Supporting Information

Kang et al. 10.1073/pnas.1219930110

SI Text

Host Range of HMO-2011. The 16S rRNA gene sequence comparison using the EzTaxon database (1) showed that strain IMCC1322 shared more than 90% sequence similarity with the following type strains of six species in the order *Rhodospirillales*: *Nisaea nitritireducens* DR41_18^T (92.3%), *Nisaea denitrificans* DR41_21^T (91.8%), *Thalassobaculum salexigens* CZ41_10a^T (91.2%), *Thalassobaculum litoreum* CL-GR58^T (90.9%), *Oceanibaculum indicum* P24^T (90.8%), and *Oceanibaculum pacificum* MC2UP-L3^T (90.6%). All these strains were obtained from culture collections (Leibniz Institute DSMZ – German Collection of Microorganisms and Cell Cultures or Korean Collection for Type Cultures) and were used to determine the host range of HMO-2011 by spot plaque assays. Bacterial strains were grown in marine broth 2216 (hereafter referred to as MB; Difco) at 25 °C until the OD_{600} reached 0.3–0.5. One milliliter of culture was mixed with 5 mL of molten top agar (MB with 0.5% Bacto agar) and poured onto bottom agar plates (MB with 1.5% Bacto agar). After 20–30 min, 10 μL of phage stock (~3×10^7 pfu·ml^{-1}) was spotted onto the plates. Plates were incubated at 25 °C for 2–3 d before plaque formation was checked. The spot assay showed that HMO-2011 did not infect any type strains of the six species tested.

ORFs Predicted to Encode Proteins for DNA Replication and Metabolism. Four ORFs encoded proteins involved in DNA replication, including a primase (ORF10), helicase (ORF11), DNA polymerase (ORF12), and endonuclease (ORF15). A ribonucleotide reductase (ORF18) and an integrase (ORF9) were also predicted. Although these genes were only distantly related to genes of other isolated phages, the gene content and gene order in this module were similar to those of marine podoviruses such as cyanophages P-SSP7 and Syn5 and, especially roseophage SIO1 (2-4).

The primase and helicase, which are typically encoded as a single protein in many phages such as T7, were separately encoded by ORF10 and ORF11, respectively, in HMO-2011. The replicative DNA helicase encoded by ORF11 had conserved residues in the catalytic domain of DnaB-like helicases, including the Walker A and B motifs (5). However, the DnaG-type primase (ORF10) had unusual spacing between conserved residues in both the zinc-binding domain and the Toprim domain, two conserved domains found in prokaryotic primases (Fig. S2*A*) (6). Similar features were found in primases predicted from several phages and many marine metagenome sequences (Fig. S2*A*) (7).

The phage endonuclease I encoded by ORF15 was similar to gene 3 of phage T7, a Holliday junction resolvase, which resolves Holliday structures and branched DNA molecules produced during phage DNA replication.

ORF9 is predicted to encode a tyrosine integrase that might mediate the site-specific integration of the phage genome into the host chromosome. The deduced amino acid sequence has a catalytic tyrosine and four of five highly conserved residues in tyrosine integrases (8, 9), suggesting the functionality of the protein. Therefore, a search for a putative integration site was performed by comparing the genome sequences of HMO-2011 and IMCC1322 using BLASTN. No match ≥11 bp was found between host genome regions surrounding tRNA genes and phage genome regions near the integrase ORF. However, it is noteworthy that a long (35 bp) match was found between ORF61 of HMO-2011 and SAR116_2109 of IMCC1322 because the use of tRNA genes as integration sites is not universal (10). Plaques formed by HMO-2011 were always clear, indicating that the probability of lysogeny is very low even if the phage can be temperate.

ORFs Predicted to Encode Proteins for Structure, Packaging, and Lysis. Nineteen ORFs were predicted to encode structural proteins. A coat protein containing a P22 coat protein domain (PF11651) was encoded by ORF47, and, in a phylogenetic analysis, it formed a well-supported cluster with coat proteins predicted from the GOS metagenome sequences (Fig. S2*B*). A portal protein that connects the head and tail and functions as a passage for DNA was encoded by ORF51. Four ORFs (ORF39, ORF42, ORF44, and ORF65) were predicted to encode structural proteins based on similarities to other phage proteins that were identified in virus particles by mass spectrometry (Table S1). ORF31 was annotated as a structural gene because it encoded a C1q domain-containing protein and the C1q domain is known to be involved in interactions with bacterial surface structures (11, 12).

Among the 19 ORFs annotated as structural proteins, 12 were predicted to encode tail proteins. Annotation of tail proteins was mainly based on sequence similarities to tail structural proteins of other phages (Table S1), some of which have been experimentally verified as structural proteins by mass spectrometry. BLAST hits with low sequence similarity or low coverage were used to annotate some tail proteins, considering the mosaicism and divergence of tail fiber genes (3, 13). Paralogous relationships were also used to annotate some tail proteins (Fig. S3) (3). Five ORFs (ORFs 29, 32, 34, 37, and 40) had identical C-terminal ends, and ORF33 and ORF41 shared N-terminal ends (Fig. S3 *A* and *B*). In addition, some ORFs were similar to paralogous proteins of other phages or environmental fosmids. For example, ORF39 was similar to ORFs 37–39 of cyanophage P-SSP7 and to cds53, cds55, and cds57 of a fosmid from the Mediterranean Sea (Fig. S3*C*). These paralogies suggested that the tail proteins of HMO-2011 evolved through duplication followed by diversification of proteins or domains, an evolutionary mechanism previously described for the structural proteins of other phages (14, 15).

In tailed phages, packaging of DNA into preassembled capsids requires a portal protein and terminase. In HMO-2011, a portal protein is encoded by ORF51, and terminase, which translocates DNA via ATP hydrolysis, is encoded by ORF52 (large subunit) and ORF55 (small subunit). Lysis of host bacteria is mediated by the concerted action of two proteins, endolysin and holin. A putative endolysin is encoded by ORF56; however, a holin-encoding ORF was not predicted.

Search for a DnaJ Central Domain in Family A DNA Polymerases. HMO-2011 ORF12 encodes a DNA polymerase containing a partial DnaJ central domain in addition to a DNA polymerase family A domain. To determine whether other family A DNA polymerases in cultured organisms contain a DnaJ central domain, we performed conserved domain searches for DNA polymerases that were predicted to be similar to ORF12 by BLASTP, PSI-BLAST, or DELTA-BLAST analyses. Among the top 30 hits in each BLAST analysis, no proteins were found to have a DnaJ central domain. Next, the Conserved Domain Architecture Retrieval Tool was used (16), which searches for similar protein sequences based on domain architecture, and can retrieve all proteins sharing at least a single domain. No protein was found to have both a DnaJ central domain and a DNA polymerase family A domain. Finally, the Pfam protein families database was searched using PfamAlyzer, which can retrieve proteins with a combination of domains specified by the user (17). The results yielded no proteins with a DnaJ central domain adjacent to DNA polymerase family A domain.

Search for Putative DNA Polymerases in Metagenomes Having a Domain Architecture Similar to that of ORF12. To retrieve metagenome sequences putatively encoding a DNA polymerase similar to ORF12, TBLASTN was performed using ORF12 of HMO-2011 as a query against "All Metagenomic Sanger Reads" and "All Metagenomic 454 Reads" databases in CAMERA (18). The results were downloaded and used to search for ORF12 homologs. In brief, metagenome sequences were regarded as having a domain architecture similar to that of ORF12 if the alignments included an 87-amino acid region of ORF12 (from 294 to 380) and two repeats of the CXXCXGXG motif. One mismatch in each of the repeats was allowed. In other words, metagenome sequences with an N-terminal region of a DNA polymerase domain preceded by a partial DnaJ central domain (Fig. 3) were considered to have a domain architecture similar to that of ORF12. A total of 629 sequences retrieved from 142 metagenome samples satisfied the above criteria (Table S2). Most of the retrieved sequences were from diverse marine habitats encompassing the Pacific, Atlantic, Indian, and Southern Oceans. Among the 629 retrieved sequences, 489 were from pyrosequencing reads.

ORF61 as an AMG. Methanesulfonic acid (MSA) is formed in the atmosphere by dimethylsulfide (DMS) oxidation and is deposited onto the surfaces of terrestrial and aquatic environments, where it can be used by microorganisms as sulfur, carbon, and energy sources (19). Considering that 25%–70% of DMS flux is oxidized to MSA (19), utilization of MSA by microorganisms can be important in the biogeochemical sulfur cycle. Methanesulfonate monooxygenase mediates the oxidation of MSA to formaldehyde and sulfite in some methylotrophic bacteria and is composed of hydroxylase, ferredoxin, and reductase components (20, 21). ORF61 of HMO-2011 encodes a hydroxylase alpha-subunit of methanesulfonate monooxygenase (MsmA). To determine whether ORF61 can be considered an auxiliary metabolic gene (AMG), we checked whether IMCC1322 is equipped with gene sets for the import and utilization of MSA. Because it is known that IMCC1322 can use formaldehyde, the oxidized product of MSA (22), we searched for MSA uptake and oxidation genes. A BLASTP search using ORF61 as a query returned two hits from IMCC1322: SAR116_2102 and SAR116_2109. Examination of the regions surrounding these two genes showed many genes involved in MSA utilization. Five genes (SAR116_2097-2101) encoded the ABC-type nitrate/sulfonate/bicarbonate transport system for MSA uptake. Six genes (SAR116_2102-2107) encoded components of methanesulfonate monooxygenase: hydroxylase alpha- and beta-subunit; ferredoxin; and reductase subunits A, B, and C. Another hydroxylase alpha-subunit was encoded by SAR116_2109. These results show that IMCC1322 can take up and oxidize MSA, which suggests that the MsmA protein encoded by ORF61 could have an effect on host metabolism during infection. Therefore, ORF61 can be regarded as an AMG. Both the MsmA proteins encoded by IMCC1322 are unusual in their sequences. SAR116_2102 is much longer than most other MsmA proteins and has a pyridine nucleotide-disulfide oxidoreductase domain in its C terminus. SAR116_2109 lacked the four cysteine and histidine residues in the Rieske domain. Interestingly, a recent metatranscriptomic study performed for coastal waters in the southeastern region of the United States suggested that the relatively active transcription of a few genes involved in MSA utilization, including SAR116_2101 and SAR116_2109, may be a distinctive characteristic of IMCC1322 and close relatives in coastal regions (23). HIMB100, another cultured isolate of the SAR116 clade, does not have genes for MSA uptake and utilization.

Search for Sequences Similar to the HMO-2011 Genome in the nr, env_nr, GOS, and BroadPhage Databases. First, we compared the HMO-2011 genome to the nonredundant (nr) and environmental nonredundant (env_nr) databases by using BLASTP. Each ORF of HMO-2011 was used as a query against the nre database (merged database of nr and env_nr) provided in the Bioinformatics Toolkit (http://toolkit.tuebingen.mpg.de) to search for similar protein sequences (E-value ≤ 0.01). When normalized by the length of each ORF and database sizes of nr or env_nr, the search results showed that all HMO-2011 ORFs had more hits in env_nr than in nr, except 6 ORFs that had no hits in either database. Twenty-nine ORFs had hits in only env_nr and 67 ORFs had their best hit in env_nr (Fig. S4*A*). Nearly all hits in env_nr were from the global ocean sampling (GOS) database. Second, a comparison between the GOS database (CAM_PROJ_GOS) and the BroadPhage database (CAM_PROJ_BroadPhage) was performed using TBLASTN at the Community Cyberinfrastructure for Advanced Microbial Ecology Research and Analysis (CAMERA) portal (https://portal.camera.calit2.net) (18). Each ORF of the HMO-2011 genome was used as a query, and all BLAST parameters were default values except for gap open cost (=11) and gap extend cost (=1). All hits satisfying the criteria of alignment length (≥20 amino acids) and bitscore (≥40) were counted. Hit counts were normalized by the length of each ORF and database sizes. All 74 ORFs had hits in both databases, and 72 ORFs had more hits in the BroadPhage database (Fig. S4*B*).

SI Materials and Methods

Cultivation of IMCC1322 and Cell Counting. The host bacterial strain, IMCC1322, was originally isolated from a tiny colony grown on 1/10 marine R2A agar that contained 1.82 g of R2A agar powder (Difco) and 13.5 g of Bacto agar (Difco) per 1 L of diluted aged seawater [distilled water:aged seawater, 2:8 (vol/vol)]. However, growth of IMCC1322 on 1/10 marine R2A agar was very slow and often inefficient. When a colony grown on agar plates was transferred to a new plate by streaking, only 5–10 colonies were visible to the naked eye after 2–3 wk of incubation at 20 °C. Furthermore, it was difficult to establish a stable broth culture from these colonies. Therefore, a large number of glycerol stocks were prepared every time a stable broth culture was obtained, and these were used as inoculum for broth media. For the experiments performed in this study, IMCC1322 was grown and maintained in a seawater-based liquid medium (mPYC; 0.5 g proteose peptone, 0.5 g yeast extract, and 0.5 g casamino acids per 1 L seawater) at 15 °C with shaking at 100 rpm. mPYC medium was prepared by adding a presterilized nutrient stock solution at a ratio of 1:50 to aged seawater that had been filtered through a 0.2-μm filter and had been autoclaved. The nutrient stock solution contained proteose peptone no. 3, yeast extract, and casamino acids (25 $g \cdot L^{-1}$ each, all from Difco) and was sterilized by autoclaving (121 °C, 20 min). Broth cultures were usually grown in 30–50 mL of mPYC medium using 125-mL baffled Erlenmeyer polycarbonate flasks equipped with vent caps (Corning) and were transferred to new medium in new flasks at a ratio of 1:10–1:20 every 5 d. The concentration of IMCC1322 cells was determined using a Guava EasyCyte flow cytometer (Millipore) after staining with SYBR Green I (Invitrogen) (24).

Morphological Characterization Using Transmission Electron Microscopy. Phage particles in culture lysate (100 mL) were pelleted by ultracentrifugation (120,000 × g, 1 h) after filter sterilization (0.2 μm). After the supernatants were discarded, the pellets were resuspended in 50 μL of SM buffer without gelatin (50 mM Tris-Cl, 0.1 M NaCl, 8 mM $MgSO_4$, pH 7.5). Concentrated phage (5 μL) was adsorbed onto carbon and formvar-coated 200-mesh copper grids (EMS) and stained with a 2% solution of uranyl acetate (Sigma). Grid examination was performed using a transmission electron microscope (CM200; Philips).

Plaque Assay. Plaque assays for the purification and titration of phage samples were performed using the double agar overlay method. The recipe for the bottom and top agars was the same as that for mPYC except that Noble agar (Difco) was added to filtered seawater before autoclaving at a final concentration of 1.5% (bottom) and 0.5% (top). For the plaque assay, phage samples (100 μL) were added to 500 μL of IMCC1322 culture, mixed briefly by vortexing, and incubated for 30–60 min at 20 °C. After incubation, 5 mL of top agar, maintained at 42 °C, was added to the mixed samples. After mixing by swirling, the mixtures were immediately poured onto the bottom agar plates and allowed to solidify. Plaques were counted after incubation at 20 °C for 1 wk. Phage samples were diluted with autoclaved seawater as necessary.

One-Step Growth Curve. Exponentially growing IMCC1322 cells (~2×10^8 cells in 0.4 mL) were mixed with HMO-2011 stock (~2×10^7 pfus in 0.1 mL) and incubated for 1 h at 20 °C. After incubation, the mixture was centrifuged (12,000 × *g*, 5 min) and the supernatant was discarded. The pellet was resuspended by vortexing after adding 0.5 mL of mPYC. Subsequently, 0.1 mL of this suspension was added to three flasks containing 30 mL of mPYC. The flasks were incubated at 20 °C (at 100 rpm) for 24 h. Culture broths were withdrawn every 2 h and used for the plaque assay as described above. Plaques were counted after incubation at 20 °C for 1 wk.

Coculture of IMCC1322 and HMO-2011. Fifteen milliliters of exponentially growing IMCC1322 culture was inoculated into 300 mL of mPYC in a 1-L Erlenmeyer flask and mixed by swirling. This mixture was dispensed into six 125-mL Erlenmeyer flasks (50 mL per flask). Phage stock solution was added to three flasks at a multiplicity of infection (MOI) of ~0.001 and mixed by swirling. The remaining three flasks were used as controls. Cultures were incubated at 15 °C with shaking at 100 rpm for 6 d. During incubation, 2 mL of culture broth was withdrawn from each flask every day and was used for the following experiments. One milliliter from all six flasks was fixed with 0.2-μm filtered formalin (final concentration, 2%) and stored at 4 °C until it was used for the determination of IMCC1322 cell numbers by flow cytometry as described above. Chloroform was added (10%, vol/vol) to 1 mL of culture broth withdrawn from the three flasks inoculated with phage stock solution and mixed by vortexing for 1 min. After centrifugation (5 min, 12,000 × *g*), the aqueous phase was recovered and stored at 4 °C until it was used for the determination of phage titers by plaque assays.

Phylogenetic Analysis of DNA Polymerase Sequences. Three Pfam alignment files [Full, National Center for Biotechnology Information (NCBI), and Metagenomics] of PF00476 (DNA polymerase family A) were downloaded and used to extract the sequence regions corresponding to the DNA polymerase family A domain from 56 DNA polymerases. These extracted regions were used as a search database for BLASTP analysis of HMO2011 ORF12 and 2 family A DNA polymerase sequences of pelagiphages, HTVC011P (AGE60547.1) and HTVC019P (AGE60596.1). Based on the BLASTP results, sequence regions corresponding to the DNA polymerase family A domain were extracted from the DNA polymerases of HMO-2011 (331–680 aa), HTVC011P (201–589 aa), and HTVC019P (201–565 aa). These three sequences were combined with the above 56 sequences and aligned using MUSCLE for tree building (25). An unrooted maximum-likelihood tree was constructed using RAxML version 7.2.8 with a bootstrap of 100 replicates (26). The command line was as follows: raxmlHPC-PTHREADS -f a -s DNApol.phy -n DNApol -m PROTGAMMAWAGF -x 0123 -# 100 -T 4.

Viromes Used for Binning Analysis: Selection Process, Brief Descriptions, and Sequence Processing. Seven viromes from the Indian and Pacific Ocean were selected for binning analyses. Four viromes from the Indian Ocean were included in our analyses without preliminary selection procedures because these were the only viral metagenomes obtained from the Indian Ocean (27). The three Pacific Ocean viromes were selected among the marine viromes deposited in the CAMERA Web site, mainly based on the proportion of reads recruited by the HMO-2011 genome (18). The CAMERA database was chosen as a starting point for the selection of viromes because it contains the BroadPhage database in addition to legacy viral metagenome data (28, 29). First, BLASTN analyses were performed using the HMO-2011 genome as a query. The search databases were "All Metagenomic 454 Reads" (All454) and "All Metagenomic Sanger Reads" (AllSanger). All BLAST parameters were the default values except for match reward (=2), mismatch penalty (=−3), gap open cost (=5), gap extend cost (=2), and low complexity filter (=F). Only the hits satisfying the criteria of alignment length (≥50) and bitscore (≥40) were counted. From the search results, the proportion of sequences similar to the HMO-2011 genome was calculated for each metagenome. Because CAMERA restricted the number of hits per query to 50,000, the results from All454 were insufficient for retrieving all sequences satisfying the criteria described above. Therefore, the top 20 metagenomes were selected based on the proportions calculated from 50,000 hits, excluding datasets from microbial fractions (>0.1 μm), animals, PCR products, mesocosm experiments, or ssDNA. These 20 metagenomes were reanalyzed separately using the same method as described above. When combined with the result from AllSanger, 11 viromes had a higher proportion (>1% of total reads) of reads similar to the HMO-2011 genome. These 11 viromes were as follows: CAM_SMPL_MOVE0902, JCVI_SMPL_1103283000058 (MOVE858), CAM_SMPL_000816, CAM_SMPL_000801, CAM_SMPL_001011, CAM_SMPL_000990, CAM_SMPL_000722, CAM_SMPL_000723, CAM_SMPL_000725, CAM_SMPL_000724, and CAM_SMPL_000727. Two viromes, CAM_SMPL_MOVE0902 and JCVI_SMPL_1103283000058 (MOVE858), were excluded from further analyses because they contained a relatively small number of sequences (5,641 and 11,496 sequences, respectively). In addition, 5 viromes from Scripps Pier (CAM_SMPL_000722, CAM_SMPL_000723, CAM_SMPL_000725, CAM_SMPL_000724, and CAM_SMPL_000727) were replaced by a single virome (CAM_S_1336) that was indicated to originate from a seawater sample collected from the same station on the same day as the above 5 viromes (30), based on the suggestion made by Dr. Matthew Sullivan (University of Arizona, principal investigator for all 6 viromes from Scripps Pier). Consequently, 5 viromes from the CAMERA database were used for binning analysis: CAM_SMPL_000816, CAM_SMPL_000801, CAM_SMPL_001011, CAM_SMPL_000990, and CAM_S_1336. However, the results of 2 viromes from station ALOHA (CAM_SMPL_000816 and CAM_SMPL_000801) are not shown in Table 1 and Fig. 4 because it was known to us, after the binning analyses were finished, that these 2 viromes were obtained using DNA samples extracted from bands of a pulsed-field electrophoresis gel on which a viral community DNA sample was resolved. CAM_SMPL_000816 was from a band around 60 kb whereas CAM_SMPL_000801 was from a band around 130 kb (Dr. Grieg Steward, University of Hawaii, principal investigagor of 2 viromes). For the readers' information, results of BLASTN analyses of 2 viromes are briefly presented below. The HMO-2011 genome was assigned by 3.3% and 2.2% of total reads, contributing 38.6% and 7.9% of the reads assigned to viruses for CAM_SMPL_000816 and CAM_SMPL_000801, respectively. Reads assigned to HMO-2011 from CAM_SMPL_000801 were highly biased toward ORF18 that encodes a ribonucleotide reductase.

Brief descriptions of the seven selected viromes are given below to present information that may be important for interpretation of the binning results. All seven viromes were generated by pyrosequencing. For the four viromes from the Indian Ocean, water samples were filtered through 0.1-μm membrane filters and concentrated by tangential flow filtration (TFF). These viral concentrates were treated with nuclease, pelleted by a sucrose cushion, and used for DNA extraction. DNA samples were fragmented and used to construct linker-amplified shotgun libraries for pyrosequencing (27). A surface seawater sample used for a virome from Scripps Pier (CAM_S_1336) was filtered through 0.22-μm membrane filters and concentrated using either TFF or $FeCl_3$ precipitation. Viral concentrates were purified with DNase treatment, and some subsamples were further purified by CsCl-step gradient or sucrose cushion ultracentrifugation. DNA samples were extracted from the purified viral concentrates and used for pyrosequencing after library construction by linker amplification (30). To our knowledge, there are no publicly available references for the remaining two viromes except the metadata provided by the CAMERA database. A virome from the northeastern subarctic Pacific (CAM_SMPL_001011) was obtained from a water sample filtered through 0.2-μm membrane filters. Template preparation methods included precipitation by $FeCl_3$, DNase treatment, and CsCl-gradient ultracentrifugation. This virome shared the latitude and longitude of the sampling station, the date of sampling, sampling depth, and principal investigator with CAM_SMPL_002238 (31). Therefore, it is highly probable that CAM_SMPL_001011 was produced using a linker-amplified shotgun library protocol similar to CAM_SMPL_002238. A virome from Southern California Bight (CAM_SMPL_000990) seems to have been obtained from a mixed sample that was prepared using at least six water samples collected from the upper mixed euphotic layer on three different sampling dates. Although only CsCl-gradient ultracentrifugation is mentioned as a template preparation method in the CAMERA database, this virome is believed to have been obtained from water samples processed as follows: filtration with a 0.22-μm filter, $FeCl_3$ flocculation, DNase treatment, CsCl-gradient ultracentrifugation, and linker amplification (communication with Dr. Jed Fuhrman, University of Southern California, principal investigator). See Table S3 for more details about the viromes selected, including the latitude and longitude of sampling stations, temperature and salinity of water samples, and sampling dates.

Sequencing reads in the seven selected viromes were quality trimmed before use for binning analysis. We downloaded the sra files from NCBI for the four Indian Ocean viromes (SRX096024, SRX096023, SRX096025, and SRX096299) and extracted fasta and qual files. Fasta files were downloaded from the CAMERA database for the remaining three viromes. Subsequent sequence processing was performed using Mothur (32). In brief, sequences were trimmed according to the following criteria: minimum length = 100, maximum number of $N = 1$, and maximum length of homopolymers = 12. For the Indian Ocean viromes, quality trimming using qual files was also performed as follows: qwindowsize = 50 and qwindowaverage = 25. Putative key, linker, or primer sequences were also trimmed. Finally, artificial replicates were removed using cd-hit-454 (version 4.6.1) with default parameters except for a 0.99 similarity cutoff. The resulting fasta files were used for binning analyses.

1. Chun J, et al. (2007) EzTaxon: A web-based tool for the identification of prokaryotes based on 16S ribosomal RNA gene sequences. *Int J Syst Evol Microbiol* 57(Pt 10): 2259–2261.
2. Rohwer F, et al. (2000) The complete genomic sequence of the marine phage Roseophage SIO1 shares homology with nonmarine phages. *Limnol Oceanogr* 45:408–418.
3. Sullivan MB, Coleman ML, Weigele P, Rohwer F, Chisholm SW (2005) Three *Prochlorococcus* cyanophage genomes: Signature features and ecological interpretations. *PLoS Biol* 3(5): e144.
4. Pope WH, et al. (2007) Genome sequence, structural proteins, and capsid organization of the cyanophage Syn5: A "horned" bacteriophage of marine *synechococcus*. *J Mol Biol* 368(4):966–981.
5. Tuteja N, Tuteja R (2004) Unraveling DNA helicases: Motif, structure, mechanism and function. *Eur J Biochem* 271(10):1849–1863.
6. Frick DN, Richardson CC (2001) DNA primases. *Annu Rev Biochem* 70:39–80.
7. Rusch DB, et al. (2007) The *Sorcerer II* Global Ocean Sampling expedition: Northwest Atlantic through eastern tropical Pacific. *PLoS Biol* 5(3):e77.
8. Nunes-Düby SE, Kwon HJ, Tirumalai RS, Ellenberger T, Landy A (1998) Similarities and differences among 105 members of the Int family of site-specific recombinases. *Nucleic Acids Res* 26(2):391–406.
9. Groth AC, Calos MP (2004) Phage integrases: Biology and applications. *J Mol Biol* 335 (3):667–678.
10. Canchaya C, Fournous G, Brüssow H (2004) The impact of prophages on bacterial chromosomes. *Mol Microbiol* 53(1):9–18.
11. Albertí S, et al. (1996) Interaction between complement subcomponent C1q and the *Klebsiella pneumoniae* porin OmpK36. *Infect Immun* 64(11):4719–4725.
12. Roumenina LT, et al. (2008) Interaction of the globular domain of human C1q with *Salmonella typhimurium* lipopolysaccharide. *Biochim Biophys Acta* 1784(9): 1271–1276.
13. Sandmeier H (1994) Acquisition and rearrangement of sequence motifs in the evolution of bacteriophage tail fibres. *Mol Microbiol* 12(3):343–350.
14. Tétart F, Repoila F, Monod C, Krisch HM (1996) Bacteriophage T4 host range is expanded by duplications of a small domain of the tail fiber adhesin. *J Mol Biol* 258 (5):726–731.
15. Lee C-N, et al. (2007) Comparison of genomes of three *Xanthomonas oryzae* bacteriophages. *BMC Genomics* 8:442.
16. Geer LY, Domrachev M, Lipman DJ, Bryant SH (2002) CDART: Protein homology by domain architecture. *Genome Res* 12(10):1619–1623.
17. Hollich V, Sonnhammer ELL (2007) PfamAlyzer: Domain-centric homology search. *Bioinformatics* 23(24):3382–3383.
18. Sun S, et al. (2011) Community cyberinfrastructure for Advanced Microbial Ecology Research and Analysis: the CAMERA resource. *Nucleic Acids Res* 39(Database issue): D546–D551.
19. Kelly DP, Murrell JC (1999) Microbial metabolism of methanesulfonic acid. *Arch Microbiol* 172(6):341–348.
20. Baxter NJ, Scanlan J, De Marco P, Wood AP, Murrell JC (2002) Duplicate copies of genes encoding methanesulfonate monooxygenase in *Marinosulfonomonas methylotropha* strain TR3 and detection of methanesulfonate utilizers in the environment. *Appl Environ Microbiol* 68(1):289–296.
21. Jamshad M, De Marco P, Pacheco CC, Hanczar T, Murrell JC (2006) Identification, mutagenesis, and transcriptional analysis of the methanesulfonate transport operon of *Methylosulfonomonas methylovora*. *Appl Environ Microbiol* 72(1):276–283.
22. Oh H-M, et al. (2010) Complete genome sequence of "*Candidatus* Puniceispirillum marinum" IMCC1322, a representative of the SAR116 clade in the *Alphaproteobacteria*. *J Bacteriol* 192(12):3240–3241.
23. Gifford SM, Sharma S, Booth M, Moran MA (2013) Expression patterns reveal niche diversification in a marine microbial assemblage. *ISME J* 7(2):281–298.
24. Stingl U, Tripp HJ, Giovannoni SJ (2007) Improvements of high-throughput culturing yielded novel SAR11 strains and other abundant marine bacteria from the Oregon coast and the Bermuda Atlantic Time Series study site. *ISME J* 1(4):361–371.
25. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32(5):1792–1797.
26. Stamatakis A (2006) RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22(21):2688–2690.
27. Williamson SJ, et al. (2012) Metagenomic exploration of viruses throughout the Indian Ocean. *PLoS ONE* 7(10):e42047.
28. Angly FE, et al. (2006) The marine viromes of four oceanic regions. *PLoS Biol* 4(11): e368.
29. Dinsdale EA, et al. (2008) Functional metagenomic profiling of nine biomes. *Nature* 452(7187):629–632.
30. Hurwitz BL, Deng L, Poulos BT, Sullivan MB (2013) Evaluation of methods to concentrate and purify ocean virus communities through comparative, replicated metagenomics. *Environ Microbiol* 15(5):1428–1440.
31. Hurwitz BL, Sullivan MB (2013) The Pacific Ocean virome (POV): A marine viral metagenomic dataset and associated protein clusters for quantitative viral ecology. *PLoS ONE* 8(2):e57355.
32. Schloss PD, et al. (2009) Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* 75(23):7537–7541.

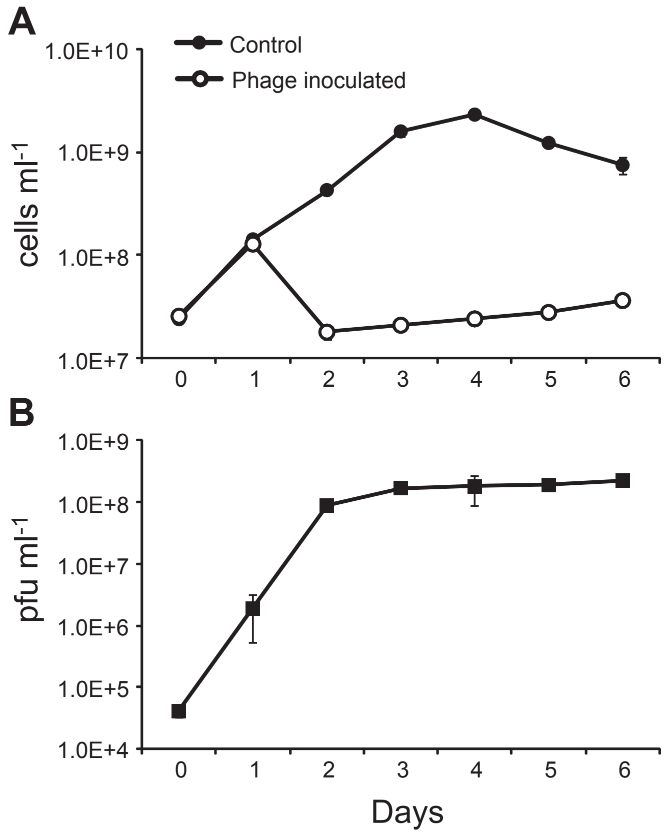

Fig. S1. Lysis of IMCC1322 by HMO-2011 during coculture. HMO-2011 was inoculated into exponentially growing IMCC1322 culture at day 0. (*A*) Concentration of IMCC1322 cells. Noninoculated samples were used as controls. (*B*) Titers of HMO-2011. All data are averages from triplicate experiments, and error bars indicate SDs.

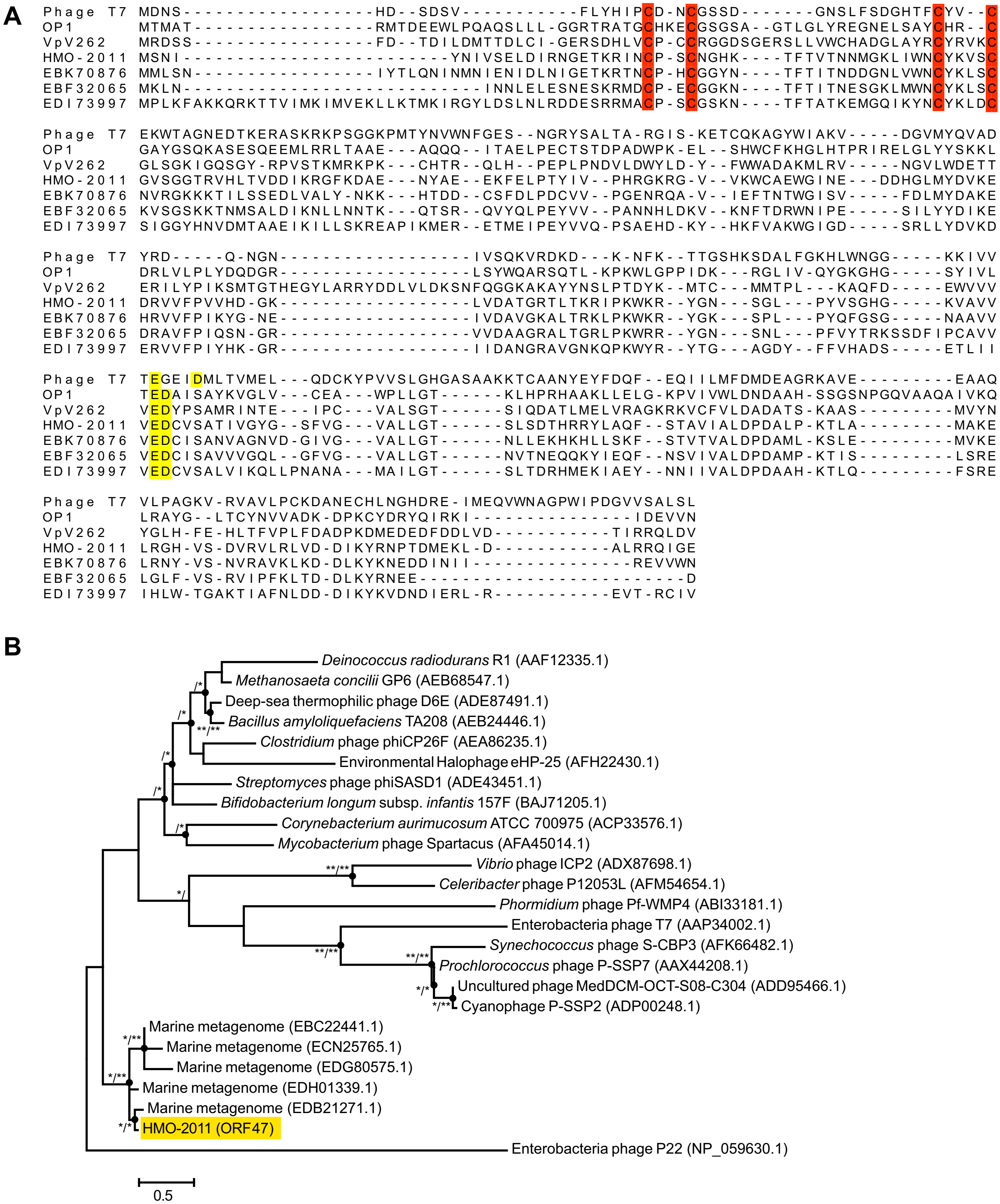

Fig. S2. (Continued)

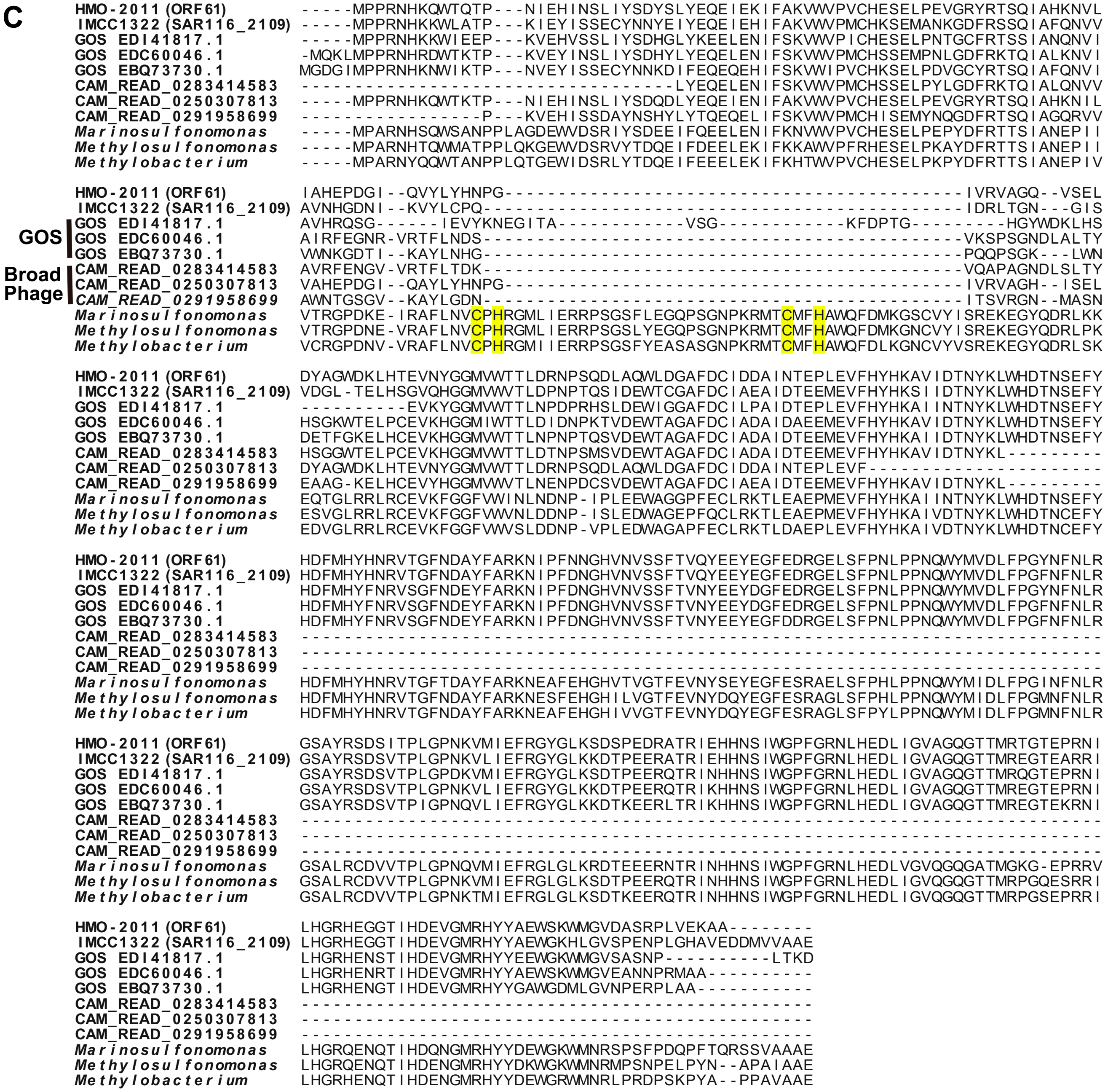

Fig. S2. Several representative ORFs that were more closely related to sequences from marine metagenomes than to those from cultured organisms. (*A*) Alignment of primase sequences. Residues in orange and yellow correspond to putative zinc-binding motifs and Toprim domain catalytic sites (PF08275), respectively. Note that 2 amino acids are inserted in the second CXXC motif of the zinc-binding domain and 3 amino acids are deleted in the catalytic motif of the Toprim domain in HMO-2011, VpV262, and the metagenome sequences, compared with phage T7. A deletion in the Toprim domain was also observed in OP1. Accession numbers: Phage T7, NP_041975.1 (1–271 amino acids); OP1 (*Xanthomonas* phage OP1), YP_453600.1; VpV262 (*Vibrio* phage VpV262), AAM28363.1. EBK70876, EBF32065, and EDI73997 were retrieved from the GOS metagenome. Alignment was obtained using T-Coffee (1). Alignment of phage T7 was manually adjusted in a short region including zinc-binding motifs, without considering overall alignment quality to more clearly show the position of cysteine or histidine residues. (*B*) Maximum-likelihood (ML) tree showing relationships among capsid proteins. Alignment using MUSCLE and tree building with ML and neighbor joining (NJ) was performed in MEGA 5.05 (2). The gap option was set to partial deletion (90% cutoff), and the robustness of branches was checked through bootstrap analyses (100 replicates) in both algorithms. Nodes recovered in both trees are indicated by black circles. Bootstrap values are indicated at the nodes with a single (≥50) or double (≥90) asterisk (ML/NJ). Note that ORF47 of HMO-2011 formed a robust cluster with marine metagenome sequences. In BLASTP analysis against the nr database of GenBank, ORF47 was shown to be most similar to a hypothetical protein of *Bifidobacterium longum* subsp. *infantis* 157F. Among known viruses, the capsid protein of *Geobacillus* phage D6E (Deep-sea thermophilic phage D6E) was most similar to ORF47. (*C*) Alignment of putative MsmA proteins. Residues in yellow correspond to cysteine and histidine amino acids conserved in the Rieske domain of MsmA proteins. Sequences marked GOS were retrieved from the env_nr database of GenBank whereas sequences marked BroadPhage were retrieved from the BroadPhageMetagenomes database at CAMERA. GenBank accession numbers for other sequences: IMCC1322 (SAR116_2109), ADE40352.1; *Marinosulfonomonas* (*Marinosulfonomonas methylotropha*), AAK84301.1; *Methylosulfonomonas* (*Methylosulfonomonas methylovora*), AAD26619.1; *Methylobacterium* (*Methylobacterium nodulans* ORS 2060), ACL62498.1. Alignment was generated using ClustalW2.

1. Notredame C, Higgins DG, Heringa J (2000) T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol* 302(1):205–217.
2. Tamura K, et al. (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol* 28(10): 2731–2739.

A

```
ORF29 -------MASELRVNTLKDAAGANSVTMTYVAGG-SAKAWVNFNGTGTPAVNESLNISSLSDDGVGKYDVNVSSAFSNAN
ORF32 ---LEYRMTSVLNVDSIAD-K-AGTGPVALTKQH-AAKGWCHANQVSVG-ITDSFNVSSISDQSAGLIYYTPTNSMSDAN
ORF34 -----MALTQVINKGIGAGNTVTGEGSATTSLQQGLAKAWVYAEADAVR--TDSFNMSSSTDHGDGDFSYGLTAAFSAVT
ORF37 LAGETIMALGKIKADTLEHST-AGSLNTSYVVQG-SAKHWCMVANGGVPSISNSLNASSVSDDSTGNFTITLASAMENTN
ORF40 --------MSTLKADTIQS---TGGGAATLTKQS-AAKVWGNLDGTGTIAIRDSFNVSSADDNGTGIYDFNYSNSMSNDD

ORF29 HSPVGSMIGDGSTSFRSYALVASLSANTASKIDVSTVNTGDGIQANTDYTSVNVHGLGDLA
ORF32 FAAC-VSVRNSSTGATFAQSLNDLHTSSRWASSS-KNSSSTGVDSQHNNIAVH----GDLA
ORF34 YSQS-GMVRTG---TLSKILTRNSDRSTSSLIAVEVGNTSDGPQNYDHDVQAH----GDLA
ORF37 QAAL-GMIANSAANPSTSNVTGTMELTTTLAVET-VNTSGTNLDVPFNQFAIN----GDLA
ORF40 YAPN-ESVSETSGFEGNLNRYPAIDGVSTSDIRFGATFSGSNYDVDELYVTVN----GDLA
```

B

```
ORF33 MPYIGKSPQNGVRNRFIYQATAGQTSFSGSDSDSKVLTYQDGLYMDVFQNGVLLKPGTDYTATTGTTVVLVTGASSGDVV
ORF41 MPYIGQKVPGSYQATKAVQRFNGDNSTTTFTLTTAVSSVQD---VLVSVDGVVQDT-TAYTIPDGTTLTFTAAPSTGTAN

ORF33 EMVAYDVFSVANSYTVTESDTRYPFKGNNSIIRLNGQSITSDVTIDADENGVSAGPITQDN-ATVTVNGYWSIV--
ORF41 IFVNYLAPQAG--TITPPAENKGNFKGG-GLFRTNAQSLTSDITILATENANVTGPFTVASGVTLTVESGGTLVTL
```

C

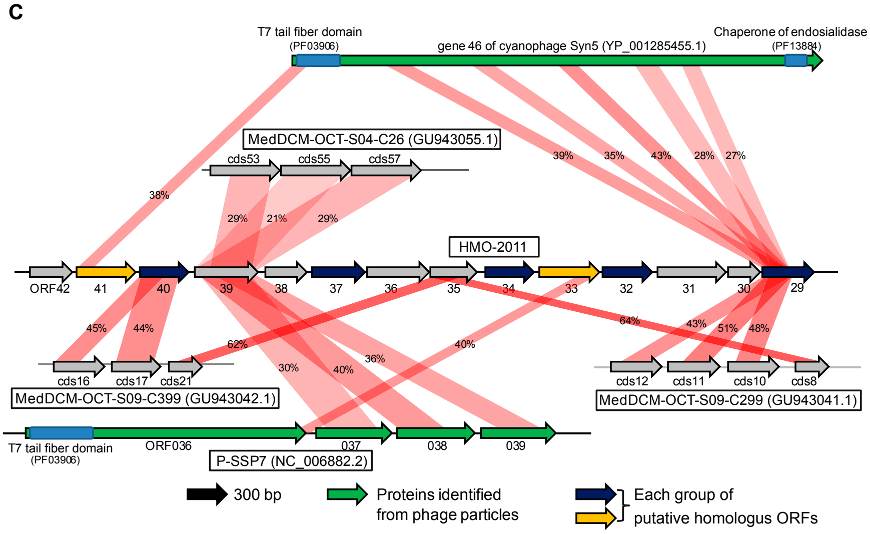

Fig. S3. Relationships based on homology among the tail-related proteins of HMO-2011, other marine phages, and marine environmental fosmids. (*A*) Alignment of the five tail proteins of HMO-2011 sharing C-terminal ends. Conserved residues are in yellow. Alignments were generated using ClustalW2. (*B*) Alignment of the two tail proteins of HMO-2011 sharing N-terminal ends. Conserved residues are in yellow. Alignments were generated using ClustalW2. (*C*) Similarities among tail-related proteins of HMO-2011, other phages, and environmental fosmids. ORFs with homology are connected by red shading. Numbers within the shading indicate percent identity at the amino acid level, with the color intensity proportional to identity. The horizontal length of the shading is approximately proportional to alignment length. GenBank or RefSeq accession numbers are provided in parentheses for each protein, genome, or fosmid. ORF lengths are drawn to scale whereas intergenic regions are not. Some homology relationships were omitted for visualization convenience.

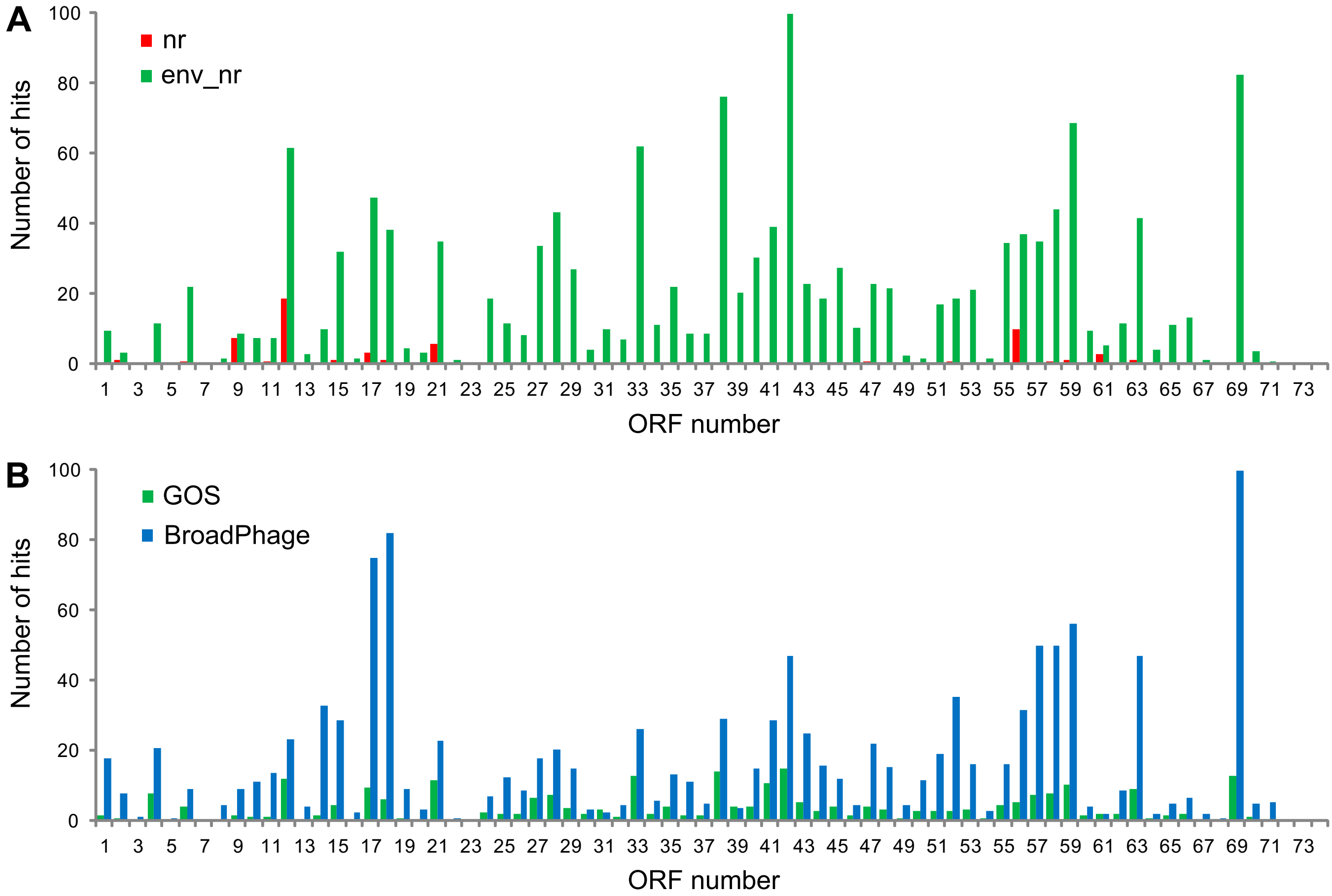

Fig. S4. BLAST analysis of HMO-2011 ORFs showing the number of similar sequences in several databases. (*A*) Normalized number of hits (*E*-value ≤ 0.01) obtained by BLASTP analysis against nre (merged nr and env_nr databases). Note that many ORFs only had hits from the env_nr database. Nearly all hits from env_nr were retrieved from the GOS database. (*B*) Normalized number of hits (bitscore ≥ 40, alignment length ≥ 20 amino acids) obtained through TBLASTN analysis against two marine metagenome databases provided by CAMERA. Nearly all ORFs had more hits from the BroadPhage database than from the GOS database. For both graphs, normalization was performed by dividing the actual number of hits with the lengths of each ORF and the sizes of each database. An arbitrary factor was multiplied to the number of normalized hits for visualization convenience. Different arbitrary factors were used for *A* and *B*.

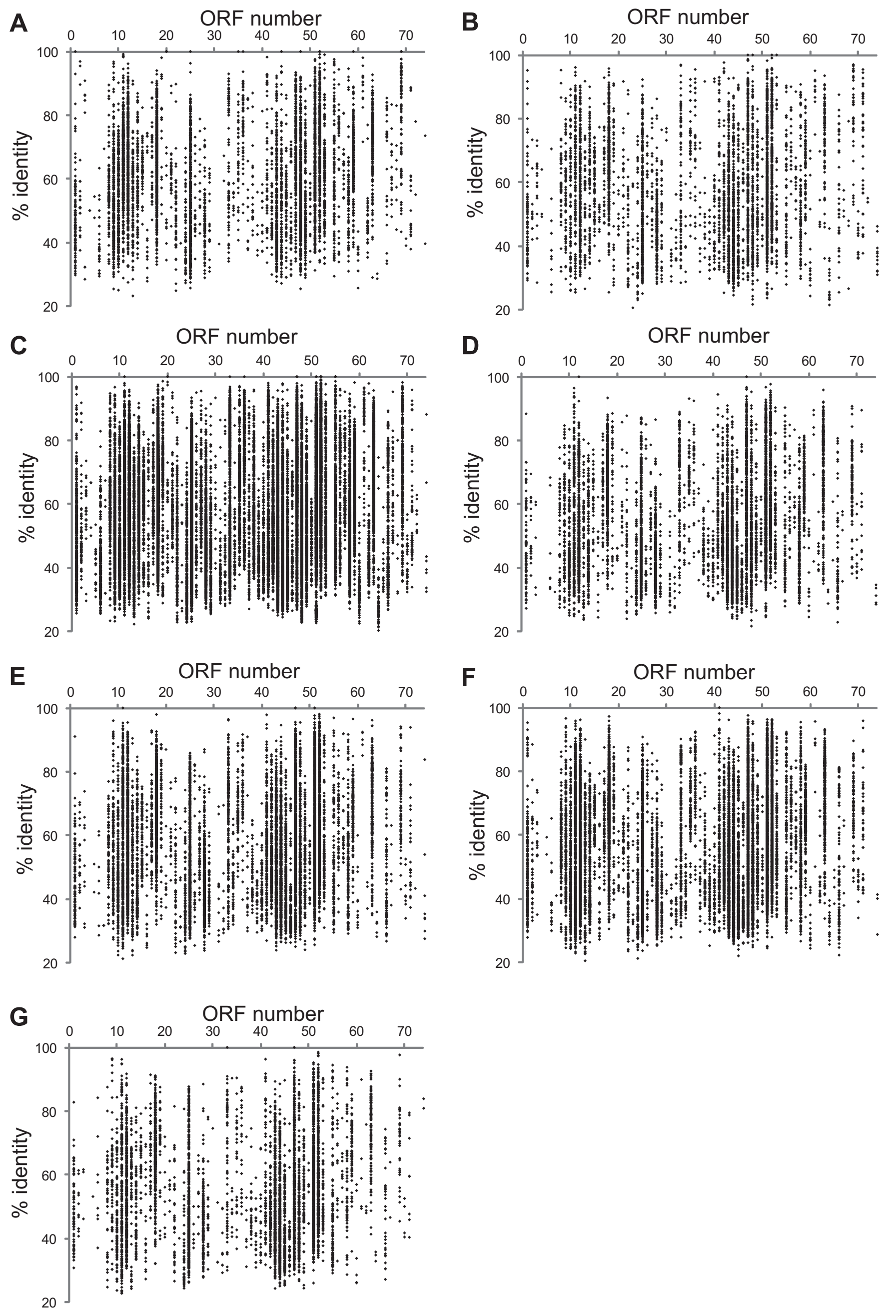

Fig. S5. Fragment recruitment plot of marine virome reads assigned to the HMO-2011 genome by BLASTX. Each virome read was plotted according to its matching ORF and sequence identity (%) at the amino acid level. (*A*) CAM_SMPL_001011, (*B*) CAM_SMPL_000990, (*C*) CAM_S_1336, (*D*) GSIOVIR108, (*E*) GSIOVIR112, (*F*) GSIOVIR117, (*G*) GSIOVIR122. See Table 1 and Table S3 for more information about the samples.

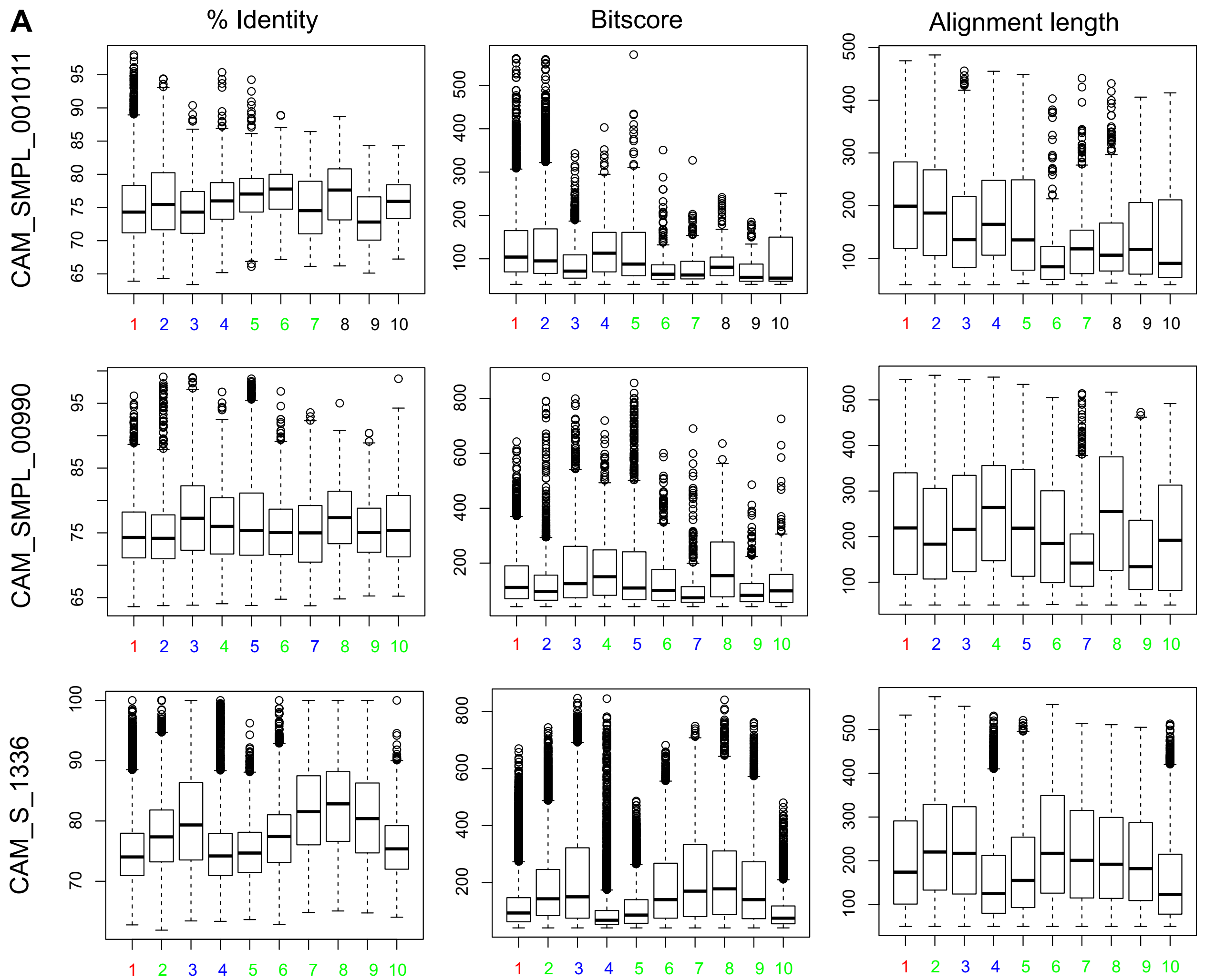

Fig. S6. (Continued)

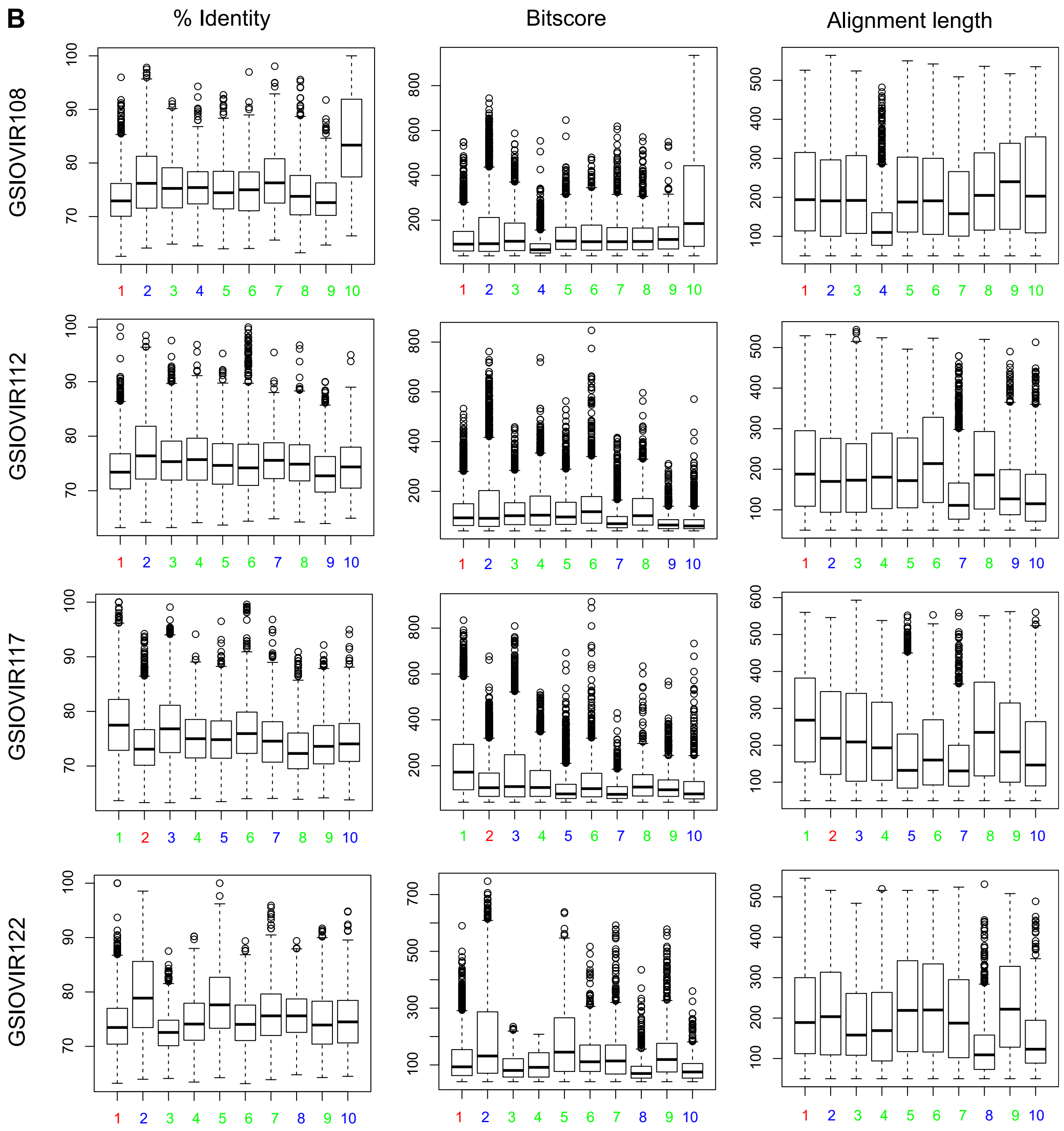

Fig. S6. (Continued)

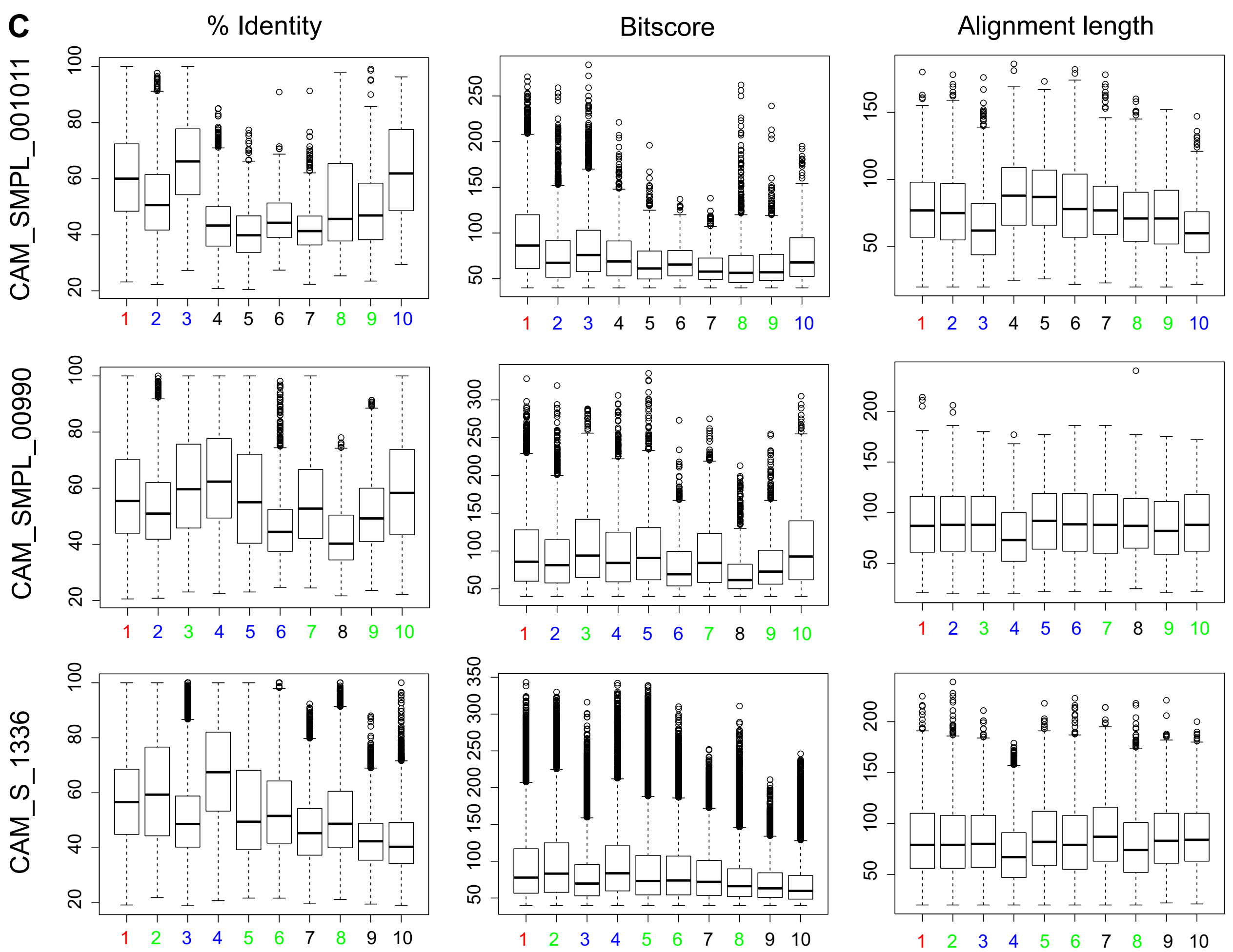

Fig. S6. (Continued)

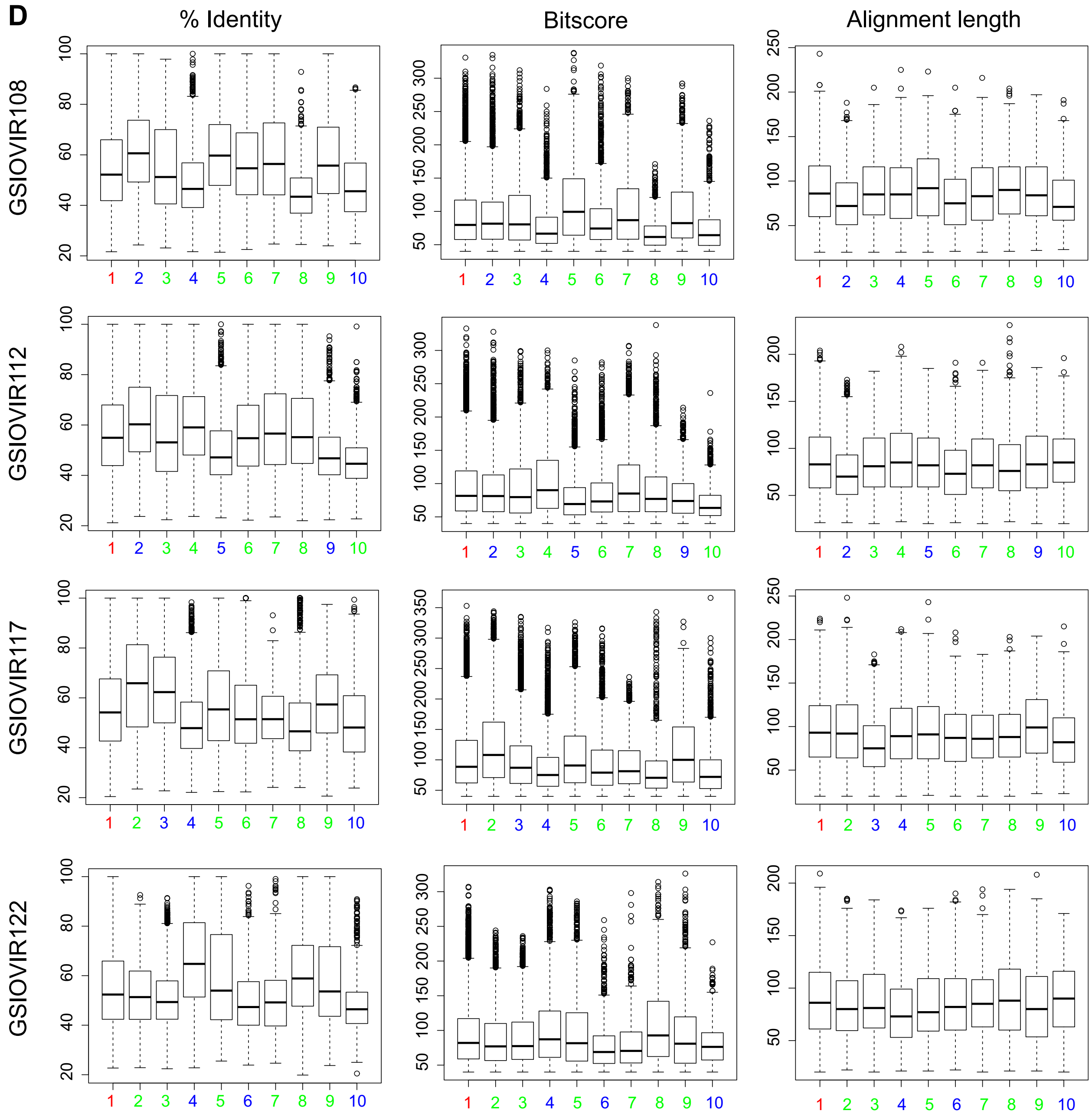

Fig. S6. Box plots showing the distribution of sequence similarities, bitscores, and alignment lengths of virome reads assigned to each highly ranked virus genome. Numbers on the *x* axes correspond to the rankings of each highly assigned virus as presented in Table S3 (refer to Table S3 for the names of the highly assigned viruses). Colors are assigned as in Table S3: HMO-2011 in red, pelagiphages in blue, cyanophages in green, and all other viruses in black. Data outside the 1.5× interquartile range, from the first or third quartile, are indicated with open circles. (*A*) Reads from three Pacific Ocean viromes assigned by BLASTN, (*B*) reads from four Indian Ocean viromes assigned by BLASTN, (*C*) reads from three Pacific Ocean viromes assigned by BLASTX, and (*D*) reads from four Indian Ocean viromes assigned by BLASTX. Sequence similarities were calculated at the nucleotide level (*A* and *B*) or at the amino acid level (*C* and *D*). Alignment lengths have been presented in nucleotides (*A* and *B*) or in amino acids (*C* and *D*).

Table S1. Annotation of ORFs predicted in the HMO-2011 genome

ORF	Position	Strand	Size, aa	Most significant hit in BLASTP against nr* (organism, GenBank accession no., no. of identical residues/aligned length/hit length, *E*-value)	Predicted function	Domain, family, signal peptide (SP), and transmembrane helix (TM)
1	433–1182	+	249	gp86 (*Mycobacterium* phage Twister, AFF28335, 73/201/290, 7E−20)	—	—
2	1550–1951	+	133	gp89 (*Mycobacterium* phage Alma, AER48797, 46/114/118, 3E−15)	—	—
3	1974–2381	+	135	Apolipoprotein *N*-acyltransferase (*Desulfobacter postgatei*, EIM62110, 18/75/519, 5.1)	—	—
4	2779–2961	+	60	Hypothetical protein (*Synechococcus* phage S-SSM7, ADO98252, 24/56/48, 3E−5)	—	—
5	3012–3353	+	113	COP1-interacting protein-like protein (*Arabidopsis thaliana*, AEE81904, 22/48/317, 0.53)	—	—
6	3495–4007	+	170	Hypothetical protein (*Neisseria bacilliformis*, EGF12106, 29/74/220, 2E−10)	—	—
7	4070–4474	+	134	Mediator of DNA damage checkpoint protein 1-like (*Apis florea*, XP_003690909, 26/99/279, 0.50)	—	SP; TM
8	4544–5380	+	278	Hypothetical protein (*Vibrio fischeri*, ACH64802, 51/240/318, 3E−3)	—	PF06067 (Domain of unknown function DUF932)
9	5377–6255	+	292	Putative phage integrase family protein (*Azospirillum brasilense*, CCC96864, 84/291/353, 3E−19)	Integrase	PF00589 (Phage integrase family)
10	6280–7059	+	259	Putative DNA primase (*Xanthomonas* phage OP1, BAE72747, 57/180/282, 2E−07)	Primase	SSF56731 (DNA primase core)
11	7064–8287	+	407	Putative replicative DNA helicase (Blood disease bacterium R229, CCA83263, 112/433/425, 2E−20)	Helicase	PF13481 (AAA domain); SSF52540 (P-loop containing nucleoside triphosphate hydrolases)
12	8284–10326	+	680	DNA polymerase I (*Desulfotomaculum nigrificans*, EGB21815, 117/357/882, 4E−35)	DNA polymerase	PF01612 (3′-5′ exonuclease); PF00476 (DNA polymerase family A)
13	10366–11268	+	300	Hypothetical protein (*Mycobacterium tuberculosis*, ZP_02549341, 28/79/251, 0.82)	—	—
14	11268–12122	+	284	DNA polymerase (Uncultured organism, AAL02212, 65/178/182, 6E−21)	—	—
15	12112–12573	+	153	Endonuclease (*Celeribacter* phage P12053L, AFM54632, 78/123/133, 3E−46)	Endonuclease	PF05367 (Phage endonuclease I)
16	12570–12938	+	122	Hypothetical protein (*Bacillus pseudofirmus*, ADC49254, 20/73/127, 1.5)	—	—
17	12935–13219	+	94	Protein 1.7 (*Yersinia pestis* phage phiA1122, AAP20506, 40/72/76, 7E−19)	—	PF11753 (Protein of unknown function DUF3310)
18	13373–15316	+	647	Hypothetical protein (*Volvox carteri*, EFJ43138, 382/646/762, 0.0) Ribonucleoside-triphosphate reductase (*Acanthocystis turfacea* Chlorella virus OR0704.3, AGE59602, 374/636/627, 0.0)	Ribonucleotide reductase	SSF51998 (PFL-like glycyl radical enzymes) cd01676 (Class II ribonucleotide reductase, monomeric form)
19	15679–16182	+	167	Golgin subfamily A member 2 (*Acromyrmex echinatior*, EGI68077, 20/68/919, 1.6)	—	—
20	16172–16411	+	79	Bifunctional aspartate kinase/diaminopimelate decarboxylase protein (*Xylella fastidiosa*, AAO28288, 15/37/868, 6.3)	—	—

Table S1. Cont.

ORF	Position	Strand	Size, aa	Most significant hit in BLASTP against nr* (organism, GenBank accession no., no. of identical residues/aligned length/hit length, *E*-value)	Predicted function	Domain, family, signal peptide (SP), and transmembrane helix (TM)
21	16386–16769	+	127	MazG nucleotide pyrophosphohydrolase domain protein (*Alistipes* sp. HGB5, EFR57928, 67/110/108, 2E−37)	Nucleotide pyrophosphohydrolase	PF03819 (MazG nucleotide pyrophosphohydrolase domain)
22	20313–16927	—	1128	Branched-chain amino acid ABC transporter periplasmic protein (*Rhodopseudomonas palustris*, ACF02740, 44/150/411, 0.35)	—	—
23	21821–20313	—	502	No hits	—	—
24	22315–21821	—	164	Vacuolar protein sorting-associated protein 18 (*Culex quinquefasciatus*, EDS44865, 30/120/572, 2.8)	—	—
25	24624–22504	—	706	Hypothetical protein (Uncultured organism MedDCM-OCT-S04-C16, ADD96023, 36/71/354, 1E−15)	—	—
26	24832–24626	—	68	Phage replication protein (*Klebsiella oxytoca*, AFN33316, 16/52/710,0.72)	—	TM
27	25008–24832	—	58	Diguanylate cyclase (*Desulfuromonas acetoxidans*, EAT14554, 14/49/353, 7.9)	—	—
28	26104–25010	—	364	Hypothetical protein (*Acidovorax citrulli*, ABM32947, 101/294/669, 4E−33) Hypothetical protein Xp10p26 (*Xanthomonas* phage Xp10, AAP58693, 85/334/498, 7E−4)	Tail structure	—
29	26502–26101	—	133	Hypothetical protein (Uncultured phage MedDCM-OCT-S09-C299, ADD94746, 42/83/132, 4E−17) Tail fiber (Cyanophage Syn5, ABP87953, multiple matches: 20/46/1351, 20/72/1351, 17/44/1351, 16/46/1351, 20/75/1351)	Tail structure	—
30	26747–26502	—	81	Hypothetical protein (*Pelagibaca bermudensis*, EAU45063, 29/78/161, 8E−7) Tail fiber protein (*Synechococcus* phage S-CBS4, AEX56016, 19/52/146, 1E−3)	Tail structure	—
31	27276–26752	—	174	Hypothetical protein (Organic Lake phycodnavirus 2, ADX06235, 34/101/1038, 4E−9)	Structure	PF00386 (C1q domain)
32	27662–27276	—	128	Hypothetical protein (Uncultured phage MedDCM-OCT-S09-C399, ADD94772, 47/128/126, 4E−4) Tail fiber (Cyanophage Syn5, ABP87953, multiple matches from PSI-BLAST: 16/103/1351, 17/103/1351, 19/77/1351, 21/74/1351, 19/94/1351)	Tail structure	—
33	28108–27647	—	153	Hypothetical protein (Uncultured marine bacterium MedDCM-OCT-S09-C145, ADD94874, 39/86/545, 2E−14) Tail fiber-like protein (*Synechococcus* phage S-SM2, ADO97376, 43/111/919, 5E−8)	Tail structure	—
34	28488–28108	—	126	Hypothetical protein (Uncultured phage MedDCM-OCT-S09-C399, ADD94772, 31/90/126, 1.8E−2) Similar to ORF32 (amino acid identity: 31%)[†]	Tail structure	—
35	28865–28500	—	121	Predicted protein (Cyanophage NATL2A-133, ADP00151, 50/80/85, 4E−21) Tail fiber assembly protein (*Pantoea stewartii* DC283, EHU00846, 21/56/197, 1.7)	Tail structure	—
36	29352–28879	—	157	MFS general substrate transporter (*Fomitiporia mediterranea*, EJD02383, 25/67/791, 0.45)		—

Table S1. Cont.

ORF	Position	Strand	Size, aa	Most significant hit in BLASTP against nr* (organism, GenBank accession no., no. of identical residues/aligned length/hit length, *E*-value)	Predicted function	Domain, family, signal peptide (SP), and transmembrane helix (TM)
37	29753–29352	—	133	Hypothetical protein (Uncultured phage MedDCM-OCT-S09-C399, ADD94772, 25/56/126, 8E−06) Tail fiber (Cyanophage Syn5, ABP87953, two matches: 34/100/1351, 16/43/1351)	Tail structure	—
38	30055–29735	—	106	Hypothetical protein (Uncultured phage MedDCM-OCT-S04-C26, ADD95062, 38/93/117, 2E−12) Tail fiber protein (*Synechococcus* phage S-CBS4, AEX56016, 23/69/146, 1.3)	Tail structure	—
39	30545–30060	—	161	Hypothetical protein (*Prochlorococcus* phage P-SSP7, AAX44218, 59/163/191, 5E−9)	Structure	—
40	30916–30542	—	124	Hypothetical protein (Uncultured phage MedDCM-OCT-S09-C299, ADD94745, 57/132/132, 6E−19) Tail fiber (Cyanophage Syn5, ABP87953, Three matches 20/56/1351, 20/53/1351, 15/46/1351)	Tail structure	—
41	31362–30913	—	149	Hypothetical protein (*Synechococcus elongatus*, ABB56777, 24/47/387, 3E−3) Tail fiber (*Pelagibacter* phage HTVC019P, AGE60615, 24/57/491, 3.7)	Tail structure	—
42	31699–31373	—	108	Hypothetical protein (Uncultured organism MedDCM-OCT-S08-C1350, ADD96228, 48/81/85, 3E−26) Virion structural protein (Cyanophage S-TIM5, AEZ65682, 43/127/126, 6E−7)	Structure	—
43	33539–31758	—	593	Hypothetical protein (*Sinorhizobium fredii*, ACP24927, 71/124/506, 3E−27) S protein (Enterobacteria phage P1, AAQ14006, 69/355/987, 2E−17 from PSI-BLAST)	Tail structure	—
44	35051–33543	—	502	Hypothetical protein (*Acidovorax* sp. CF316, EJE52265, 99/387/682, 9E−11) Structural protein (*Brucella* phage Pr, AEY69748, 26/163/645, 6E−6 from PSI-BLAST)	Structure	—
45	35776–35060	—	238	Thioredoxin-disulfide reductase (*Desulfotomaculum kuznetsovii*, AEG16215, 37/121/302, 0.42)	—	—
46	36318–35833	—	161	Conserved hypothetical protein (*Vibrio parahaemolyticus*, EED25475, 17/62/174, 5.2)	—	—
47	37392–36331	—	353	Hypothetical protein (*Bifidobacterium longum* subsp. *infantis*, BAJ71205, 107/336/285, 6E−38)	Capsid protein	PF11651 (P22 coat protein - gene protein 5)
48	38502–37633	—	289	Hypothetical protein (*Haemophilus influenzae*, ZP_01797928, 49/185/284, 2E−3)	—	—
49	41060–38772	—	762	Hypothetical protein (*Actinobacillus minor*, EEV25019, 51/122/143, 1E−22)	—	—
50	41223–41047	—	58	Hypothetical protein (*Taylorella asinigenitalis*, AEP36223, 17/50/72, 0.12)	—	—
51	43324–41189	—	711	Hypothetical protein (*Desulfovibrio*, sp. 6_1_46AFAA, EGW50096, 185/652/717, 8E−61) 94 kDa protein (gp59) (*Escherichia* phage N4, ABK54420, 53/203/763, 4E−5)	Portal protein	—
52	44894–43335	—	519	Protein of unknown function DUF264 (*Thermosinus carboxydivorans*, EAX47548, 177/432/845, 4E−88)	Terminase, large subunit	PF03237 (Terminase-like family)
53	45174–44773	—	133	Hypothetical protein (*Ruminococcus obeum*, EDM86090, 21/79/645, 0.54)	—	—
54	45128–45220	+	30	Preprotein translocase, YajC subunit (*Rothia dentocariosa*, EFJ77340, 11/20/112, 0.40)		SP; TM

Table S1. Cont.

ORF	Position	Strand	Size, aa	Most significant hit in BLASTP against nr* (organism, GenBank accession no., no. of identical residues/aligned length/hit length, *E*-value)	Predicted function	Domain, family, signal peptide (SP), and transmembrane helix (TM)
55	45541–45167	—	124	Hypothetical protein (*Staphylococcus haemolyticus*, BAE05668, 34/114/126, 8E−3)	Terminase, small subunit	PF03592 (Terminase small subunit) (predicted by HHpred)
56	46026–45541	—	161	Family 24 glycoside hydrolase (*Arcobacter nitrofigilis*, ADG93351, 61/150/138, 3E−27)	Lysozyme	PF00959 (Phage lysozyme)
57	46136–46023	—	37	Hypothetical protein (Uncultured organism MedDCM-OCT-S09-C94, ADD96476, 20/36/68, 3E−6)		TM
58	46664–46281	—	127	Hypothetical protein (*Pseudomonas* phage PA11, YP_001294624, 53/123/119, 2E−28)		2 TM
59	46903–46664	—	79	Hypothetical protein (*Pseudomonas fulva*, AEF20734, 37/76/172, 6E−16)	—	—
60	47355–46900	—	151	Hypothetical protein (*Anolis carolinensis*, XP_003214352, 35/116/1347, 0.28)		TM
61	48439–47357	—	360	Methanesulfonate monooxygenase, hydroxylase alpha (large) subunit (*Candidatus* Puniceispirillum marinum 1322, ADE40352, 269/355/364, 0.0)	Methanesulfonate monooxygenase(?)	PF00848 (Ring hydroxylating alpha subunit)
62	48669–48439	—	76	Predicted protein (*Hordeum vulgare* subsp. *vulgare*, BAJ90467, 25/74/882, 0.15)	—	SP
63	49025–48666	—	119	Hypothetical protein (*Modestobacter marinus*, CCH87769, 31/55/194, 2E−14)	—	—
64	49369–49031	—	112	Hypothetical protein (*Geobacillus* sp. Y412MC61, YP_003251799, 40/110/109, 9E−4)	—	—
65	50628–49369	—	419	Hypothetical protein (*Synechococcus* phage S-RIM8 A.HR1, AFB17591, 275/413/436, 0. 0) Virion structural protein (Cyanophage S-TIM5, AEZ65693, 48/299/1472, 2E−18 from PSI-BLAST)	Structure	—
66	51297–50770	—	175	Hypothetical protein (*Aphanizomenon* sp. NH-5, ACG63808, 45/171/191, 8E−10)	—	—
67	51632–51297	—	111	Hypothetical protein (*Clostridium saccharolyticum*, ADL04451, 20/52/189, 0.17)	—	—
68	52641–51604	—	345	No hits	—	—
69	52759–52466	—	97	Hypothetical protein (*Synechococcus* phage S-CBS2, ADF42402, 47/90/96, 3E−20)	—	—
70	52959–52759	—	66	Hypothetical protein (Uncultured alpha proteobacterium, ADI19141, 15/51/65, 4.0)	—	SP
71	53175–52960	—	71	Hypothetical protein (*Vibrio shilonii*, ZP_01866522, 22/52/289, 1.1)	—	—
72	53665–53928	+	87	Metalloendopeptidase (*Zobellia galactanivorans*, CAZ96295, 13/27/1261, 6.9)	—	—
73	53907–53776	—	43	Hypothetical protein (*Trypanosoma brucei*, EAN79303, 14/36/1056, 4.1)	—	—
74	54168–54785	+	205	3-deoxy-D-manno-octulosonic-acid transferase (*Ruegeria* sp. R11, EEB72743, 27/76/398, 0.73)	—	—

*Additional BLASTP or PSI-BLAST results were presented for some ORFs to show the basis of functional annotations.
†Paralogous relationships were used for functional assignments of some tail structure genes if necessary (see also Fig. S3).

Table S2. Number of metagenome sequences putatively encoding a DNA polymerase with a partial DnaJ central domain

Sample name or ID in CAMERA	No. of sequences	Geographic position and description	Sample name or ID in CAMERA	No. of sequences	Geographic position and description
ANTARCTICAAQUATIC_ SMPL_ SITE232_0.1UM	2	Ace Lake, Antarctica	CAM_SMPL_SRA022117	1	Ross Sea, McMurdo Sound, Antarctica
ANTARCTICAAQUATIC_ SMPL_ SITE6	1	Ace Lake, Antarctica	CAM_SMPL_SRA022118	1	Ross Sea, McMurdo Sound, Antarctica
CAM_S_1336	72	North American West Coast	CAM_SMPL_SRA022119	1	Ross Sea, McMurdo Sound, Antarctica
CAM_S_596	4	Atlantic Ocean	CAM_SMPL_SRA022120	9	Ross Sea, McMurdo Sound, Antarctica
CAM_SMPL_000709	1	Coral, Wreck Reef	CAM_SMPL_SRA022121	22	Ross Sea, McMurdo Sound, Antarctica
CAM_SMPL_000719	3	Hydrothermal vent, Pacific	CAM_SMPL_SRA022122	6	Ross Sea, McMurdo Sound, Antarctica
CAM_SMPL_000720	4	Hydrothermal vent, Pacific	CAM_SMPL_SRA022123	14	Ross Sea, McMurdo Sound, Antarctica
CAM_SMPL_000721	7	Hydrothermal vent, Pacific	CAM_SMPL_SRA022124	21	Ross Sea, McMurdo Sound, Antarctica
CAM_SMPL_000722	12	Pacific Ocean	CAM_SMPL_SRA022142	1	Southern Ocean
CAM_SMPL_000723	9	Pacific Ocean	CAM_SMPL_SRA022147	3	East Antarctica Plain
CAM_SMPL_000724	6	Pacific Ocean	CAM_SMPL_SRA022150	9	CEAMARC-61, 34 nm from Antarctica coast
CAM_SMPL_000725	17	Pacific Ocean	CAM_SMPL_SRA022153	5	CEAMARC-8, bottom sample, 16 nm from Antarctica
CAM_SMPL_000726	2	Pacific Ocean	CAM_SMPL_SRA022155	2	CEAMARC-43, 6.3 nm from Antarctica
CAM_SMPL_000727	5	Pacific Ocean	CAM_SMPL_SRA022156	1	CEAMARC-43. 6.3 nm from Antarctica
CAM_SMPL_000816	11	Pacific Ocean	CAM_SMPL_SRA022158	2	CEAMARC-49A, Deep sample 1 nm from Mertz Glacier
CAM_SMPL_000825	4	Equatorial Atlantic	CAM_SMPL_SRA022159	3	CEAMARC-49A, Deep sample 1 nm from Mertz Glacier
CAM_SMPL_000833	3	Suboxic marine basin, Pacific	CAM_SMPL_SRA022161	3	CEAMARC-59- Deep sample
CAM_SMPL_000834	2	Pacific: Gulf of California	CAM_SMPL_SRA022166	2	Polynya
CAM_SMPL_000837	1	Pacific Ocean	CAM_SMPL_SRA022170	3	CEAMARC-70, 16 miles from Antarctica coast
CAM_SMPL_000841	1	Eel River	CAM_SMPL_SRA022172	1	Iceberg-4, 250 meters away from 35 km long and 18 km wide iceberg
CAM_SMPL_000844	1	Estuary, Atlantic	CAM_SMPL_SRA022173	1	Transect from Antarctic to Hobart
CAM_SMPL_000957	7	Pacific Ocean	CAM_SMPL_SRA022174	1	Open Ocean Transect
CAM_SMPL_000960	1	Atlantic Ocean	CAM_SMPL_SRA022175	1	Open Ocean Transect
CAM_SMPL_000961	2	Pacific Ocean: Southern California Bight	CAM_SMPL_SRA022178	1	Open Ocean Transect
CAM_SMPL_000966	10	Chesapeake bay station 858	CAM_SMPL_SRA022180	16	Ross Sea, McMurdo Sound, Antarctica
CAM_SMPL_000972	1	North Sea, Atlantic	CAM_SMPL_SRA022181	14	Ross Sea, McMurdo Sound, Antarctica
CAM_SMPL_000974	1	Atlantic	CAM_SMPL_SRA022192	2	Southern Ocean
CAM_SMPL_000990	13	Pacific Ocean: Southern California Bight	CAM_SMPL_SRA022193	4	Southern Ocean
CAM_SMPL_000994	7	Chesapeake bay station 858	CAM_SMPL_SRA022199	1	Southern Ocean
CAM_SMPL_001000	4	Chesapeake bay station 858	CAM_SMPL_SRA022200	1	Southern Ocean
CAM_SMPL_001003	3	Pacific Ocean	CAM_SMPL_SRA022201	6	Southern Ocean
CAM_SMPL_001014	3	Pacific Ocean: Southern California Bight	CAM_SMPL_SRA022202	1	Southern Ocean
CAM_SMPL_001589	1	Biosphere2 ocean	CAM_SMPL_SRA022206	1	Open Ocean Transect
CAM_SMPL_001739	8	Oregon Coast	GS000c	1	Sargasso Sea
CAM_SMPL_001740	2	Oregon Coast	GS000d	1	Sargasso Sea
CAM_SMPL_001742	1	Oregon Coast	GS002	16	North American East Coast

Table S2. Cont.

Sample name or ID in CAMERA	No. of sequences	Geographic position and description	Sample name or ID in CAMERA	No. of sequences	Geographic position and description
CAM_SMPL_001743	5	Oregon Coast	GS004	1	North American East Coast
CAM_SMPL_001745	2	Oregon Coast	GS005	6	North American East Coast
CAM_SMPL_001746	6	Oregon Coast	GS006	2	North American East Coast
CAM_SMPL_001748	1	Oregon Coast	GS007	1	North American East Coast
CAM_SMPL_001749	6	Oregon Coast	GS008	5	North American East Coast
CAM_SMPL_001750	4	Oregon Coast	GS009	3	North American East Coast
CAM_SMPL_001751	5	Oregon Coast	GS010	5	North American East Coast
CAM_SMPL_001752	4	Oregon Coast	GS013	2	North American East Coast
CAM_SMPL_001768	1	Oregon Coast	GS017	4	Caribbean Sea
CAM_SMPL_GS108	1	Coccos Keeling, Inside Lagoon	GS018	1	Caribbean Sea
CAM_SMPL_GS112	3	Indian Ocean	GS019	3	Caribbean Sea
CAM_SMPL_SRA022044	5	Botany Bay, Australia	GS023	3	Eastern Tropical Pacific
CAM_SMPL_SRA022077	1	Southern Ocean	GS025	7	Eastern Tropical Pacific
CAM_SMPL_SRA022079	1	New Comb Bay, Antarctica	GS026	2	Galapagos Islands
CAM_SMPL_SRA022081	1	Antarctica Open water	GS027	9	Galapagos Islands
CAM_SMPL_SRA022083	6	CEAMARC-27 Antarctica Shelf-64 nm off continent.	GS028	7	Galapagos Islands
CAM_SMPL_SRA022085	1	CEAMARC-61, 34 nm from Antarctica coast	GS029	10	Galapagos Islands
CAM_SMPL_SRA022087	1	CEAMARC-8, bottom sample, 16 nm from Antarctica	GS030	4	Galapagos Islands
CAM_SMPL_SRA022088	1	CEAMARC-8, bottom sample, 16 nm from Antarctica	GS031	4	Galapagos Islands
CAM_SMPL_SRA022089	1	CEAMARC-37, 18 nm from Antarctica	GS032	5	Galapagos Islands
CAM_SMPL_SRA022092	1	CEAMARC-47, 2nm from Antarctica and 9.5 nm from Mertz Glacier	GS033	17	Galapagos Islands
CAM_SMPL_SRA022093	6	CEAMARC-49A, Deep sample 1 nm from Mertz Glacier	GS034	1	Galapagos Islands
CAM_SMPL_SRA022094	1	CEAMARC-59	GS035	2	Galapagos Islands
CAM_SMPL_SRA022095	2	CEAMARC-59	GS036	2	Galapagos Islands
CAM_SMPL_SRA022096	2	CEAMARC-59- Deep sample	GS047	2	Tropical South Pacific
CAM_SMPL_SRA022097	1	Polyna-W-Compass-B Time series	GS048a	3	Polynesia Archipelagos
CAM_SMPL_SRA022099	3	CASO-15, Off continental shelf	GS110a	1	Indian Ocean
CAM_SMPL_SRA022101	1	CEAMARC-12, 10 nm from shore of Dumont d''urville, french station	GS115	1	Indian Ocean
CAM_SMPL_SRA022102	1	CEAMARC-70, 16 miles from Antarctica coast	GS116	1	Indian Ocean
CAM_SMPL_SRA022104	3	Iceberg-4, 250 meters away from 35 km long and 18 km wide iceberg	GS119	1	Indian Ocean
CAM_SMPL_SRA022107	1	Open Ocean Transect	HF_SMPL_BATS216_20M_SG	1	Bermuda time Series BATS Station 20m
CAM_SMPL_SRA022111	1	Open Ocean Transect	HF_SMPL_HOT179_25M_SG	1	Hawaii Ocean Time-series Station ALOHA
CAM_SMPL_SRA022112	9	Open Ocean Transect	HF_SMPL_HOT186_75M_GDNA	1	Hawaii Ocean Time-series Station ALOHA

Table S2. Cont.

Sample name or ID in CAMERA	No. of sequences	Geographic position and description	Sample name or ID in CAMERA	No. of sequences	Geographic position and description
CAM_SMPL_SRA022115	6	Ross Sea, McMurdo Sound, Antarctica	MOVE0902	2	Chesapeake Bay
CAM_SMPL_SRA022116	3	Ross Sea, McMurdo Sound, Antarctica	MOVE858	2	North American East Coast

Table S3. Highly assigned viruses in the binning of virome reads

Metagenome sample		CAM_SMPL_001011 (7.8/16.3)*	CAM_SMPL_000990 (12.8/22.4)	CAM_S_1336 (8.8/17.7)	GSIOVIR108 (8.0/17.2)	GSIOVIR112 (8.4/17.8)	GSIOVIR117 (12.4/23.0)	GSIOVIR122 (6.7/15.8)
Investigator		Sullivan, M. B.	Fuhrman, J. A.	Sullivan, M. B.	Williamson, S. J.	Williamson, S. J.	Williamson, S. J.	Williamson, S. J.
Sampling date		4–2-2009	19–8,13–5,11–3-2009	7–4-2009	3–8-2005	8–8-2005	9–9-2005	30–9-2005
Latitude/longitude		50N 145W	33.55N 118.4W	32.87N 117.25W	12.06S 96.53E	8.30S 80.23E	4.39S 55.31E	30.54S 40.25E
Temperature, °C		5.96	11.97–19.24	14.5	25.8	26.6	26.4	20.2
Salinity, ppt		32.56	33.30–33.48	—	32.4	32.5	35.5	35.8
Preparation		$FeCl_3$, DNase, CsCl	CsCl	See ref. 1	See ref. 2	See ref. 2	See ref. 2	See ref. 2
Binning Methods	Rankings	Viruses and their contributions						
BLASTN	1	HMO-2011 (25.3)†	HMO-2011 (13.3)	HMO-2011 (10.3)	HMO-2011 (13.8)	HMO-2011 (15.4)	S-SM2 (15.7)	HMO-2011 (14.4)
	2	HTVC010P (13.9)	HTVC008M (11.6)	S-SM2 (8.7)	HTVC010P (9.4)	HTVC010P (8.0)	HMO-2011 (13.8)	HTVC010P (8.6)
	3	HTVC008M (8.8)	HTVC010P (7.2)	HTVC010P (8.2)	P-SSM2 (5.8)	P-SSP2 (5.7)	HTVC010P (6.9)	S-CBS1 (7.5)
	4	HTVC019P (2.6)	S-SM2 (6.6)	HTVC008M (4.7)	HTVC008M (4.9)	P-SSM2 (5.2)	P-SSM2 (5.1)	S-CBS3 (4.7)
	5	S-SM2 (2.0)	HTVC019P (5.4)	P-SSM2 (3.9)	P-SSP2 (4.7)	P-SSP7 (4.7)	HTVC008M (4.9)	S-SM2 (4.1)
	6	P-SSM2 (1.9)	P-SSM2 (3.7)	S-TIM5 (3.7)	S-SM2 (4.4)	NATL1A-7 (4.5)	S-SSM7 (2.8)	NATL1A-7 (4.1)
	7	S-SSM7 (1.5)	HTVC011P (2.6)	Syn33 (2.5)	S-SSM7 (3.8)	HTVC008M (4.5)	HTVC011P (2.3)	P-SSP2 (3.9)
	8	SIO1 (1.4)	S-RSM4 (2.0)	S-MbCM6 (2.4)	P-SSP7 (3.8)	S-SM2 (3.9)	NATL1A-7 (2.3)	HTVC008M (3.5)
	9	GAP32 (1.2)	S-SSM7 (1.8)	Syn1 (2.3)	NATL1A-7 (3.0)	HTVC011P (3.8)	P-SSP2 (2.1)	P-SSP7 (3.1)
	10	Ehv 86 (1.2)	NATL1A-7 (1.7)	S-SSM7 (2.3)	P-RSM4 (3.0)	HTVC019P (3.3)	HTVC019P (1.9)	HTVC011P (3.1)
	11	P12053L (1.1)	Syn1 (1.5)	P12053L (1.8)	HTVC011P (2.4)	S-SSM7 (3.3)	P-SSP7 (1.8)	HTVC019P (2.6)
	12	BpV1 (1.0)	P-SSM7 (1.5)	SIO1 (1.8)	NATL2A-133 (2.2)	NATL2A-133 (2.4)	P-RSM4 (1.5)	NATL2A-133 (2.5)
	13	Herpesvirus 2 (1.0)	Syn33 (1.3)	GAP32 (1.8)	HTVC019P (2.1)	P-RSM4 (2.2)	S-RSM4 (1.4)	P-SSM2 (2.4)
	14	RaK2 (0.9)	BV-PW1 (1.1)	S-ShM2 (1.6)	P-SSM4 (1.9)	9515–10a (2.2)	NATL2A-133 (1.4)	9515–10a (1.5)
	15	BV-PW1 (0.9)	P-SSP2 (1.1)	S-RSM4 (1.5)	P-HM1 (1.7)	P-HM1 (1.6)	S-CBS1 (1.4)	S-TIM5 (1.5)
BLASTX	1	HMO-2011 (19.0)	HMO-2011 (13.6)	HMO-2011 (9.1)	HMO-2011 (16.1)	HMO-2011 (16.4)	HMO-2011 (15.4)	HMO-2011 (14.6)
	2	HTVC008M (9.6)	HTVC008M (10.7)	S-SM2 (6.5)	HTVC010P (7.5)	HTVC010P (5.9)	S-SM2 (11.1)	S-CBS1 (8.3)
	3	HTVC010P (9.5)	S-SM2 (5.7)	HTVC008M (6.1)	S-SM2 (4.8)	S-SM2 (4.3)	HTVC010P (5.4)	S-CBS3 (7.5)
	4	GAP32 (4.4)	HTVC010P (5.5)	HTVC010P (5.6)	HTVC008M (4.5)	P-SSP7 (4.3)	HTVC008M (4.9)	HTVC010P (5.3)
	5	RaK2 (3.0)	HTVC019P (5.3)	S-TIM5 (4.9)	P-SSP7 (3.6)	HTVC008M (4.1)	P-SSM2 (3.7)	S-SM2 (3.1)
	6	phiJL001 (2.7)	HTVC011P (3.5)	P-SSM2 (3.7)	S-SSM7 (3.5)	S-SSM7 (3.5)	S-SSM7 (3.1)	HTVC008M (2.9)
	7	YuA (2.6)	P-SSM2 (3.0)	GAP32 (3.5)	P-SSM2 (3.4)	P-SSM2 (3.4)	S-CBS1 (2.4)	S-TIM5 (2.9)
	8	S-SM2 (1.5)	P12024L (2.8)	S-SSM7 (2.9)	S-TIM5 (2.6)	P-SSP2 (3.3)	S-TIM5 (2.3)	P-SSP7 (2.8)
	9	S-TIM5 (1.8)	S-SSM7 (2.3)	RaK2 (2.2)	P-SSP2 (2.3)	HTVC011P (3.2)	P-SSP7 (2.2)	P-SSP2 (2.5)
	10	HTVC019P (1.8)	S-RSM4 (1.8)	P12024L (2.0)	HTVC019P (2.2)	S-TIM5 (3.1)	HTVC019P (2.0)	HTVC011P (2.4)
	11	S-SSM7 (1.5)	S-TIM5 (1.6)	P12053L (2.0)	phiJL001 (2.2)	HTVC019P (2.8)	HTVC011P (2.0)	NATL1A-7 (2.3)
	12	P12053L (1.5)	GAP32 (1.5)	Syn1 (1.6)	YuA (2.1)	phiJL001 (2.2)	P-SSP2 (1.6)	HTVC019P (2.0)
	13	P-SSM2 (1.2)	Syn1 (1.2)	S-CRM01 (1.6)	HTVC011P (2.0)	NATL1A-7 (2.1)	S-SM1 (1.5)	9515–10a (1.9)
	14	P-HM2 (0.9)	P-SSP7 (1.0)	S-RSM4 (1.5)	P-RSM4 (1.7)	9515–10a (2.1)	S-CBS3 (1.5)	S-CBS4 (1.8)
	15	HTVC011P (0.8)	P12053L (1.0)	HTVC019P (1.4)	NATL1A-7 (1.4)	S-CBS1 (1.9)	phiJL001 (1.4)	NATL2A-133 (1.7)

Colors were assigned to viruses according to host taxonomy. HMO-2011 (a SAR116 phage) is in red. All cyanophages are in green, whereas all pelagiphages (starting with "HTVC") are in blue. The remaining viruses are in black and their full names and hosts are as follows: BpV1, *Bathycoccus* sp. RCC1105 virus BpV1; BV-PW1, *Cafeteria roenbergensis* virus BV-PW1; Ehv 86, *Emiliania huxleyi* virus 86; GAP32, *Cronobacter* phage vB_CsaM_GAP32; Herpesvirus 2, Saimiriine herpesvirus 2; P12024L, *Persicivirga* phage P12024L; P12053L, *Celeribacter* phage P12053L; phiJL001, *Alphaproteobacteria* phage phiJL001; RaK2, *Enterobacteria* phage vB_KleM-RaK2; SIO1, Roseophage SIO1; YuA, *Pseudomonas* phage YuA.

*Numbers in parentheses (BLASTN/BLASTX, in %) are the proportions of virome reads assigned to viral genomes, which was calculated as follows: (number of reads assigned to viral genomes)/(total number of reads for the sample).

†Numbers in parentheses (in %) are the contributions of each virus genome calculated as follows: (number of reads assigned to each virus)/(number of reads assigned to viral genomes).

1. Hurwitz BL, Deng L, Poulos BT, Sullivan MB (2013) Evaluation of methods to concentrate and purify ocean virus communities through comparative, replicated metagenomics. *Environ Microbiol* 15(5):1428–1440.
2. Williamson SJ, et al. (2012) Metagenomic exploration of viruses throughout the Indian Ocean. *PLoS ONE* 7(10):e42047.