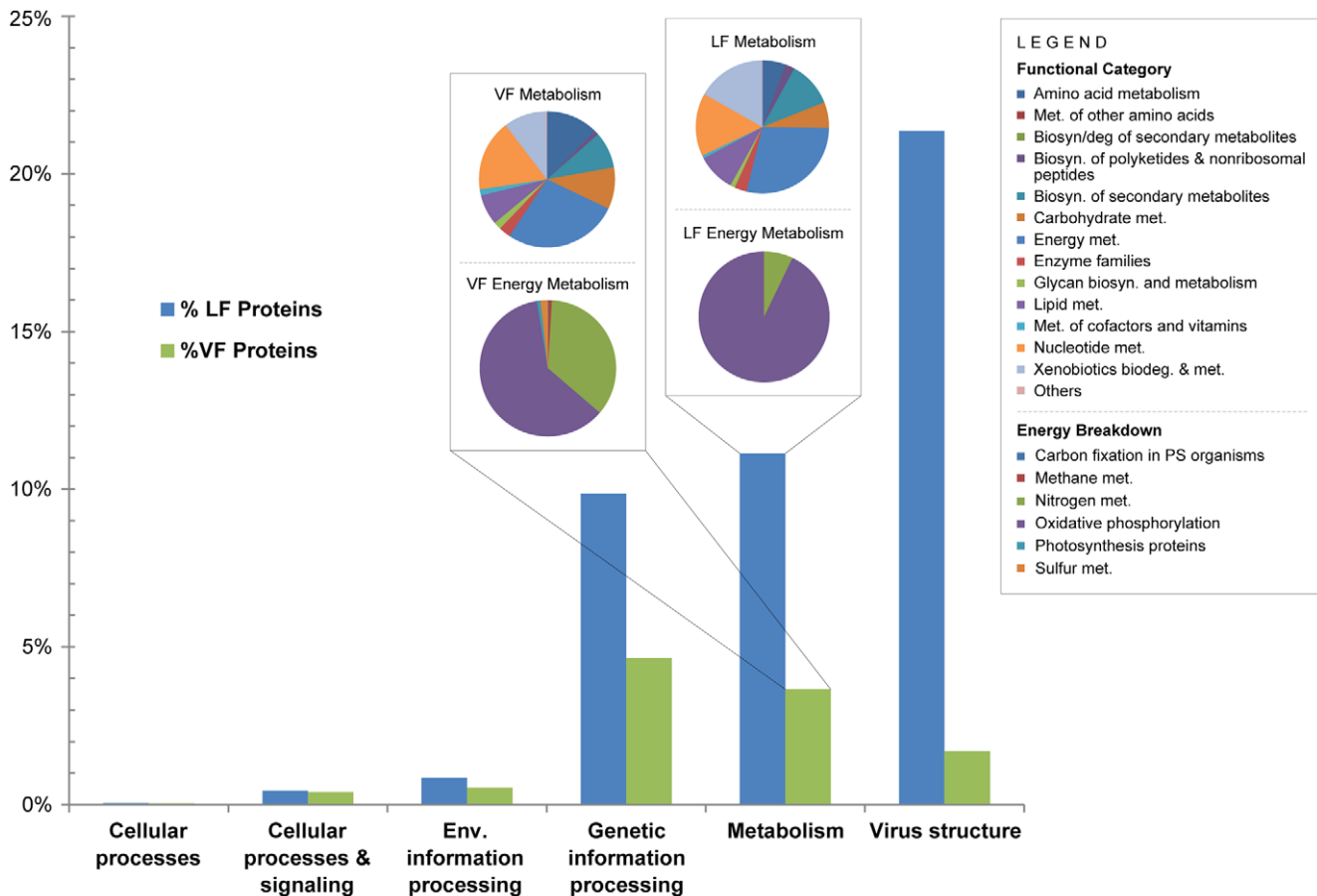


**Figure 3. A phylogenetic tree built from amino acid sequences using FastTree for the gene encoding the large subunit terminase.** Selected reference sequences are colored red and green. Representative Indian Ocean VF and LF sequences are colored pink and purple respectively. GOS Phase I sequences are colored blue. Confidence values are displayed on the tree.  
doi:10.1371/journal.pone.0042047.g003

those associated with photosystems I and II [14,56–59], and the presence of viral PS genes has been documented in numerous marine metagenomic studies [7,12,13,15,18]. However, only a small proportion of VF sequences (0.42%) could be mapped to proteins involved in photosynthesis based on KEGG classification of protein clusters. A direct BLAST analysis of VF and LF sequences using PSI and PSII genes collected from cyanophage

genomes (PSII) as well as *Prochlorococcus* and *Synechococcus* (PSI) (Table S6) did reveal the presence of additional viral PS genes. The PSII genes *psbA* and *psbD* (total = 6,877) far outnumbered the PSI genes that were previously noted in the marine environment including *psaA*, *psaB*, *psaC*, *psaD*, *psaE* and *psaK* (total = 371) (Table S7). Viral PSI genes were also noted in previous analyses of GOS microbial metagenomic data including





**Figure 4. Functional characterization of Indian Ocean viral sequences from the viral and larger fractions of metagenomic data in the context of KEGG pathways.** The inset pie charts represent the breakdown of the Metabolism super-pathway (top) and Energy metabolism pathway (bottom). The percentages of viral sequences attributed to the Energy metabolism pathway are indicated on the Metabolism pie charts. doi:10.1371/journal.pone.0042047.g004

17 samples collected from the Indian Ocean [14,15]. It is hypothesized that viral PSI components may facilitate electron donation from alternative sources other than plastocyanin to the PSI of their hosts, thereby increasing ATP generation for replication [14]. The discrepancy in the abundance of viral PSII versus PSI genes in the Indian Ocean data suggests that cyanophage may benefit more from carrying PSII genes, which have been shown to supplement photosynthesis in culture [60,61].

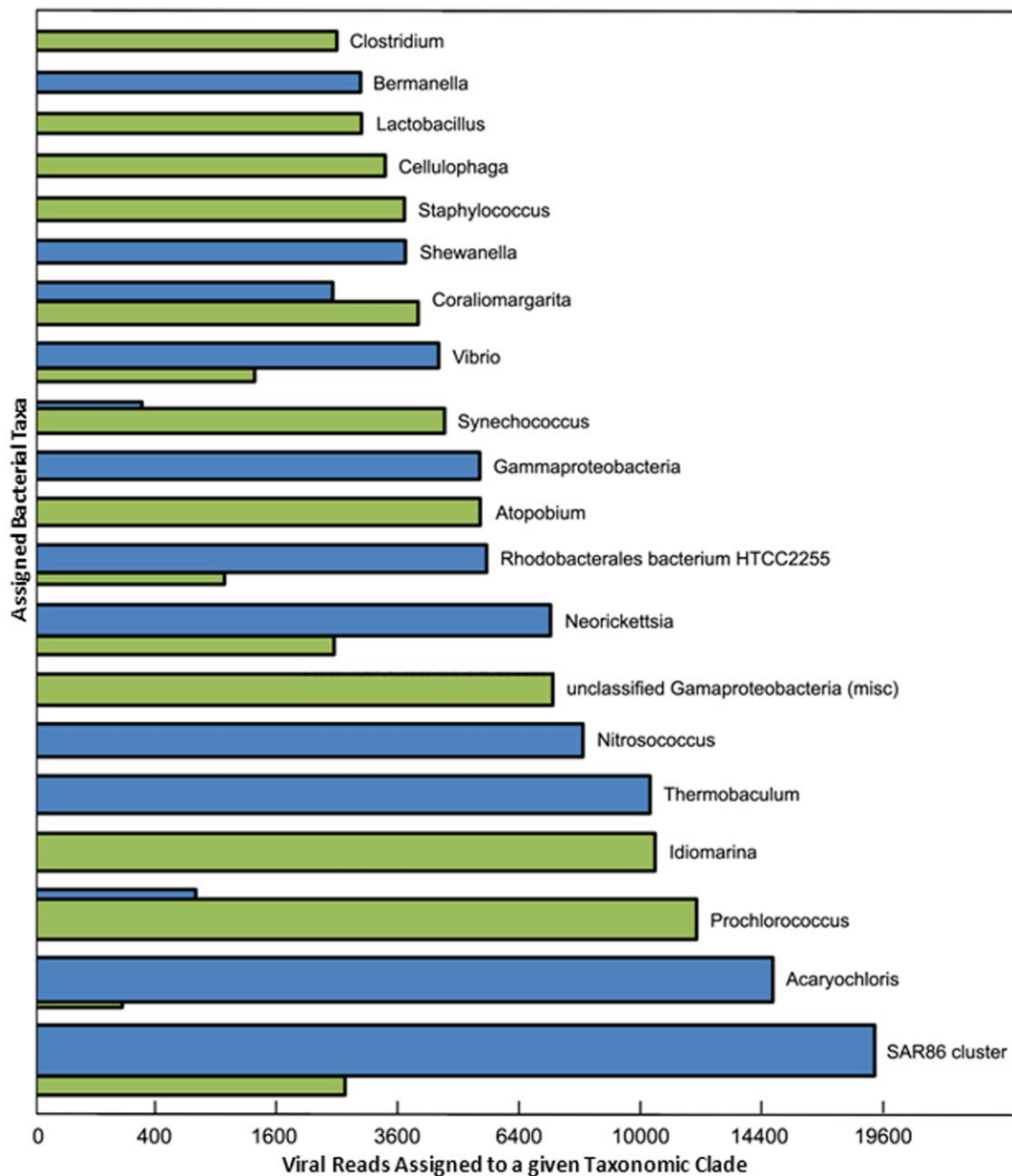
Carbonic anhydrase (CA) enzymes were also present in the viral data (VF = 24; LF = 1). CA is responsible for catalyzing the conversion of carbonic acid to CO<sub>2</sub> and is utilized by a diverse group of marine phytoplankton as part of their inorganic carbon concentration mechanism (CCM) to support photosynthetic carbon fixation [62–64]. Again, this is the first report of viral CAs in the marine environment to the best of our knowledge. Together with the PSI and PSII genes, the presence of these viral CAs suggests that the repertoire of viral-encoded photosynthesis-related genes is broader than previously recognized.

#### Identification of potential virus-host relationships

There is a growing body of evidence based on analyses of fully sequenced viral genomes that phage tend to mimic the polynucleotide sequence composition of their hosts [65–68]. On this evidence, we developed the first classification method to predict the taxonomy of bacterial hosts for uncharacterized viral metagenomic sequences that does not rely on homology or

sequence alignment; rather it compares sequence composition between viruses and bacteria. Briefly, our MGTAXA method involves three steps [69]: (I) trains one Glimmer Interpolated Context Model (ICM) [70] for every taxonomic node represented by at least one available bacterial reference sequence; (II) scores each metagenomic viral sequence against all models; and (III) picks the model with the highest score as representing the taxonomy of the putative host. This follows the approach of the Phymm bacterial classifier [21] where it was used to assign taxonomy to bacterial metagenomic sequences. The benchmarking of this classification scheme on temperate phages resulted in 97% accuracy at the phylum level and 89% at the genus level of the predicted host taxonomy with 2% rejection rate of the testing samples, and on all phages regardless of replication strategy - in 76% accuracy at the phylum level and 50% at the genus level with 6% rejection rate (see Methods S1, for detailed method description and benchmarking protocol). If the method were selecting candidate hosts from all bacterial genomes in RefSeq in an unbiased way, a randomly generated genus-level assignment would have an average accuracy of 0.2%.

Two approaches were taken to predict the bacterial hosts: 1) direct assignment – the ICMs trained on RefSeq (NCBI) genomes and 2) a transitive assignment – the hosts were predicted using ICMs trained on large metagenomic bacterial scaffolds (>100 kb) from the cellular fraction, to which a bacterial taxonomy was in turn assigned based on scoring against RefSeq ICMs (see Methods



**Figure 5. Bar chart demonstrating the predicted taxonomy of bacterial host genera for assembled Indian Ocean viral metagenomic data.** The area of each bar is proportional to the total number of viral reads contained in the contigs assigned to a particular host genus. Direct assignments are colored green and transitive assignments are colored blue. Direct assignments were made by assigning viral contigs to the host models built on NCBI RefSeq genomes. Transitive assignments were made by assigning viral contigs to models built on large bacterial contigs, and then assigning taxonomy to these bacterial contigs (as described in the text). Twenty top taxa in pairs of green and blue are shown according to the maximum count from either direct or transitive assignment method, respectively. If no bar is shown for one of the methods, the count was zero. doi:10.1371/journal.pone.0042047.g005

S1, for details on cellular fraction contig selection and taxonomic assignment). The essence of the transitive assignment process is that it restricts the model set to the bacterial genomes that are most abundant in a given metagenome (at least those that assembled into large scaffolds) before using that model set for host assignment. The value of the transitive method is that it recruits viral contigs to microbial scaffolds obtained from the same environment. On the other hand, depending on the dynamics of the phage-host system as well as sequence variability, the hosts for the assembled viruses might not necessarily be the best assembled microbes. In that case, the direct assignments can still be more

accurate and informative. Thus, we consider these two approaches to function synergistically in the designation of virus:host classifications.

The method was subsequently applied to all 5 Kb or longer contigs assembled from the Indian Ocean 454 viral libraries. **Figure 5** demonstrates the results of MGTAXA predictions of putative host taxonomy for the viruses present in the Indian Ocean metagenomic dataset. We found that even when all RefSeq bacterial genomes were used for host prediction (i.e. direct assignment) there was a preferential selection of microbial taxa that are indigenous to the marine environment including

*Prochlorococcus* and *Idiomarina*. Assignments to NCBI unclassified Gammaproteobacteria (which encompasses diverse marine isolates including the uncultivated SAR86 cluster), and to a lesser degree, the SAR86 cluster itself were also prevalent [71]. This behavior further supports the validity of our host prediction method already demonstrated on the compiled benchmark. The transitive assignments redistributed the number of assigned viral reads per contig towards fewer overall bacterial genera, generally focusing on those that demonstrated the highest abundance in microbial metagenomic data (**Table S8**) such as the SAR86 cluster (119 scaffolds) and *Rhodobacterales* HTCC2255 (33 scaffolds). This result was expected by the design of the transitive assignment methodology. However, this trend was not universal, as demonstrated by the number of *Prochlorococcus* assignments for which 38 bacterial scaffolds were assigned (**Figure 5; Table S8**). The viral contigs that were assigned to these putative host taxa by the direct method were largely reassigned to several different bacterial genera by the transitive method (**Table S9**). Transitive assignments to *Candidatus Pelagibacter* and the SAR11 cluster were also sparse despite the prevalence of bacterial scaffolds attributed to these organisms (87 and 33 respectively) (**Table S8**). The top predicted host for Indian Ocean viral assemblages using transitive methodology was the SAR86 cluster. The SAR86 cluster encompasses a group of uncultivated, proteorhodopsin-containing Gammaproteobacteria [72] and can comprise a significant portion of the microbial communities in various marine environments [73,74]. Since the bacterial members of the SAR86 cluster remain wild, no phages infective for these organisms have been reported. However, it is unlikely that these microbes are completely resistant to viral infection. Indeed, our results indicate otherwise and suggest that SAR86 virus-host interactions prevailed in the tropical Indian Ocean at the time of sampling. Due to the lack of specific SAR86 virus isolates and their corresponding genomic information, none of our viral metagenomic data could be specifically assigned to these putative viruses through our taxonomic analyses. This limitation further highlights the benefit of the homology-independent predictions of host taxonomy by revealing previously unknown, yet potentially significant virus-host relationships. Cyanobacteria within the *Acaryochloris* genus became the second most abundant predicted host for viruses in the Indian Ocean through transitive assignment despite the fact that only 10 bacterial scaffolds were assigned to this organism. *Acaryochloris* exists as either a free-living organism or as a symbiont of higher organisms (including macroalgae and ascidians) and is unique among the cyanobacteria since its main photosynthetic pigment is Chl*d* rather than Chl*a* [75,76]. Two *Acaryochloris* phage have also been isolated and their genomes sequenced and recently described by Chan and coworkers [77] who documented the unique presence of mitochondrial DNA polymerase. The authors of this study found homologs of this phage-encoded gene in the GOS microbial data, including the Indian Ocean, further suggesting that these cyanophage (and their hosts) are present in this environment. It's possible that these viruses were more abundant in the Indian Ocean at the time of sampling than indicated by our analyses with related sequences not receiving a definitive taxonomic assignment as discussed previously.

## Summary

This is the first study to examine the Indian Ocean virome using holistic metagenomic approaches. Since our analyses were not constrained to the “viral fraction” of samples, we were able to gain a much more comprehensive understanding of virus diversity, total gene complement and functional potential as well as virus-host relationships. Significant taxonomic differences were evident

between viruses represented in the VF versus the LF. An enrichment of cyano-myoviruses and viruses that likely infect eukaryotic phytoplankton or heterotrophic protists was found in the LF; while podo- and siphoviruses were prevalent in the VF. Similarly, notable differences in functional potential were evident by the distribution of abundances within metabolic pathways. The presence of putative viral genes potentially involved in nitrogen metabolism, carbon fixation and oxidative phosphorylation suggests that viruses infecting autotrophic and heterotrophic microbes may influence host cell physiology through diverse mechanisms in the Indian Ocean. Predicted virus-host relationships suggest that members of the SAR86 cluster and the cyanobacteria *Acaryochloris* and *Prochlorococcus* represent the dominant hosts for viruses in the Indian Ocean, providing insight into the types of viruses that putatively possess these metabolic capabilities.

## Materials and Methods

### Sample and metadata collection & size fractionation

Surface water samples (~400 L) were collected from 17 sites from the tropical Indian Ocean between August and October, 2005 aboard the S/V Sorcerer II. Two additional sites, GS148 and GS149 (~200 L), were sampled off the island of Zanzibar, Tanzania using alternate vessels. All field studies conducted within the EEZ of foreign nations received Marine Research Permits as required under the U.N. Convention on Law of the Sea, and as required, separate agreements to access genetic resources. The locations were not privately owned or protected in any way and the field studies did not involve endangered or protected species. A YSI (model 6600) was used to measure the physical environmental parameters including water temperature, salinity, dissolved oxygen and sample depth. Sub-samples were collected for dissolved nutrient analyses as described previously [23] and were processed by the Virginia Institute of Marine Sciences (VIMS) Analytical Service Center. The microbial community was first passed through a 20 µm Nytex pre-filter and then size fractionated by serial filtration through 3.0 µm, 0.8 µm and 0.1 µm membrane filters (Pall Life Sciences, East Hills, NY). Filters were preserved and stored as described previously [23].

### Virus concentration & purification

The viral fraction of water samples (i.e. <0.1 µm) was concentrated using tangential flow filtration (TFF) as described previously [22]. Viral concentrates (VCs) were cryo-preserved through the addition of glycerol (10% final concentration) and frozen at -20°C. VCs were transferred to and stored at -80°C upon return to the J. Craig Venter Institute (JCVI) until further processed. VCs were further concentrated, treated with nuclease to remove free DNA and pelleted through a sucrose cushion as previously described [22]. To check for cellular contamination a 16S rRNA gene PCR was used with positive (*E. coli* cells) and negative (DEPC water) samples. Gel electrophoresis was performed on 2 µl (of 50 µl total reaction volume) on a 0.8% agarose gel stained with SybrGold (Invitrogen). If no discernible 16S rRNA gene product (~1500 bp) was visualized the samples were further processed through DNA extraction.

### DNA extraction, library construction, sequencing and post-processing of data

Methods describing DNA extraction from filters, construction of clone libraries, template preparation, and automated cycle sequencing can be found in Rusch *et al.*, 2007. Viral DNA was extracted from purified VCs and modified linker amplified shotgun libraries (LASLs) were constructed as described previously

[22]. In addition to Sanger libraries, 454 Titanium libraries were prepared from amplified viral DNA (LASLs) and pyrosequenced at JCVI. Briefly, genomic DNA was fragmented and size-selected to a range of 500–800 bp. Linkers were ligated to DNA fragments for use as priming sites during subsequent amplification reactions. Three replicate amplification reactions were completed using 15 total cycles to reduce biases and amplification of potential cellular contamination from kit reagents. Adaptors were ligated onto the fragments and used as priming sites for emulsion PCR. Amplified samples were purified using AMPure beads to remove small DNA fragments and sequencing was performed using the 454 GS FLX Titanium sequencing platform. Viral metagenomic sequences were trimmed of any linker sequence left over during LASL production. Additionally, artificial replicates were screened for and removed from all 454 data using an approach described by Gomez-Alvarez (2009) [78].

### Identification of viral sequences in the filter fractions of data & taxonomic profiling

The Automated Phylogenetic Inference System (APIS) was used for the taxonomic classification of viral predicted proteins as well as the identification of viral proteins within the larger size classes of data [25]. APIS automated the process of calculating sequence similarity, alignment, and phylogenetic inference for each protein in a given dataset. Each predicted protein was compared to an in-house curated database (phyloDB), which consists of proteins from complete (or nearly complete) genomes and selected Sanger-sequenced viral metagenomes, using BLASTP. Full-length sequences of significant BLASTP hits were retrieved and then a multiple alignment was generated using MUSCLE. From this alignment, a neighbor-joining tree was produced using QuickTree [79] to determine the phylogenetic placement of the query sequence by comparing the taxonomic classification of the sequence(s) that clade with the query. If the taxonomic information differed among these clading sequences, this was noted and the classification of the query was limited to the higher taxonomic ranks where they were in agreement. To identify viral sequences from the larger size class (0.1–20  $\mu\text{m}$ ) of organisms, all proteins classified via APIS as viral at the Kingdom level were used in further analyses.

A BLAST-based approach was used, which is part of the Viral Metagenome Annotation Pipeline (VMGAP) [80]. Environmental sequences were compared against NCBI AllGroup.niaa database (BLASTP) using an e-value of  $\leq 1e-10$  and identity  $\geq 50\%$ , the top hit was noted and taxonomic information transferred to each metagenomic protein.

### Functional characterization of viral sequences

**ORF prediction and viral annotation.** Open reading frames were predicted as described previously [81], and were based on a combination of naïve 6-frame translations and MetaGeneAnnotator [82], an *ab initio* gene finder program. The predicted protein coding sequences were then annotated using the Viral Metagenome Annotation Pipeline (VMGAP), described by Lorenzi and coworkers [80].

**Annotation of protein clusters and KEGG classification.** ORFs predicted on 454 sequence reads were mapped to protein clusters as follows. RPS-BLAST was first used to compare each protein sequence against a database of Position Specific Scoring Matrices (PSSM) representing clusters containing more than 20 proteins. Then, proteins were assigned to clusters based on the PSSM that gave the highest bit score with an e-value  $\leq 1 \times 10^{-3}$ . For those protein sequences that did not produce any significant hit against the PSSM database, a BLASTP search was conducted against a database of proteins belonging to clusters

containing fewer than 20 members. Proteins were then assigned to clusters based on the hit having the lowest e-value with at least 40% identity and 70% coverage. Protein sequences that did not result in any significant hit remained unassigned.

Annotation was performed on all protein clusters [33] from the individual predicted protein annotations within each cluster. All protein annotations were counted within, and across, all clusters and an uncorrected p-value was calculated using Fisher's exact test indicating the probability of random association of each annotation with each cluster. The annotation, description, coverage and p-value were given for each of the 3 best annotations. For clusters that did not receive annotation by this method, a different strategy was taken based on additional searches and relative ranking of the results (Methods S1, for details). To bin protein clusters into KEGG classes, all cluster names having a match that was 100% identical to any KEGG class definition were assigned to the corresponding three levels of classification from the KEGG Pathway Database. When more than one possible classification class were available at a particular level, all classes representing  $<75\%$  of that level were labeled "unspecific". Clusters without any direct KEGG association were examined for the presence of specific keywords and then classified following the same three-level classification. Otherwise, clusters were binned as unclassified for each level.

**Large subunit terminase phylogeny.** Environmental sequences and references were retrieved from two proteins clusters, both containing putative genes encoding terminase enzymes. The average length of all sequences was calculated and sequences below 40% of the average were removed. Non-viral GOS Phase I sequences were reduced through clustering and only the representative sequence used in subsequent steps. Existing hidden Markov model (HMM) profiles, including fragment HMMs, obtained in cluster annotation or peptide annotation were identified and the HMM that accounted for the majority of viral sequences was selected for alignment using HMMER (<http://hmmer.org/>). The multiple alignments were processed through a gap-filtering step to remove sequences that contain 60% or more gaps. Phylogenetic trees were constructed using the program FastTree [83] and processed using Archaeopteryx (<http://phylosoft.org/archaeopteryx/>).

### Statistical analyses

Principal Component Analysis (PCA) and Canonical Component Analysis (CCA) were performed using the R statistical program. For PCA, library proportional abundances of viral sequences for each protein cluster were calculated and used to build the centered matrix. For CCA, oceanographic metadata was included to assess correlation between these and protein cluster proportional abundances.

### Metagenomic assembly

The Sanger and 454 data from each viral library were assembled independently using the Newbler GS De Novo Assembler, version 2.3, with a minimum identity threshold of 86%. In addition to virus-specific assemblies, a comprehensive global assembly was also conducted on all Sanger reads produced from Phase I of GOS and all Indian Ocean size fractions as described by Rusch et al. 2007 using the Celera Assembler version 5.3 with a minimum identity threshold of 86%. A minimum overlap length of 40 bp was used for all assemblies.

### Genotype diversity estimates

A local command line version of the PHACCS program [26] was used to estimate viral genotypic diversity. The 454 datasets were randomly sub-sampled 8–10 times prior to assembly to approximate the same amount of base pairs as the corresponding

Sanger datasets and the resulting contig spectrums were used as input. The average genome size used was 50 Kbp for all, except where the 454-subsampling used additional genome sizes to assess the effect on diversity estimates, and the minimum overlap was 40 bp. Averages and standard deviations were calculated in Excel.

### Virus-host classification

As this is the first report of an alignment-independent method that aims to predict the virus:host relationships in the metagenomic datasets, detailed methodology and benchmarking is reported in Methods S1. Briefly, sequence composition similarities were used to bin viral contigs with 1) bacterial hosts from RefSeq and 2) the cellular fraction metagenomic sequences corresponding to this dataset. Hosts were described at the genus taxonomic level; if the lowest level reached for assignment was order or higher, the viral contigs were labeled as ‘unassigned’. The implementation parallelized for a distributed computing cluster is available as part of our open-source software package MGTAXA [69].

### Accession numbers

All Sanger-generated viral data was submitted to the NCBI Trace Archives. These include the raw reads from viral metagenomes as well as the predicted protein sequences extracted from the larger size fractions of data. All 454-generated viral data was submitted to the NCBI Short Read Archive (SRA), corresponding to accession numbers SRX096024, SRX096023, SRX096025, and SRX096299. The microbial and viral assemblies used for the analysis of virus-host classification were submitted under NCBI's Project ID 19733.

### Supporting Information

**Methods S1 Additional methods including, i) Alternative cluster annotation, ii) Identification of viral photosynthesis genes, iii) Classification of Virus:Host relationships.**  
(DOCX)

**Figure S1 Heatmap showing delineation of viral sequences from large vs. viral fraction.** Abundance of sequences from each site within selected functional groups at the level-2 classification of Gene Ontology (GO) was used to generate heatmap. Hierarchical clustering of sites indicates a grouping of large (blue) versus viral (green) fraction. Functional groups are color-coded as follows: cellular processes (purple), cellular processes and signaling (blue), environmental information processing (green), genetic information processing (yellow), metabolism (orange), phage structure (red).  
(TIF)

**Table S1 Indian Ocean sample characteristics and oceanographic metadata.**  
(XLSX)

**Table S2 Viral sequences retrieved from the larger size classes.**  
(XLSX)

### References

1. Wommack KE, Colwell RR (2000) Virioplankton: viruses in aquatic ecosystems. *Microbiol Mol Biol Rev* 64: 69–114.
2. Suttle CA (2007) Marine viruses - major players in the global ecosystem. *Nature Rev Microbiol* 5: 801–812.
3. Fuhrman JA (1999) Marine viruses and their biogeochemical and ecological effects. *Nature* 399(6736): 541–548.
4. Suttle CA (2005) Viruses in the sea. *Nature* 437: 7057 356–61
5. Rohwer F, Thurber RV (2009) Viruses manipulate the marine environment. *Nature* 459: 207–212.
6. Breitbart M, Salamon P, Andresen B, Mahaffy JM, Segall AM, et al. (2002) Genomic analysis of uncultured marine viral communities. *Proc Natl Acad Sci USA* 99: 14250–14255.
7. Bench SR, Hanson TE, Williamson KE, Gosh D, Radosovich M, et al. (2007) Metagenomic characterization of Chesapeake Bay virioplankton. *Appl Environ Microbiol* 73: 7629–7641.

**Table S3 Assembly statistics for Indian Ocean viral metagenomic libraries.**  
(XLSX)

**Table S4 Top five most abundant cyanophage homologs based on taxonomic assignments.**  
(XLSX)

**Table S5 Functional characterization of VF and LF viral sequences by KEGG category.**  
(XLSX)

**Table S6 List of reference sequences and genomes used to identify Indian Ocean viral photosystem genes.**  
(XLSX)

**Table S7 Viral photosystem genes present in the Indian Ocean metagenomic data.**  
(XLSX)

**Table S8 GOS Bacterial Scaffold Taxonomic Assignments based on MGTAXA.**  
(XLSX)

**Table S9 Re-distribution of viral contigs to bacterial taxa from direct to transitive assignments.**  
(XLSX)

**Table S10 NCBI RefSeq prokaryotic genomes used for building reference models.**  
(XLS)

**Table S11 Benchmarking accuracy for predicting the host taxonomy.**  
(XLSX)

**Table S12 Benchmarking set of virus:host pairs compiled from NCBI RefSeq.**  
(XLS)

**Table S13 Summary counts for the benchmark virus:host pairs in Table S12.**  
(XLSX)

### Acknowledgments

We would like to deeply thank Jeff Hoffman for sample and metadata collection and processing, Cynthia Andrews-Pfannkoch for DNA extraction from filters and construction of metagenomic clone libraries, Johannes Goll for his assistance with METAREP; Jeff Hoover for his assistance with the viral metagenomic annotation pipeline; Harald Kattnig for his contribution to phylogenetic tree building, Doug Rusch for his contribution to selecting bacterial scaffolds for MGTAXA host classification and Jasmine Pollard for her assistance with figure preparation.

### Author Contributions

Conceived and designed the experiments: SJW JCV. Performed the experiments: SJW LZA DWF. Analyzed the data: SJW LZA HAL MT DB JPM AT SY. Contributed reagents/materials/analysis tools: SJW LZA HAL JPM AT SY. Wrote the paper: SJW LZA HAL AT SY.

8. Angly FE, Felts B, Breitbart M, Salamon P, Edwards RA, et al. (2006) The marine viromes of four oceanic regions. *PLoS Biol* 4: e368.
9. Breitbart M, Felts B, Kelley S, Mahaffy JM, Nulton J, et al. (2004) Diversity and population structure of a near-shore marine-sediment viral community. *Proc Royal Soc London Bio* 271: 565–574.
10. 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.
11. Dinsdale EA, Edwards RA, Hall D, Angly F, Breitbart M, et al. (2008) Functional metagenomic profiling of nine biomes. *Nature* 455: 629–632.
12. Williamson SJ, Rusch DB, Yooshep S, Halpern AL, Heidelberg KB, et al. (2008) The Sorcerer II Global Ocean Sampling Expedition: metagenomic characterization of viruses within aquatic microbial samples. *PLoS ONE* 3: 1 e1456.
13. Sharon I, Tzahor S, Williamson S, Shmoish M, Man-Aharonovich D, et al. (2007) Viral photosynthetic reaction center genes and transcripts in the marine environment. *The ISME J* 1: 492–501.
14. Sharon I, Alperovitch A, Rohwer F, Haynes M, Glaser F, et al. (2009) Photosystem I gene cassettes are present in marine virus genomes. *Nature* 461: 258–262.
15. Sharon I, Battchikova N, Aro EM, Giglione C, Meinel T, et al. (2011) Comparative metagenomics of microbial traits within oceanic viral communities. *The ISME J* 5: 1178–1190.
16. McDaniel L, Breitbart M, Moberley J, Long A, Haynes M, et al. (2008) Metagenomic analysis of lysogeny in Tampa Bay: implications for prophage gene expression. *PLoS ONE* 3(9): e3263.
17. Williamson SJ, Cary SC, Williamson KE, Helton RR, Bench SR, et al. (2008) Lysogenic virus-host interactions predominate at deep-sea diffuse-flow hydrothermal vents. *The ISME J* 2: 1112–1121.
18. 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.
19. Clokie MRJ, Millard AD, Mehta JY, Mann NH (2006) Virus isolation studies suggest short-term variations in abundance in natural cyanophage populations of the Indian Ocean. *J Mar Biol Assoc UK* 86: 499–505.
20. Massana R, del Campo J, Dinter C, Sommaruga R (2007) Crash of a population of the marine heterotrophic flagellate *Cafeteria roenbergensis* by viral infection. *Environ Microbiol* 9(11): 2660–2669.
21. Brady A, Salzberg SL (2009) Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. *Nat Methods* 6(9): 673–676.
22. Andrews-Pfannkoch C, Fadrosch DW, Thorpe J, Williamson SJ (2010) Hydroxyapatite-Mediated Separation of Double-Stranded DNA, Single-Stranded DNA, and RNA Genomes from Natural Viral Assemblages. *Appl Environ Microbiol* 76: 5039–5045.
23. Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson SJ, et al. (2007) The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biol* 5: e77.
24. Badger JH, Eisen JA, Ward NL (2005) Genomic analysis of *Hyphomonas neptunium* contradicts 16S rRNA gene-based phylogenetic analysis: implications for the taxonomy of the orders 'Rhodobacterales' and 'Caulobacterales'. *Int J Syst Evol Microbiol* 55: 1021–1026.
25. Allen LZ, Allen EE, Badger JH, McCrow JP, Paulsen IT, et al. (2012) Influence of nutrients and currents on the genomic composition of microbes across an upwelling mosaic. *The ISME J* 6: 1403–1414.
26. 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.
27. Sandaa RA (2008) Burden or benefit? Virus-host interactions in the marine environment. *Research in microbiology* 159: 374–381.
28. Steward GF, Montiel JL, Azam F (2000) Genome size distributions indicate variability and similarities among marine viral assemblages from diverse environments. *Limnol and Oceanogr* 45:1697–1706.
29. Raoult D, Audic S, Robert C, Abergel C, Renesto P, et al. (2004) The 1.2-megabase genome sequence of mimivirus. *Science* 306: 1344–1350.
30. Monier A, Larsen JB, Sandaa RA, Bratbak G, Claverie JM, et al. (2008) Marine mimivirus relatives are probably large algal viruses. *Virology* 5: 12.
31. Ghedin E, Claverie JM (2005) Mimivirus relatives in the Sargasso sea. *Virology* 2: 62.
32. Claverie JM, Grzela R, Lartigue A, Bernadac A, Nitsche S, et al. (2009) Mimivirus and Mimiviridae: Giant viruses with an increasing number of potential hosts, including corals and sponges. *J Invertebr Pathol* 101: 172–180.
33. Yooshep S, Sutton G, Rusch DB, Halpern AL, Williamson SJ, et al. (2007) The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. *PLoS biology* 5: e16.
34. Rao VB, Feiss M (2008) The Bacteriophage DNA Packaging Motor. *Annual review of genetics* 42: 647–681.
35. Draper B, Rao VB (2007) An ATP hydrolysis sensor in the DNA packaging motor from bacteriophage T4 suggests an inchworm-type translocation mechanism. *J Mol Biol* 369: 79–194.
36. Mitchell MS, Matsuzaki S, Imai S, Rao VB (2002) Sequence analysis of bacteriophage T4 DNA packaging/terminase genes 16 and 17 reveals a common ATPase center in the large subunit of viral terminases. *Nucleic Acids Res* 30: 4009–4021.
37. Casjens SR, Gilcrease EB, Winn-Stapley DA, Schickmaier P, Schmiegler H, et al. (2005) The generalized transducing *Salmonella* bacteriophage ES18: Complete genome sequence and DNA packaging strategy. *J Bacteriol* 187: 1091–1104.
38. Filee J, Pouget N, Chandler M (2008) Phylogenetic evidence for extensive lateral acquisition of cellular genes by Nucleocytoplasmic large DNA viruses. *BMC Evo Biol* 8: 320.
39. Filee J, Siguier P, Chandler M (2007) I am what I eat and I eat what I am: acquisition of bacterial genes by giant viruses. *Trends Genet* 23: 10–15.
40. Fischer MG, Allen MJ, Wilson WH, Suttle CA (2010) Giant virus with a remarkable complement of genes infects marine zooplankton. *Proc Natl Acad Sci USA* 107: 19508–19513.
41. Wilson WH, Schroeder DC, Allen MJ, Holden MTG, Parkhill J, et al. (2005) Complete genome sequence and lytic phase transcription profile of a Coccolithovirus. *Science* 309: 1090–1092.
42. Baumann S, Sander A, Gurmon JR, Yanai-Balser GM, Van Etten JL, et al. (2007) Chlorella viruses contain genes encoding a complete polyamine biosynthetic pathway. *Virology* 360: 209–217.
43. DeAngelis PL, Jing W, Graves MV, Burbank DE, VanEtten JL (1997) Hyaluronan synthase of chlorella virus PBCV-1. *Science* 278: 1800–1803.
44. Delaroque N, Muller DG, Bothe G, Pohl T, Knippers R, et al. (2001) The complete DNA sequence of the *Ectocarpus siliculosus* virus EsV-1 genome. *Virology* 287: 112–132.
45. Landstein D, Graves MV, Burbank DE, DeAngelis P, Van Etten JL (1998) Chlorella virus PBCV-1 encodes functional glutamine: Fructose-6-phosphate amidotransferase and UDP-glucose dehydrogenase enzymes. *Virology* 250: 388–396.
46. Wilson WH, Van Etten JL, Allen MJ (2009) The Phycodnaviridae: the story of how tiny giants rule the world. *CTMI* 328: 1–42.
47. Van Etten JL, Lane LC, Dunigan DD (2010) DNA Viruses: The Really Big Ones (Giruses). *Annu Rev of Microbiol* 64: 83–99.
48. Moreau H, Piganeau G, Desdevises Y, Cooke R, Derelle E, et al. (2010) Marine Prasinovirus Genomes Show Low Evolutionary Divergence and Acquisition of Protein Metabolism Genes by Horizontal Gene Transfer. *J Virol* 84: 12555–12563.
49. Muro-Pastor MI, Reyes JC, Florencio FJ (2005) Ammonium assimilation in cyanobacteria. *Photosynth Res* 83: 135–150.
50. Moore JK, Doney SC, Lindsay K (2004) Upper ocean ecosystem dynamics and iron cycling in a global three-dimensional model. *Global Biogeochem Cy* 18: 4.
51. Alperovitch-Lavy A, Sharon I, Rohwer F, Aro EM, Glaser F, et al. (2011) Reconstructing a puzzle: existence of cyanophages containing both photosystem-I and photosystem-II gene suites inferred from oceanic metagenomic datasets. *Environ Microbiol* 13: 24–32.
52. Bratbak G, Egge JK, Heldal M (1993) Viral mortality of the marine alga *Emiliania-Huxleyi* (Haptophyceae) and termination of algal blooms. *Mar Ecol-Prog Ser* 93: 39–48.
53. Angly F, Youle M, Nosrat B, Srinagesh S, Rodriguez-Brito B, et al. (2009) Genomic analysis of multiple Roseophage SIO1 strains. *Environ Microbiol* 11: 2863–2873.
54. Sullivan MB, Coleman ML, Weigle P, Rohwer F, Chisholm SW (2005) Three *Prochlorococcus* cyanophage genomes: signature features and ecological interpretations. *PLoS Biol* 3: e144.
55. Casas V, Kline DI, Wegley L, Yu YN, Breitbart M, et al. (2004) Widespread association of a Rickettsiales-like bacterium with reef-building corals. *Environ Microbiol* 6: 1137–1148.
56. 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.
57. Lindell D, Sullivan MB, Johnson ZI, Tolonen AC, Rohwer F, et al. (2004) Transfer of photosynthesis genes to and from *Prochlorococcus* viruses. *Proc of the Natl Acad Sci USA* 101: 11013–11018.
58. Mann NH, Cook A, Millard A, Bailey S, Clokie M (2003) Marine ecosystems: Bacterial photosynthesis genes in a virus. *Nature* 424: 741–741.
59. Mann NH, Clokie MRJ, Millard A, Cook A, Wilson WH, et al. (2005) The genome of S-PM2, a "photosynthetic" T4-type bacteriophage that infects marine *Synechococcus* strains. *J Bacteriol* 187: 3188–3200.
60. Lindell D, Jaffe JD, Johnson ZI, Church GM, Chisholm SW (2005) Photosynthesis genes in marine viruses yield proteins during host infection. *Nature* 438: 86–89.
61. Clokie MR, Shan J, Bailey S, Jia Y, Krusch HM, et al. (2006) Transcription of a 'photosynthetic' T4-type phage during infection of a marine cyanobacterium. *Environ Microbiol* 8: 827–835.
62. Seitzinger S, Harrison JA, Bohlke JK, Bouwman AF, Lowrance R, et al. (2006) Denitrification across landscapes and waterscapes: a synthesis. *Ecol Appl* 16: 2064–2090.
63. Badger MR, Price GD, Long BM, Woodger FJ (2006) The environmental plasticity and ecological genomics of the cyanobacterial CO<sub>2</sub> concentrating mechanism. *J Exp Bot* 57: 249–265.
64. Reinfelder JR (2011) Carbon Concentrating Mechanisms in Eukaryotic Marine Phytoplankton. *Annu Rev Mar Sci* 3: 291–315.
65. 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.

66. Lucks JB, Nelson DR, Kudla GR, Plotkin JB (2008) Genome landscapes and bacteriophage codon usage. *Plos Comput Biol* 4: e1000001.
67. Ahmed S, Saito A, Suzuki M, Nemoto N, Nishigaki K (2009) Host-parasite relations of bacteria and phages can be unveiled by oligostickness, a measure of relaxed sequence similarity. *Bioinformatics* 25: 563–570.
68. Robins H, Krasnitz M, Barak H, Levine AJ (2005) Relative-entropy algorithm for genomic fingerprinting captures host-phage similarities. *J Bacteriol* 187: 8370–8374.
69. MGTAXA website. Available at: <http://andreyto.github.com/mgtaxa/>. Accessed 2011 May.
70. Delcher AL, Harmon D, Kasif S, White O, Salzberg SL (1999) Improved microbial gene identification with GLIMMER. *Nucleic Acids Res* 27: 4636–4641.
71. Dupont CL, Rusch DB, Yooshep S, Lombardo MJ, Richater AR, et al. (2011) Genomic insights to SAR86, an abundant and uncultivated marine bacterial lineage. *The ISME J* 6: 1186–1199.
72. Sabehi G, Beja O, Suzuki MT, Preston CM, DeLong EF (2004) Different SAR86 subgroups harbour divergent proteorhodopsins. *Environ Microbiol* 6: 903–910.
73. Eilers H, Pernthaler J, Glockner FO, Amann R (2000) Culturability and in situ abundance of pelagic bacteria from the North Sea. *Appl Environ Microbiol* 66: 3044–3051.
74. Beman JM, Steele JA, Fuhrman JA (2011) Co-occurrence patterns for abundant marine archaeal and bacterial lineages in the deep chlorophyll maximum of coastal California. *The ISME J* 5: 1077–1085.
75. Chan YW, Nenninger A, Clokie SJH, Mann NH, Scanlan DJ, et al. (2007) Pigment composition and adaptation in free-living and symbiotic strains of *Acaryochloris marina*. *FEMS Microbiol Ecol* 61: 65–73.
76. Kuhl M, Chen M, Ralph PJ, Schreiber U, Larkum AWD (2005) A niche for cyanobacteria containing chlorophyll d. *Nature* 433: 820.
77. Chan YW, Mohr R, Millard AD, Holmes AB, Larkum AW, et al. (2011) Discovery of Cyanophage Genomes Which Contain Mitochondrial DNA Polymerase. *Mol Biol and Evol* 28: 2269–2274.
78. Gomez-Alvarez V, Teal TK, Schmidt TM (2009) Systematic artifacts in metagenomes from complex microbial communities. *The ISME J* 3: 1314–1317.
79. Howe K, Bateman A, Durbin R (2002) QuickTree: building huge Neighbour-Joining trees of protein sequences. *Bioinformatics* 18: 1546–1547.
80. Lorenzi HA, Hoover J, Inman J, Safford T, Murphy S, et al. (2011) The Viral MetaGenome Annotation Pipeline (VMGAP): An automated tool for the functional annotation of viral Metagenomic shotgun sequencing data. *Stand Genomic Sci* 4: 418–429.
81. Tanenbaum DM, Goll J, Murphy S, Kumar P, Zafar N, et al. (2010) The JCVI standard operating procedure for annotating prokaryotic metagenomic shotgun sequencing data. *Stand Genomic Sci* 2: 229–237.
82. Noguchi H, Taniguchi T, Itoh T (2008) MetaGeneAnnotator: Detecting species-specific patterns of ribosomal binding site for precise gene prediction in anonymous prokaryotic and phage genomes. *DNA Res* 15: 387–396.
83. Price MN, Dehal PS, Arkin AP (2010) FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS ONE* 5: e9490.