# Environmental Viral Genomes Shed New Light on Virus-Host Interactions in the Ocean

Yosuke Nishimura, Hiroyasu Watai, Takashi Honda, Tomoko Mihara, Kimiho Omae, Simon Roux, Romain Blanc-Mathieu, Keigo Yamamoto, Pascal Hingamp, Yoshihiko Sako, et al.

AMERICAN SOCIETY FOR MICROBIOLOGY | mSphere®

RESEARCH ARTICLE
Ecological and Evolutionary Science

# Environmental Viral Genomes Shed New Light on Virus-Host Interactions in the Ocean

**Yosuke Nishimura,[a,b] Hiroyasu Watai,[b] Takashi Honda,[b] Tomoko Mihara,[a] Kimiho Omae,[b] Simon Roux,[c] Romain Blanc-Mathieu,[a] Keigo Yamamoto,[d] Pascal Hingamp,[a,e] Yoshihiko Sako,[b] Matthew B. Sullivan,[c,f] Susumu Goto,[a] Hiroyuki Ogata,[a] Takashi Yoshida[b]**

Institute for Chemical Research, Kyoto University, Uji, Kyoto, Japan[a]; Graduate School of Agriculture, Kyoto University, Kyoto, Japan[b]; Department of Microbiology, the Ohio State University, Columbus, Ohio, USA[c]; Research Institute of Environment, Agriculture and Fisheries, Osaka Prefecture, Osaka, Japan[d]; CNRS, IGS UMR 7256, Aix Marseille Université, Marseille, France[e]; Department of Civil, Environmental and Geodetic Engineering, the Ohio State University, Columbus, Ohio, USA[f]

**ABSTRACT** Metagenomics has revealed the existence of numerous uncharacterized viral lineages, which are referred to as viral "dark matter." However, our knowledge regarding viral genomes is biased toward culturable viruses. In this study, we analyzed 1,600 (1,352 nonredundant) complete double-stranded DNA viral genomes (10 to 211 kb) assembled from 52 marine viromes. Together with 244 previously reported uncultured viral genomes, a genome-wide comparison delineated 617 genus-level operational taxonomic units (OTUs) for these environmental viral genomes (EVGs). Of these, 600 OTUs contained no representatives from known viruses, thus putatively corresponding to novel viral genera. Predicted hosts of the EVGs included major groups of marine prokaryotes, such as marine group II *Euryarchaeota* and SAR86, from which no viruses have been isolated to date, as well as *Flavobacteriaceae* and SAR116. Our analysis indicates that marine cyanophages are already well represented in genome databases and that one of the EVGs likely represents a new cyanophage lineage. Several EVGs encode many enzymes that appear to function for an efficient utilization of iron-sulfur clusters or to enhance host survival. This suggests that there is a selection pressure on these marine viruses to accumulate genes for specific viral propagation strategies. Finally, we revealed that EVGs contribute to a 4-fold increase in the recruitment of photic-zone viromes compared with the use of current reference viral genomes.

**IMPORTANCE** Viruses are diverse and play significant ecological roles in marine ecosystems. However, our knowledge of genome-level diversity in viruses is biased toward those isolated from few culturable hosts. Here, we determined 1,352 nonredundant complete viral genomes from marine environments. Lifting the uncertainty that clouds short incomplete sequences, whole-genome-wide analysis suggests that these environmental genomes represent hundreds of putative novel viral genera. Predicted hosts include dominant groups of marine bacteria and archaea with no isolated viruses to date. Some of the viral genomes encode many functionally related enzymes, suggesting a strong selection pressure on these marine viruses to control cellular metabolisms by accumulating genes.

**KEYWORDS** genome, marine ecosystem, metabolism, metagenomics, virus

Viruses outnumber microbes such as bacteria in the oceans (1), and the destructive lytic infections caused by viruses are thought to have crucial effects on energy and nutrient cycles driven by marine microorganisms (2, 3). Genomics-based research has

**Received** 7 December 2016 **Accepted** 2 February 2017 **Published** 1 March 2017

**Citation** Nishimura Y, Watai H, Honda T, Mihara T, Omae K, Roux S, Blanc-Mathieu R, Yamamoto K, Hingamp P, Sako Y, Sullivan MB, Goto S, Ogata H, Yoshida T. 2017. Environmental viral genomes shed new light on virus-host interactions in the ocean. mSphere 2:e00359-16. https://doi.org/10.1128/mSphere.00359-16.

**Editor** Hideyuki Tamaki, National Institute of Advanced Industrial Science and Technology



Address correspondence to Hiroyuki Ogata, ogata@kuicr.kyoto-u.ac.jp, or Takashi Yoshida, yoshiten@kais.kyoto-u.ac.jp.

been a powerful approach used to clarify the biology of viruses, including their infection strategies as well as their ecological significance (4–7). However, the diversity of viral genomes is still underrepresented in publically available genome databases (8, 9). For example, SAR11 (*Pelagibacterales*) and SAR116 are major marine prokaryotic components, but only four and one phage genomes have been sequenced for these bacteria, respectively (10, 11). Cyanophages, for which about 100 genomes have already been characterized, are the sole exception.

To address the issue of the paucity of viral genomic data, Roux et al. analyzed publicly available prokaryotic genome sequence data to mine marine and nonmarine viral genomes that have been sequenced along with the genomes of their hosts (12). They identified 12,498 viral DNA sequences (either long fragments or whole circular genomes) representing 264 predicted new genera.

Culture-independent viral metagenomics is also an effective research option for analyzing viral genomes in complex marine microbial communities (9, 13–16). A decisive advantage of viral metagenomics stems from the small genomes of viruses. Viral genomes have so far been assembled from the metagenomes of the following viral types: RNA viruses (17, 18), single-stranded DNA (ssDNA) viruses (19–26), and double-stranded DNA (dsDNA) viruses (27–29). Among these viruses, the genomes of dsDNA viruses have been the most difficult to assemble from metagenomes because of their relatively large genomes. However, recent advances in the construction of libraries (30), sequencing technologies, and bioinformatics software have resulted in the generation of larger assemblies. For example, 7 complete dsDNA viral genomes have been reported for a hypersaline lake (27), 18 for the deep-sea hydrothermal vent plumes (28), and 54 for glacial cryoconite holes (29). An interesting approach involved the construction of metagenomic fosmid libraries from virus-infected prokaryotes, which revealed 1 (31) and 42 (32) complete viral DNA genomes for solar salterns and 208 marine tailed-phage genomes (33). These studies indicated that marine viral metagenomics investigations have advanced from focusing on environmental genetics (i.e., collections of genes) to analyzing environmental genomics (i.e., collections of complete genomes), helping to unveil the evolutionary histories, life cycles, and metabolic strategies of individual viruses. In this study, we analyzed nine novel marine viral metagenomes (i.e., viromes) generated using a benchtop Illumina/MiSeq sequencer as well as previously published large-scale viromes (9). We identified 1,352 nonredundant complete viral genomes, the vast majority of which corresponded to previously unidentified viral lineages.

## RESULTS AND DISCUSSION

**Choice of assemblers.** We generated nine viromes (Osaka Bay viromes [OBVs]; 8.5 M read pairs; 2.4 Gbp) from water samples collected over a 24-h period in Osaka Bay, Japan (see Materials and Methods). We first compared four assemblers (SPAdes [34], metaSPAdes, IDBA-UD [35], and Ray Meta [36]) regarding their ability to assemble viromes. SPAdes, metaSPAdes, and IDBA-UD clearly outperformed Ray Meta in terms of the total size of >10-kb contigs (Table 1). Of the first three assemblers, SPAdes (11.9 Mb) produced the largest assemblies (i.e., metaSPAdes, 6.8 Mb; IDBA-UD, 5.3 Mb). Regarding assembly error rates assessed by REAPR (37), SPAdes (8.48 regions/kb), metaSPAdes (8.73), and IDBA-UD (8.80) had similar error rates, which were slightly higher than that of Ray Meta (6.42). Most (99.97%) of these errors were short insertion/deletions (REAPR type 1 and type 3 errors), while there were very few (0 to 0.00662 regions/kb) scaffolding errors (type 2 and type 4 errors) (Table 1). On the basis of these results, we chose SPAdes as the best assembler for the following analyses.

**Forty-six genomes assembled from the Osaka Bay viromes.** Given that the nine samples were collected at the same location over a short period and that the reads were relatively long (i.e., 2 × 150 or 2 × 300 bp), a coassembly consisting of the pooled nine samples was also prepared. The coassembly resulted in 879 contigs (>10 kb) that likely originated from dsDNA viruses (see Materials and Methods). Of these, 46 (28.5 to 192 kb; average, 54.2 kb) were assembled in a circular form (see Fig. S1 in the

**TABLE 1** Comparison of four assemblers

| Parameter | Value | | | |
|---|---|---|---|---|
| | SPAdes | metaSPAdes | IDBA | Ray |
| Assembly size (for contigs >10 kb) | 11,869,699 | 6,818,200 | 5,264,822 | 471,387 |
| REAPR error types[a] | | | | |
| FCD error (type 1) | 0.01490 | 0.01045 | 0.01083 | 0.00470 |
| FCD error over a gap (type 2) | 0.00000 | 0.00000 | 0.00000 | 0.00000 |
| Low-coverage error (type 3) | 8.46559 | 8.71562 | 8.78596 | 6.41814 |
| Low-coverage error over a gap (type 4) | 0.00414 | 0.00662 | 0.00000 | 0.00000 |
| Total no. of errors | 8.48463 | 8.73268 | 8.79678 | 6.42284 |

[a]Error values are presented as the number of times the error occurs per 1 kb for contigs longer than 1 kb. Type 1 and 3 errors were associated with short insertion/deletions. Type 2 and 4 errors were associated with scaffolding errors (e.g., chimeric assemblies). FCD, fragment coverage distribution.

supplemental material). Thus, we refer to these 46 contigs as environmental viral genomes (EVGs).

The EVGs did not contain any scaffolding errors (REAPR type 2 and type 4 errors), indicating high structural integrity for the contigs. To further assess the integrity of these EVGs, we mapped the contigs assembled from individual viromes on the EVGs. Of the 46 EVGs, 16 were totally covered by the contigs from individual assemblies, thus decreasing the possibility of artefactual chimeras due to coassembly for these 16 EVGs. The remaining 30 EVGs contained 1 to 24 regions (229 in total) that were supported only by coassembly and were not observed in the individually assembled contigs. We randomly selected 21 such weakly supported regions and tested the coassemblies by PCR assays (using the environmental DNA samples as a template) and sequencing. The results verified all of the tested regions of the coassembled contigs (Fig. S2A). Furthermore, 18 of the 46 EVGs exhibited complete or nearly complete genomic colinearity with closely related reference genomes (Fig. S2B; see Materials and Methods for the definition of genomic colinearity) or with the other independently determined EVGs described below (Fig. S2C). These results further corroborated the accuracy of the overall structure of the EVG assemblies.

**SNPs and nucleotide diversity.** Each of the individual EVGs likely corresponds to genomes of closely related viruses because the sequence assemblies were obtained from environmental viral populations. To assess the genetic diversity of each EVG, we analyzed single nucleotide polymorphisms (SNPs) and calculated the nucleotide diversity of each EVG. Nucleotide sites containing SNPs that were supported by at least one read were present in genomes at a rate of 0.558 to 7.897% (median, 2.473%) (see Table S1A in the supplemental material). The nucleotide diversity of EVGs was 0.073 to 1.734% (median, 0.423%). These results are within the ranges for genomes from the same viral species (38). We conclude that each of the EVGs represents a consensus genome of a viral species.

**One thousand five hundred genomes assembled from the *Tara* Oceans viromes.** Prompted by the detection of 46 OBV-EVGs in a modest sequencing effort, we applied our genome assembly and complete genome identification protocol to the *Tara* Oceans viromes (TOV), which consist of 43 viromes representing 26 oceanic locations (9). Given the wide geographic areas and seasons covered by these samples and the large volume of sequence data for individual TOV samples (i.e., average, 50 M reads; 2 × 100 bp), we assembled these 43 viromes individually. We obtained 1,554 TOV-EVGs (i.e., circular complete contigs, 10 to 211 kb) with a predicted viral origin. Only 64 were detected as complete in the previously reported original TOV assemblies (9), and 85.6% of the remaining EVGs (i.e., 1,275 EVGs) were detected in the original assemblies as smaller contigs with less than half the size of the contigs in these new assemblies. Clustering on the basis of the nucleotide sequence identity among the OBV-/TOV-EVGs resulted in 1,352 nonredundant complete genomes (i.e., 46 OBV-EVGs and 1,306 TOV-EVGs).

After discarding possible eukaryotic virus genomes, we obtained 1,567 complete genomes that were likely of prokaryotic dsDNA viral origin (45 OBV-EVGs and 1,522 TOV-EVGs; see Materials and Methods). Of these genomes, 1,404 (89.6%) were predicted to encode homologs of tailed-virus hallmark proteins (i.e., terminase large subunits [89.5%], major capsid proteins [34.4%], or portal proteins [60.2%]), suggesting that the genomes were derived from tailed viruses. Of the remaining 163 EVGs, 72 were predicted to encode integrase homologs.

**Diversity of environmental viral genomes.** To investigate the global novelty offered by culture-independent viral genome sequencing efforts, we compiled a set of 1,811 EVGs (>10 kb) composed of the 45 OBV-EVGs, the 1,522 TOV-EVGs, and 244 EVGs from other studies (29, 33, 39). We also compiled a set of 2,429 prokaryotic dsDNA viral genomes (>10 kb) from cultured viruses, which are referred to here as reference viral genomes (RVGs) (Fig. S3; Table S1B).

We first generated a viral proteomic tree (40) on the basis of genomic similarity scores (denoted by $S_G$) derived from tBLASTx scores. The $S_G$ value is 1 when two genomes in a comparison are identical and decreases to 0 when a tBLASTx search fails to detect any sequence similarities. The viral proteomic tree revealed a clear separation between EVG and RVG clades, with most of the EVGs grouped with other EVGs and not with the RVGs (Fig. 1). We also used average linkage clustering of the EVGs/RVGs to delineate operational taxonomic units (i.e., genomic OTUs [gOTUs]) on the basis of the $S_G$ value, with six different clustering cutoff values (Fig. 2 for cutoff $S_G = 0.15$ and Fig. S4 for all cutoff values from 0.1 to 0.9). The EVG-containing gOTUs outnumbered the RVG-containing gOTUs at five of six tested $S_G$ cutoff values. For example, we observed a 1.6-fold EVG-to-RVG gOTU overrepresentation ratio at $S_G = 0.3$ (Fig. S4A). The proteomic tree and comparative genome maps are available at http://www.genome.jp/viptree/EVG2017.

**Genus-level operational taxonomic units.** We analyzed the viral taxonomic classification of the RVGs and evaluated the correspondence between viral genera and gOTUs using different $S_G$ cutoff values. The $S_G$ values between 0.07 and 0.2 were associated with relatively high adjusted Rand index values (i.e., > 0.79), and $S_G = 0.15$ (adjusted Rand index = 0.847) was determined to be the most accurate cutoff value for a genus-level classification (Fig. S5). With this cutoff value, we obtained 1,087 gOTUs for the EVGs/RVGs. The 2,429 RVGs were distributed across 487 gOTUs, whereas the 1,811 EVGs were distributed across 617 gOTUs (i.e., 1.27-fold-higher richness), with only 1.4% of the total gOTUs containing both EVGs and RVGs (Fig. 2B). Therefore, the EVGs potentially represent 600 new viral genera. Of the 600 gOTUs, 497 were composed exclusively of OBV-/TOV-EVGs. To complement this analysis, we added 11,779 mined viral genomes (MVGs; genome sizes, >10 kb) (12). We observed only a limited overlap of gOTUs among the EVGs, RVGs, and MVGs (i.e., only two gOTUs with sequences from all three sets), and 590 genus-level gOTUs remained specific to the EVGs.

**Virus-host interactions. (i) Host prediction on the basis of genomic similarity.** Because of the dissimilarity between EVGs and RVGs, host predictions on the basis of similarities to known viral genomes (i.e., RVGs) were difficult to make. Using information regarding RVG hosts, we calculated an optimal $S_G$ threshold that separated viruses into those that infect similar hosts and those that do not. The threshold was a $S_G$ value of >0.2937 (>90% precision) for the prediction of pairs of viruses infecting host organisms that are evolutionarily related at the genus level (Fig. S6). With this cutoff, we predicted host groups for only 29 of 1,811 EVGs (2 OBV-EVGs, 13 TOV-EVGs, and 14 other EVGs; Table S1C). Of the 29 EVGs, 18, 10, and 1 were predicted to be cyanophages, *Pelagibacter* phages, and *Pseudoalteromonas* phages, respectively. Two additional host prediction methods based on tRNA genes and clustered regularly interspaced short palindromic repeat (CRISPR) spacer sequences (41) failed to predict possible hosts for the EVGs. However, the physical linkage of genes on the EVGs provided additional clues about their hosts and biology. In the following sections, we describe virus-host interactions inferred from the genomic contexts of EVGs.

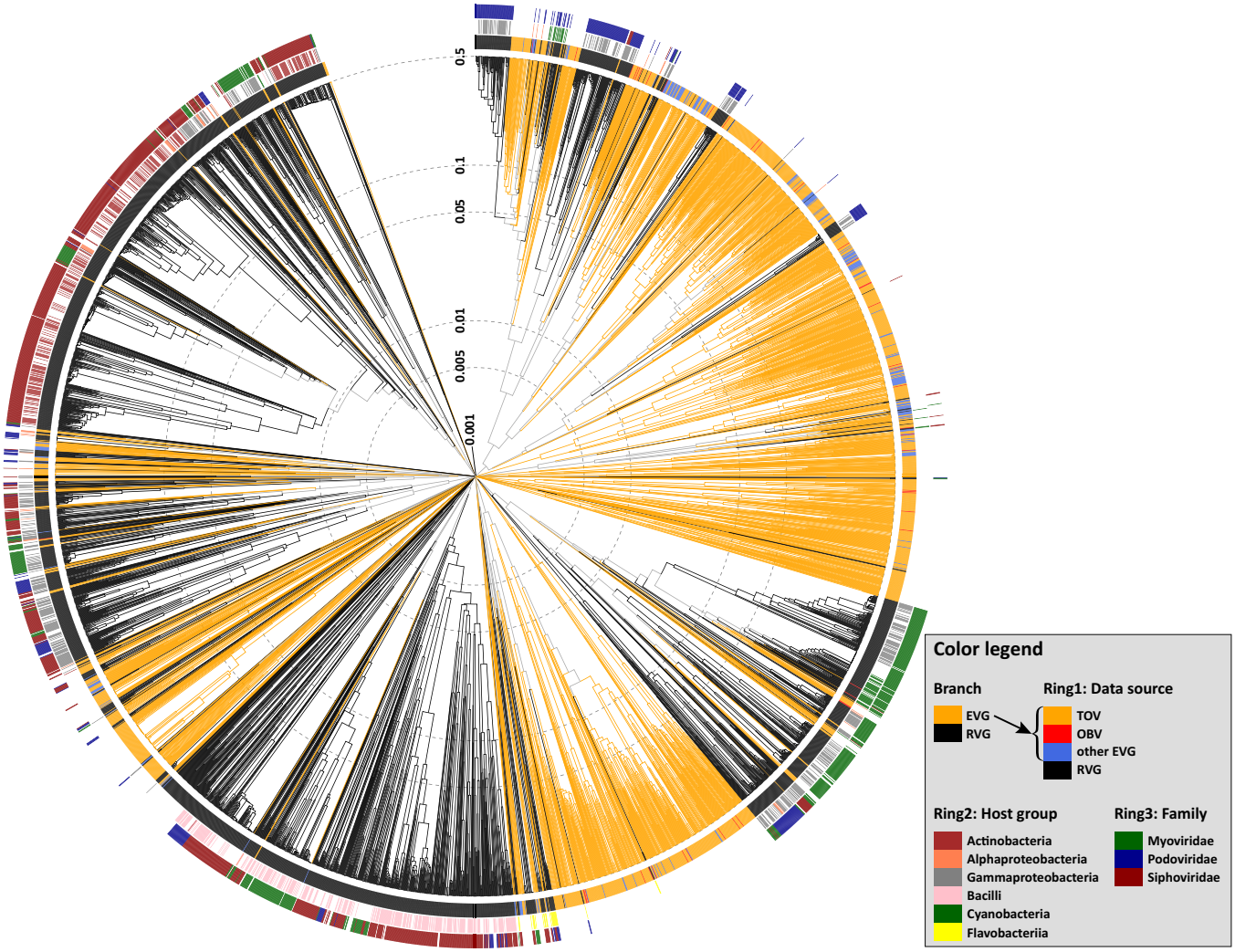


**FIG 1** Proteomic tree. The dendrogram represents proteome-wide similarity relationships among 4,240 prokaryotic dsDNA virus genomes. Branches are colored orange (EVG, environmental viral genome) or black (RVG, reference viral genome), and branch lengths are indicated using a logarithmic scale. TOV, *Tara* Oceans viromes; OBV, Osaka Bay viromes. The tree is midpoint rooted. Rings outside the dendrogram represent, from inside to outside, sources of genome data, taxonomic groups of known hosts, and viral family classifications.

**(ii) MGII viruses.** Four previously undescribed lineages that likely infect uncultivable marine group II (MGII) *Euryarchaeota* species were revealed in the proteomic tree. These four clades were exclusively composed of OBV/TOV-EVGs, with 18, 13, 23, and 4 EVGs in clades 1, 2, 3, and 4, respectively (Fig. 3). Phylogenetic analyses of the DNA polymerases encoded in those EVGs strongly support the existence of the four clades identified in the proteomic tree (Fig. 4A). These clades were grouped with homologs from haloviruses and euryarchaea. Identifications of archaeal hosts for the 58 EVGs were also supported by their gene content. Of the genes in the EVGs with homologs in cellular organisms, an average of 36.1% (14.3 to 60.0%) were most closely matched to archaeal proteins. Additionally, one to five tailed-virus structural protein homologs were detected in each of the EVGs (Table S1D). Archaeal tailed viruses have been detected only in *Euryarchaeota* species (42), with the exception of a provirus of *Nitrososphaera viennensis* (*Thaumarchaeota*) isolated from soil (43).

We observed that the EVGs contained chaperonin genes (Fig. 3A). Thirty-eight of the 58 EVGs encode chaperonin homologs, even though chaperonin genes have rarely been identified in sequenced viral genomes (i.e., only 7 of the 2,429 RVGs encode chaperonins). In some viruses, chaperonins, which are usually provided by the hosts, are responsible for the correct assembly of viral particles (44). All 18 EVGs in clade 1 encode

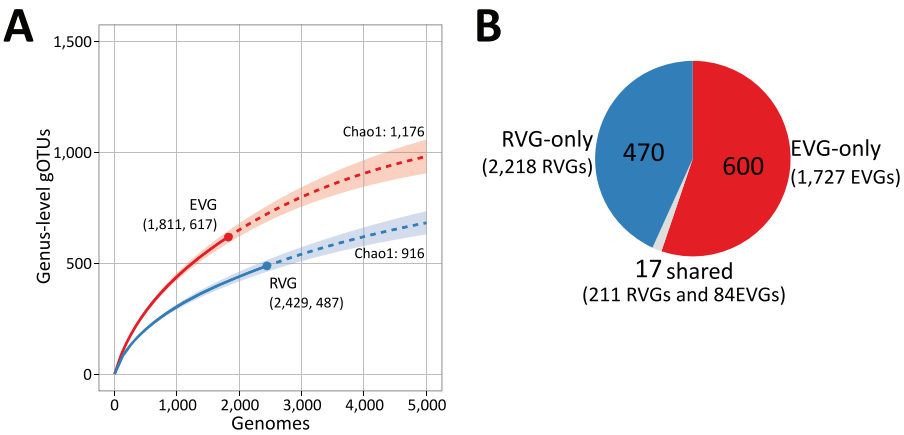


**FIG 2** Genus-level genomic OTU (gOTU) richness. The genome-wide similarity score ($S_G$) cutoff for clustering was set to 0.15 (i.e., viral-genus-level cutoff). The EVGs and RVGs were clustered together, and subsets of the EVGs and RVGs were then constructed by extracting each member. (A) Rarefaction curves for the number of gOTUs. Rarefaction curves are presented with shading representing 95% confidence intervals obtained from 100 bootstrap replicates using the R package iNEXT (107). Dashed curves represent extrapolations to 5,000 genome sequences. Numbers in parentheses represent the number of genomes and gOTUs. Chao1 richness estimates for the EVGs and RVGs are indicated. (B) Proportions of genus-level gOTU clusters. Colors represent the following cluster categories: EVG-only clusters (red), RVG-only clusters (blue), and shared clusters (gray).

archaeon-type chaperonin homologs (i.e., thermosome; group II chaperonin), while 20 EVGs in clades 2 to 4 encode bacterium-type chaperonin homologs (i.e., GroEL; group I chaperonin). We detected both groups of chaperonin genes in the MGII genomes (45, 46). The group I and group II chaperonin sequences from the EVGs were grouped with these MGII chaperonins (Fig. 4B), suggesting that MGII species serve as hosts for these environmental viruses.

The following three archaeal taxa are abundant in the marine water column: marine group I *Thaumarchaeota* (MGI), MGII, and marine group III *Euryarchaeota* (MGIII) (47). Of these, currently cultivated representatives exist only in MGI (48). The members of MGII are abundant in particle-rich surface waters (49, 50), while those of MGIII have been observed almost exclusively in deep seas (47). A recent study revealed that MGII members can temporarily become the most abundant (up to 40%) prokaryotic components in the days following a spring bloom (51). The 58 EVGs were derived from surface or deep chlorophyll maximum viromes, suggesting their photic-zone habitat. These observations and the genomic context described above suggest that the 58 EVGs represent genomes of tailed viruses infecting MGII *Euryarchaeota* species.

**(iii) A SAR86 phage encoding IscU.** Iron-sulfur (Fe-S) cluster proteins are involved in a variety of biological processes, including gene regulation, electron transfer, catalytic reactions, and oxygen-iron sensing (52). In a previous study, Fe-S cluster assembly protein genes (e.g., *sufA* and *iscU*) were identified as auxiliary metabolic genes (AMGs) of photic-zone viromes (15, 53). However, the lack of complete genome data hindered further characterizations of the viruses carrying these genes. We identified 16 OBV/TOV-EVGs with Fe-S cluster assembly protein genes, including 14 EVGs containing an Fe-S cluster A-type carrier (ATC) gene (54) and 6 EVGs carrying the IscU gene (Fig. 5A). These genomes are scattered across four groups of viruses in the proteomic tree, and many of their close relatives (i.e., other EVGs and *Pelagibacter* phage HTVC008M in Fig. 5A) do not contain these genes. The ATC and IscU proteins function as scaffolds in which Fe and S atoms are assembled into Fe-S clusters (55, 56). Phylogenetic trees of IscU (Fig. 5B) and ATC (Fig. S7) revealed that all six EVG-encoded IscU genes form a clade with gammaproteobacterial homologs. Of these, an IscU gene from OBV_N00005 was phylogenetically closely related to homologs from SAR86 (57), suggesting that SAR86 members represent potential hosts for OBV_N00005. The prevalence of these viral genes in photic-zone viromes (15) appears to be linked to the wide distribution of these bacteria.

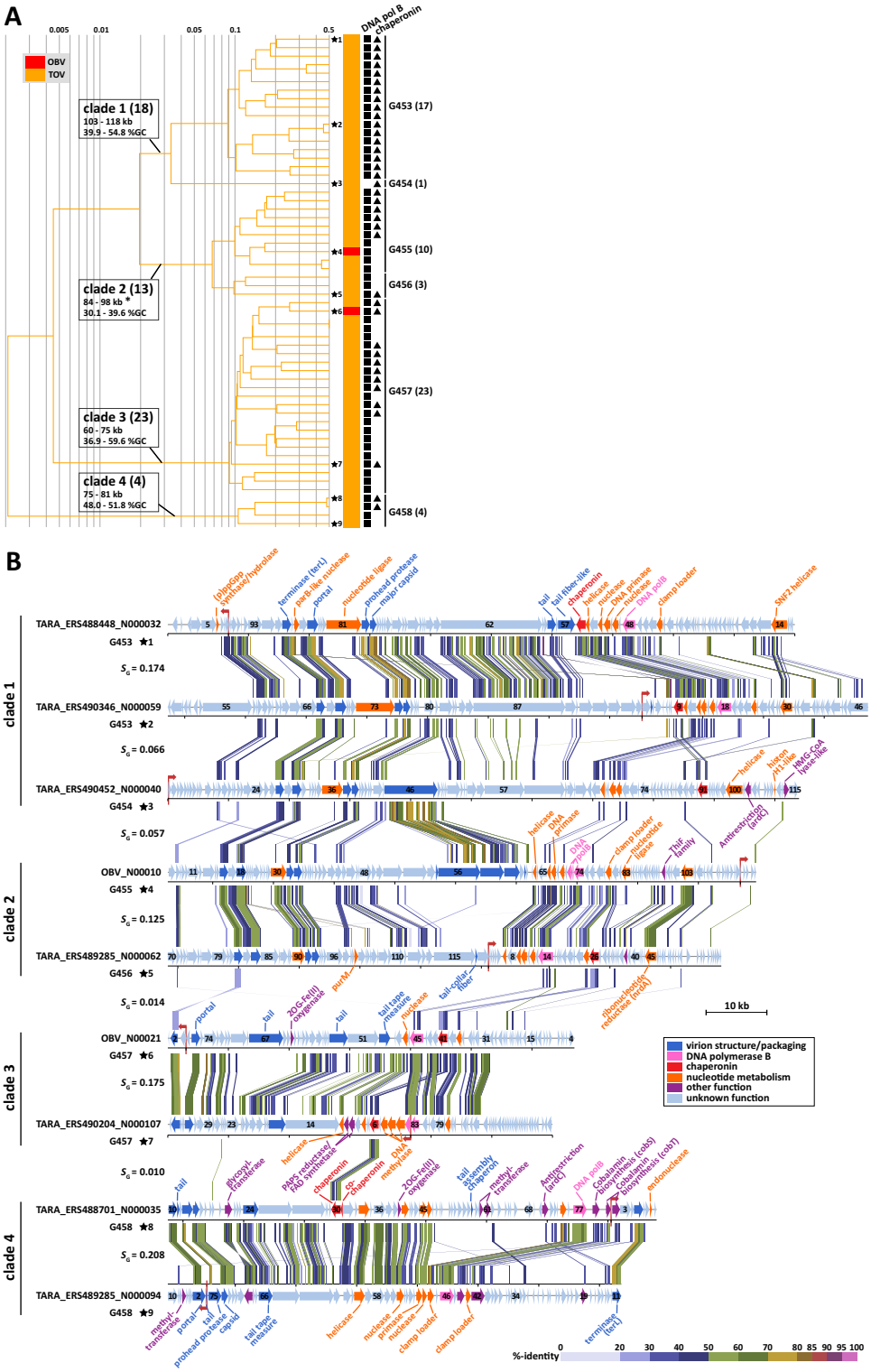


**FIG 3** Fifty-eight putative archaeal virus genomes. (A) Part of the proteomic tree with 2 OBV-EVGs (red) and 56 TOV-EVGs (orange), predicted to be derived from euryarchaeal tailed viruses infecting marine group II (MGII) species. Genomes with genes encoding DNA polymerase B (squares) and chaperonin (triangles) are indicated. Clade names and genus-level gOTUs are indicated. Numbers in parentheses represent the number of genomes of each clade or gOTU. The ranges of genome sizes and percent G+C contents for each clade are presented, with the exception that clade 2 includes a long contig (121 kb; asterisk). Branch lengths are logarithmically scaled from the root of the entire proteomic tree in Fig. 1. (B) Genome map of nine archaeal viral genomes that are indicated by stars in panel A. The sequences are circularly permuted and/or reversed. Red arrows indicate the original start position of the sequences. Putative gene functions are indicated. All tBLASTx alignments are represented by colored lines between the two genomes. The color scale represents tBLASTx percent identity.

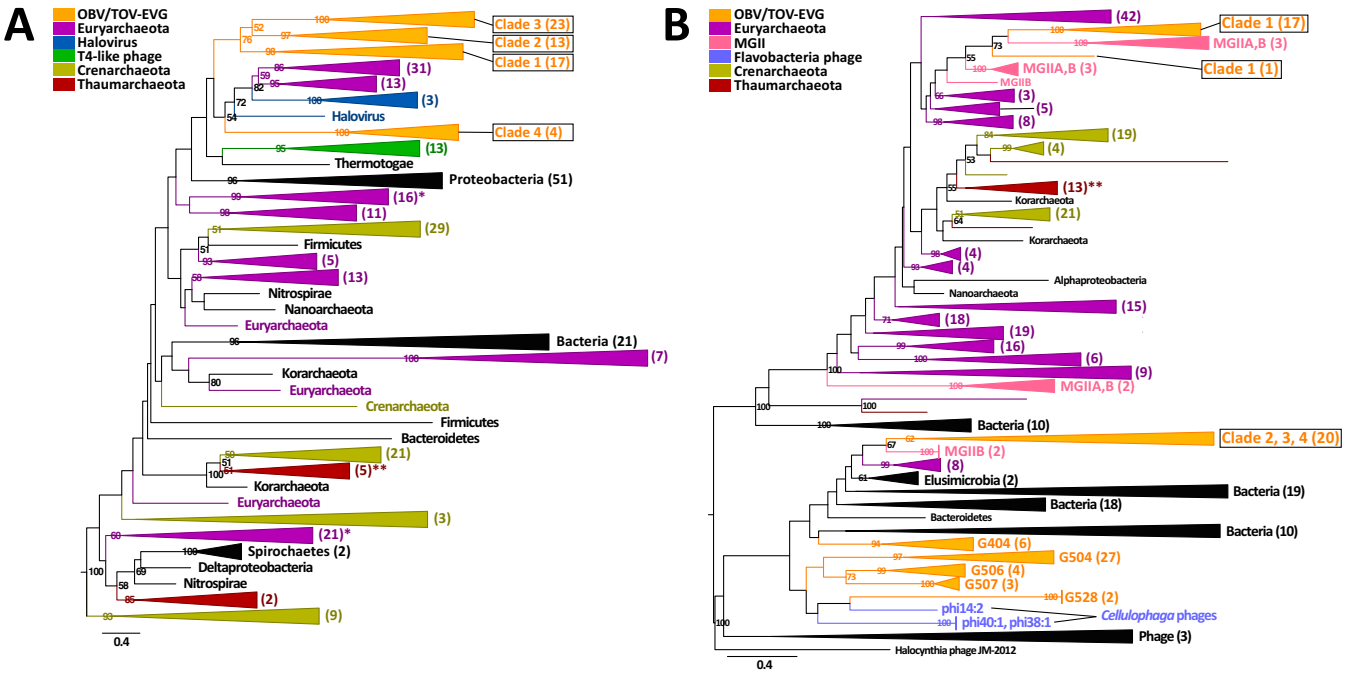


**FIG 4** Gene phylogenetic trees of DNA polymerase B and chaperonin. (A) Maximum likelihood tree of DNA polymerase B. The tree is rooted by four distant bacterial sequences (not shown) and includes 348 sequences. (B) Maximum likelihood tree of chaperonin. The tree is midpoint rooted and includes 381 sequences. In panels A and B, numbers in parentheses represent the number of sequences in each collapsed node. Colors represent taxonomies. Asterisks indicate collapsed nodes that include MGII (*) and MGI (**) sequences. The scale bar refers to the estimated number of amino acid substitutions per site. Numbers near the nodes represent bootstrap percentages of >50%. MGIIA and MGIIB indicate sequences from reported genomes (45 and 46, respectively).

In addition to the Fe-S scaffolding proteins, some of the EVGs encode several Fe-S cluster proteins that use Fe-S clusters as prosthetic groups, such as radical S-adenosylmethionine (SAM) superfamily enzymes (58) and CRISPR-associated Cas4 exonucleases (59, 60). The EVGs also encode proteins involved in the metabolism of Fe-S cluster proteins, such as glutaredoxins (Grx), the phenylacetyl-coenzyme A oxygenase component PaaD (61, 62), and ClpP, which is a serine protease targeting Fe-S cluster proteins (15). A notable example is the T4-like TARA_ERS488813_N000010 (183 kb; group *iv* in Fig. 5A), which includes an ATC gene, 12 genes for radical SAM superfamily enzymes, and *cas4*, *grx*, and *paaD* (16 genes in total; Table S1E). Other T4-like EVGs encoding ATC and/or IscU proteins contain two to seven additional Fe-S-related genes. Of these genes, *paaD* has not been previously associated with a virally encoded protein and thus represents a novel AMG. These observations suggest that Fe-S cluster assembly proteins encoded in these viral genomes function as a part of Fe-S cluster-related metabolic processes involving not only host proteins but also many virally encoded proteins.

**(iv) A novel cyanophage lineage.** The RVG set included 114 cyanophage genomes, which were grouped into 17 viral-genus-level gOTUs. There were no other RVGs classified into these gOTUs. Of these 17 gOTUs, 5 included 34 EVGs (i.e., 3 OBV-EVGs, 16 TOV-EVGs, and 15 previously described EVGs [33]), which are likely to have been derived from cyanophages or their relatives. Screening all EVGs with 11 photosynthesis-related AMGs (see Materials and Methods) led to the identification of 11 predicted cyanophage EVGs, of which 10 were included in the gOTUs mentioned above (Table S1F). The remaining EVG (i.e., TARA_ERS489084_N000023; gOTU G241), which carries *psbA* and *hli*, formed a singleton gOTU and represents a new cyanophage group. To characterize the approximate abundances of these 18 cyanophage gOTUs (149 genomes; Table 2), we mapped the TOV and OBV reads on these putative cyanophage genomes. The following five most abundant gOTUs represented >98% of the total cyanophage content: (i) G386, including T4-like myoviruses (35.1%); (ii) G14, including podoviruses (33.7%); (iii) G234, including a siphovirus and dwarf myoviruses (23.4%); (iv)

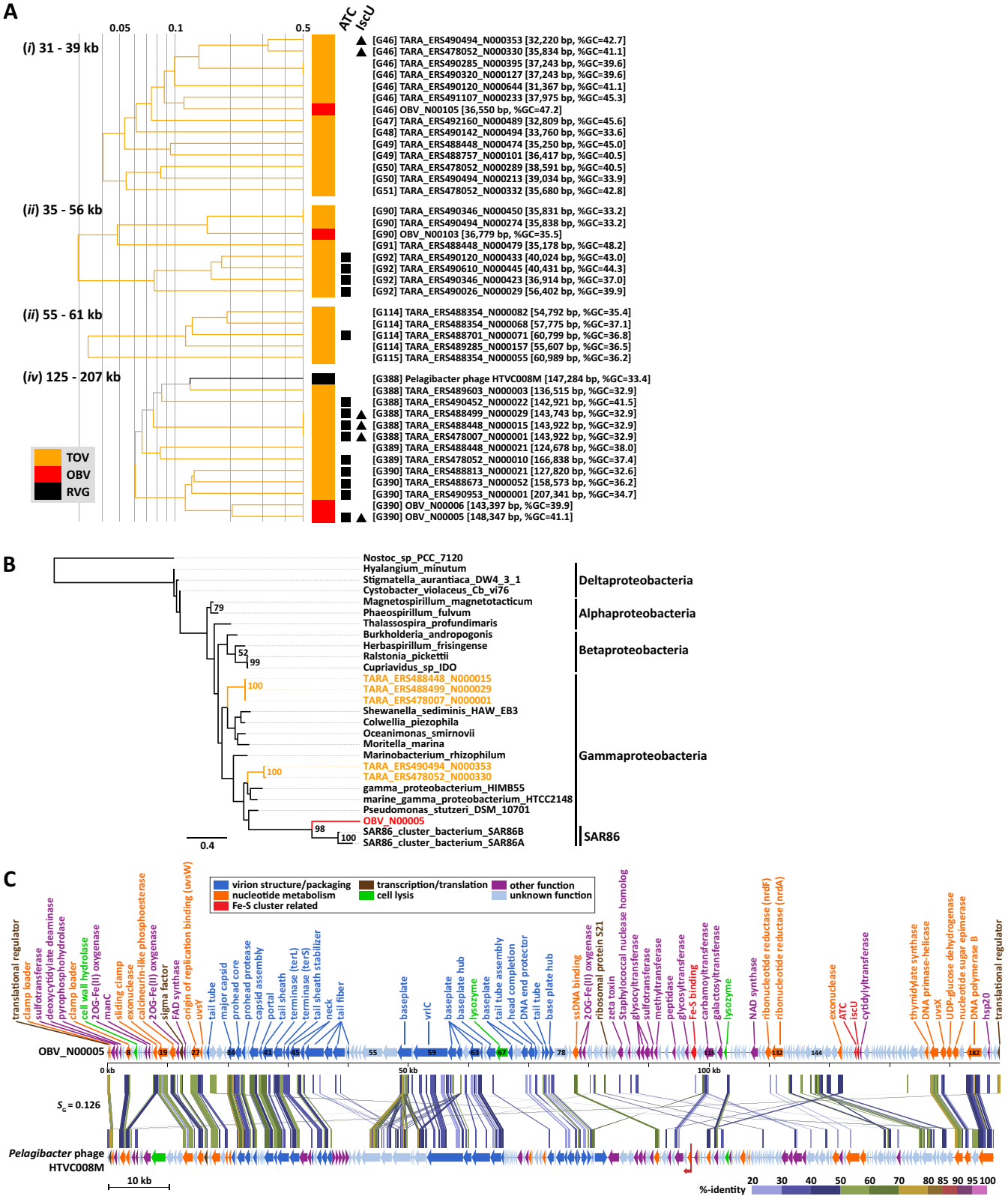


**FIG 5** Genomes with Fe-S cluster assembly-related genes. (A) Four parts of the proteomic tree with genomes carrying Fe-S cluster assembly genes (i.e., ATC [■] and IscU [▲] genes). Branch lengths are logarithmically scaled as described for Fig. 3A. Genus-level gOTUs and genome identifiers (IDs), lengths, and percent G+C compositions are indicated. (B) Maximum likelihood tree of IscU genes. The tree contains protein sequences encoded in OBV_N00005 (red), five TOV-EVGs (orange), and 21 *Proteobacteria* and cyanobacterial genomes (black). The scale bar refers to the estimated number of amino acid substitutions per site. Numbers close to the nodes represent bootstrap percentages of >50%. The tree is rooted by the cyanobacterial *Nostoc* species sequence. (C) Genome map of OBV_N00085 and *Pelagibacter* phage HTVC008M. The HTVC008M sequence is circularly permuted at 97,000 bp and reversed. A red arrow indicates the original

(Continued on next page)

**TABLE 2** Photosynthetic genes and abundance of cyanophage genomes

| Genus-level gOTU | No. of EVGs | No. of RVGs | Photosynthetic gene(s) in EVG | Photosynthetic gene(s) in RVG | FPKM[a] | % abundance[b] | Most abundant RVG |
|---|---|---|---|---|---|---|---|
| G14 | 7 | 21 | *hli*, *psbA* | *hli*, *psbA* | 3,484.7 | **33.7** | *Prochlorococcus* phage P-GSP1 |
| G15 | 1 | 1 | *hli* | | 334.6 | **3.2** | *Prochlorococcus* phage P-RSP2 |
| G234 | 16 | 3 | | | 2,419.5 | **23.4** | Cyanophage MED4-117 |
| G237 | 1 | 1 | | | 35 | 0.3 | *Synechococcus* phage S-CBS4 |
| G238 | 6 | 1 | *hli*, *ptoX* | | 340.7 | **3.3** | *Synechococcus* phage S-EIV1 |
| G241 | 1 | 0 | *hli*, *psbA* | | 48.6 | 0.5 | |
| G242 | 1 | 1 | *hli*, *psbA* | *hli* | 10.5 | 0.1 | *Synechococcus* phage S-CBS2 |
| G243 | 0 | 1 | | | 3.3 | 0 | Cyanophage P-SS2 |
| G277 | 0 | 1 | | *nblA* | 0 | 0 | *Planktothrix* phage PaV-LD |
| G278 | 0 | 2 | | *nblA* | 1.8 | 0 | *Microcystis aeruginosa* phage Ma-LMM01 |
| G386 | 2 | 72 | *cpeT*, *hli*, *petE*, *psbA*, *psbD*, *ptoX* | *cpeT*, *hli*, *ho1*, *pcyA*, *pebS*, *petE*, *petF*, *psbA*, *psbD*, *ptoX* | 3,622.7 | **35.1** | *Synechococcus* phage S-SM2 |
| G387 | 0 | 1 | | *hli*, *psbA* | 0.1 | 0 | *Synechococcus* phage S-CRM01 |
| G402 | 0 | 1 | | *hli*, *petE*, *psbA*, *psbD*, *ptoX* | 21.8 | 0.2 | Cyanophage S-TIM5 |
| G769 | 0 | 2 | | | 0 | 0 | Cyanophage PP |
| G770 | 0 | 1 | | | 0 | 0 | *Anabaena* phage A-4L |
| G771 | 0 | 1 | | | 0 | 0 | *Phormidium* phage Pf-WMP4 |
| G818 | 0 | 2 | | *nblA* | 0 | 0 | *Phormidium* phage MIS-PhV1B |
| G1074 | 0 | 2 | | | 2.1 | 0 | *Synechococcus* phage S-CBS1 |

[a]The FPKM for each gOTU was calculated as the average of the sum of FPKMs of the genomes in the gOTU across different samples. In calculating the average, the nine OBV samples were treated as a single sample to avoid any bias toward a local region.
[b]Abundance represents a normalized FPKM (the sum is equal to 100), and values of >3% are indicated in bold.

G238, including *Synechococcus* phage S-EIV1 (63) (3.3%); and (v) G15, including *Prochlorococcus* phage P-RSP2 (3.2%) (Tables 2 and S1B for the list of genomes). Thus, marine cyanophage genomes are well represented in the current databases.

**(v) Diverse marine *Bacteroidetes* phages.** *Bacteroidetes* is one of the most abundant bacterial phyla in the oceans (e.g., 30% of the bacterioplankton during phytoplankton blooms) (64). Members of this phylum are involved in the decomposition and remineralization of phytoplankton biomass (65). A recent study revealed that an algal bloom is followed by the presence of a rapid succession of diverse *Flavobacteriaceae* bacteria (64). To the best of our knowledge, the genomes of the following 38 phages infecting marine *Bacteroidetes* (*Flavobacteriaceae*) have been described: psychrophilic *Flavobacterium* phage 11b (66), *Croceibacter* phage P2559S (67), 2 *Persicivirga* phages (68), 31 *Cellulophaga* phages (69), *Flavobacterium* phage 1/32 (70), and 2 *Polaribacter* phages (71). *Polaribacter* was reported to be abundant following a spring phytoplankton bloom (64), while *Cellulophaga* phages (31 of 38) likely represent a "rare biosphere" rather than abundant marine phages (69). We detected two groups (i.e., groups 1 and 2) of putative *Flavobacteriaceae* phage genomes (i.e., 5 RVGs, 8 OBV-EVGs, 222 TOV-EVGs, and 9 EVGs from another study; Fig. 6). Group 1 and group 2 consisted of 29 and 25 gOTUs, respectively. Of these, 23 and 21 gOTUs were exclusively composed of OBV/TOV-EVGs. Of the genes in the OBV/TOV-EVGs having homologs in cellular organisms, 64.4% (15.8% to 92.3%) on average for the members of group 1 and 32.4% (10.5% to 59.1%) on average for the members of group 2 were most similar to *Bacteroidetes* genes. For example, the gene20 sequence of OBV_N00025 (group 2, G506; Fig. 6B) was most similar to the RNA polymerase sigma-70 factor sequence of a *Flavobacteria* strain from marine surface water (WP_009781949; *Leeuwenhoekiella blandensis*; E value = 1e-30) (72, 73). Genomes of these groups also encode conserved virion structural or morphogenetic proteins. For the members of group 1, we detected putative portal gene homologs in 148 EVGs (93.7%) and prohead protease homologs in 145 EVGs

**FIG 5** Legend (Continued)
start position of the HTVC008M sequence. Putative gene functions of OBV_N00005 and HTVC008M (described in reference 10) are indicated. All tBLASTx alignments are represented by colored lines between the two genomes. The color scale represents tBLASTx percent identity. FAD, flavin adenine dinucleotide; NAD, nicotinamide adenine dinucleotide.

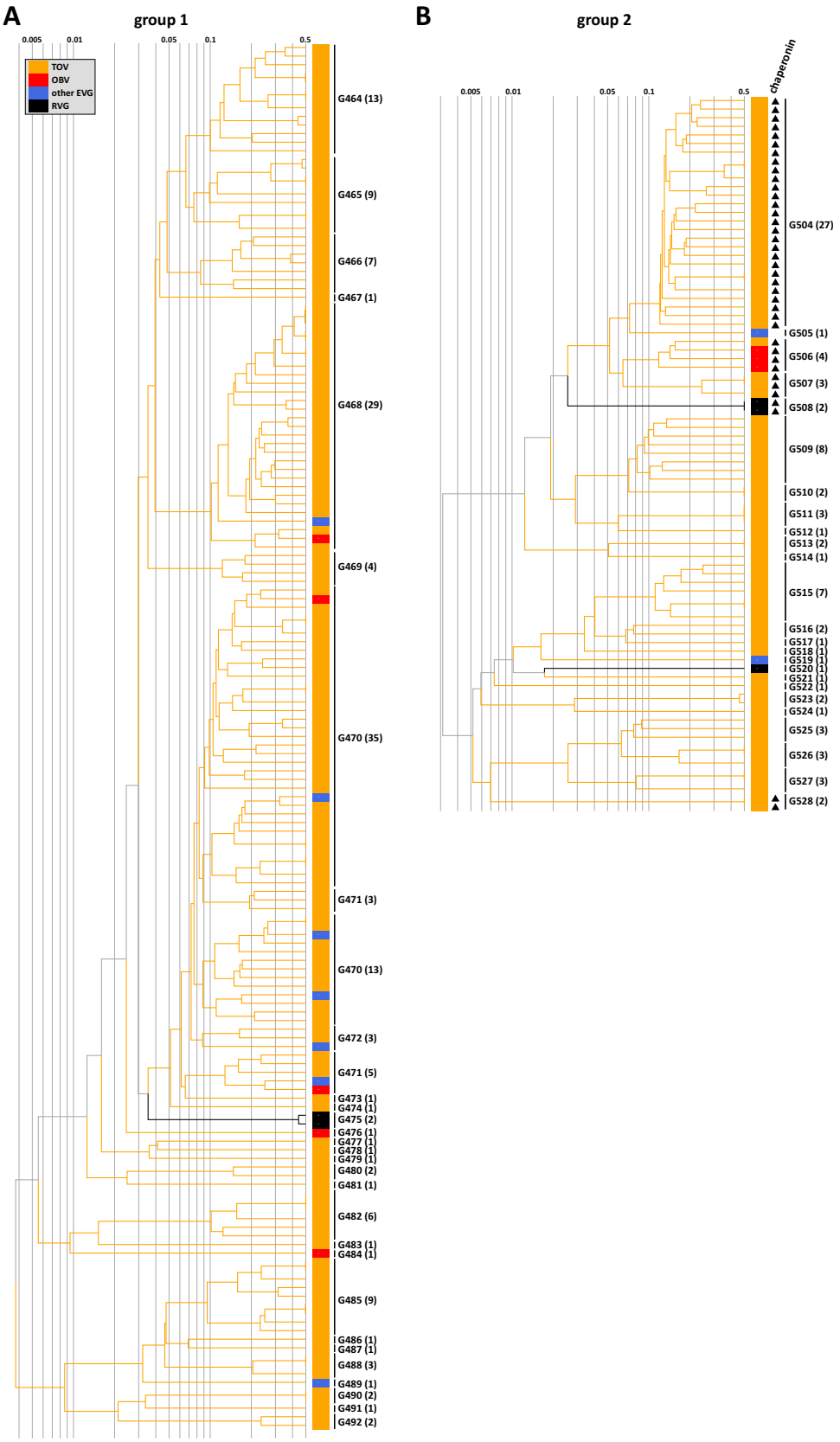


**FIG 6** Two parts of the proteomic tree with EVGs of putative *Flavobacteriaceae* phages. Branch lengths are logarithmically scaled as described for Fig. 3A. Genus-level gOTUs are indicated. Numbers in parentheses represent the number of genomes in each gOTU. (A) Group 1 distributed in 29 gOTUs, including two *Persicivirga* phages (black), 5 OBV-EVGs (red), 147 TOV-EVGs (orange), and 7 other EVGs (blue). (B) Group 2

(Continued on next page)

(91.8%). For the members of group 2, we detected homologs of two to six structural proteins of *Cellulophaga* phage phi38:1 (i.e., a member of group 2) in 78 EVGs (100%). Additionally, we detected GroEL homologs in 36 EVGs of the members of group 2 (Fig. 6B) which were phylogenetically related to the homologs in *Cellulophaga* phages (Fig. 4B). Therefore, these EVGs probably correspond to viruses of *Flavobacteriaceae* species and may prove to be useful genetic markers for studying viruses affecting bacterial decomposer communities.

**(vi) A virus potentially enhancing the adaptation of its host.** Isocitrate lyase (AceA) and malate synthase (AceB) catalyze two reactions in the glyoxylate shunt, which bypasses the $CO_2$-generating steps of the tricarboxylic acid cycle and enables the net assimilation of carbon from acetyl-coenzyme A (acetyl-CoA), leading to gluconeogenesis (i.e., generation of glucose) and cell growth (74, 75). We identified an *aceBA* operon in a TOV-EVG (TARA_ERS478052_N000008; 179 kb; see Table S1G for gene description) that included homologs of three structural genes from T4-like phages. Our genomic similarity and gene composition analysis did not provide any clue about the host of this virus. A previous study detected *aceA* and *aceB* in ocean viromes (14), but this is the first time, to our knowledge, that an *aceBA* operon has been observed in a complete viral genome. The genome also encoded six enzymes (i.e., Gmd, WcaG, ManC, NeuA, KdsA, and WaaG) for the biosynthesis of lipopolysaccharides (LPS) and capsular polysaccharides, important components of bacterial cell wall and capsule (76, 77). Previous studies identified LPS synthesis genes in temperate and lytic phages and proposed that these genes function to modify cell surface compositions to prevent other viruses from attaching to the cell during the lysogenic or pseudolysogenic phase, in the latter of which a lytic process is halted due to suboptimal host cell growth (78, 79). Following this "lock out" hypothesis, the *aceBA*-carrying virus (i.e., TARA_ERS478052_N000008) should have a provirus phase, and AceA and AceB may function to promote the growth of host cells. gene40 of the TOV-EVG was predicted to encode a homolog of zeta toxin proteins (Table S1G) thought to be involved in a toxin-antitoxin system. Toxin-antitoxin systems enhance the stability of plasmids and prophages by postsegregational killing (80). This corroborates the existence of a lysogenic phase of this virus, though there was no other evidence for lysogeny in the viral genome. It should be further noted that the function of LPS is not limited to protection of the cell from viral infection but that LPS on bacterial outer membrane confers tolerance of temperature and oxidative stresses as well as resistance to antibiotics (81). Therefore, *aceBA* and the cell wall biogenesis genes in the TOV-EVG may contribute to a host's survival and environmental adaptation by altering carbon metabolism and cell surface compositions during the lysogenic phase.

**(vii) Temperate phages of SAR116.** Our analysis also unveiled phage genomes likely infecting members of the SAR116 clade, which is one of the most abundant marine bacterial lineages (11). OBV_N00085 (40 kb) and three closely related TOV-EVGs (40 to 41 kb; $S_G$ for OBV_N00085 = 0.25 to 0.26) exhibited clear collinearity with an approximately 40-kb genomic segment from "*Candidatus* Puniceispirillum marinum" IMCC1322 of the SAR116 clade (class: *Alphaproteobacteria*) (Fig. 7 for OBV_N00085) (82). This suggests that these EVGs are derived from temperate phages infecting SAR116 or related bacteria. These genomes consistently encode integrases.

**(viii) Phages related to SAR11 phages.** Seven EVGs (OBV_N00073, three TOV-EVGs, and three other EVGs; 39 to 42 kb) exhibited high genome-wide sequence similarities to *Pelagibacter* podovirus HTVC019P (10) ($S_G$ = 0.34 to 0.44; 42 kb; a dot plot comparing OBV_N00073 and HTVC019P is presented in Fig. S2B). On the basis of the $S_G$ values (i.e., >0.2937; estimated precision, >90%), we predict that these EVGs infect host species in the genus *Pelagibacter* (Table S1C). Another *Pelagibacter* podovirus (i.e.,

**FIG 6** Legend (Continued)
distributed in 25 genus-level gOTUs, including two *Cellulophaga* phages (phi40:1 and phi38:1; black; G508), IAS virus (black; G520), 3 OBV-EVGs (red), 75 TOV-EVGs (orange), and 2 other EVGs (blue). Genomes encoding chaperonins are indicated by a triangle.

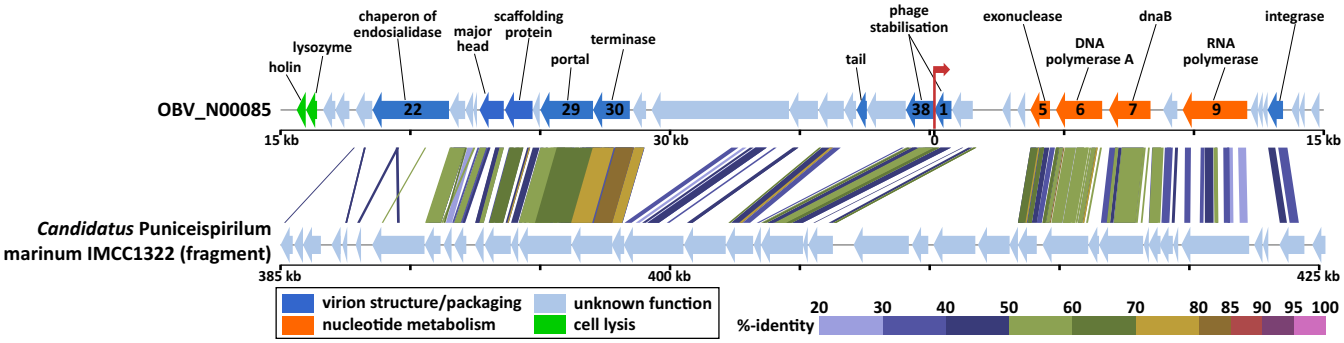


**FIG 7** Genomic alignment between the whole sequence of OBV_N00085 and a genomic region (385,000 to 425,000 bp) of "*Candidatus* Puniceispirillum marinum" IMCC1322. The OBV_N00085 sequence is circularly permuted at 15,000 bp for clarity, and a red arrow indicates the original start position of the sequence. Putative gene functions and function categories of OBV_N00085 are indicated by texts and colors. All tBLASTx alignments are presented. The color scale represents tBLASTx percent identity.

HTVC010P), which is believed to be a member of the most abundant virus subfamily in the biosphere (10), was classified in a different group of the proteomic tree together with 102 EVGs (OBV_N00107, 77 TOV-EVGs, and 24 other EVGs; 31 to 73 kb; Fig. S8). These 102 genomes carry homologs of HTVC010P structural protein genes. The G+C content of the HTVC010P genome is 32% (10), while the EVGs of this group contain higher levels of G+C content (i.e., 31 to 57%). Low levels of G+C content (i.e., 28.6 to 32.3%) are a common genomic feature of the SAR11 clade members (83). Since high levels of correlation between the G+C content of prokaryotic viruses and that of their hosts were previously reported (84, 85), the variation in the levels of their G+C content suggests that the viruses in this group infect a wide range of host species.

**Environmental viral genomes as a reference during marine virome analyses.** We mapped protein sequences and raw reads from independently generated photic virome data (i.e., the Pacific Ocean viromes [POV]) (86) on the RVGs and EVGs. The RVG set recruited 4.70% of the POV proteins, while the EVG/RVG union set recruited 22.6% of the proteins (i.e., a 4.8-fold increase; Fig. 8A). At the nucleotide sequence level, the RVG set recruited 1.02% of the POV reads, while the EVG/RVG union set recruited 4.20%

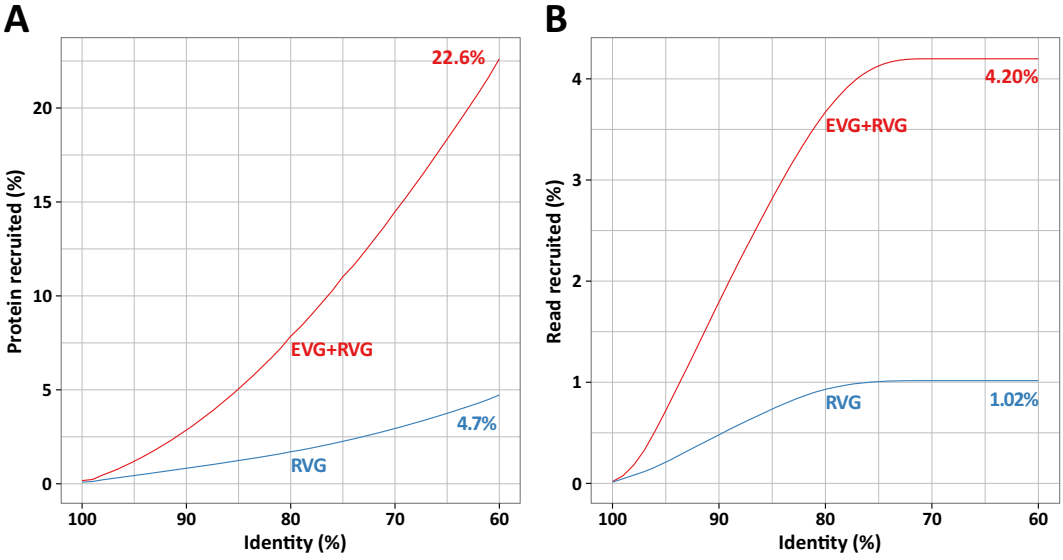


**FIG 8** Recruitment of photic POV sequences to RVGs (blue) and to a pool of EVGs and RVGs (red). Mappings were performed with tBLASTn (proteins) and BLASTn (reads). In both mappings, the initial filtering of hits involved an E value of <1e−3, and an additional filtering was based on ≥60% identity and ≥80% alignment length of the query sequence. (A) Recruitment of proteins. (B) Recruitment of reads.

of the reads (i.e., a 4.1-fold increase; Fig. 8B). Thus, the EVGs serve as an effective additional reference viral genome data set for exploring viromes from photic oceans.

**Conclusion.** From the assemblies of 52 marine viromes, we obtained 1,567 circular complete genomes that are most likely of prokaryotic dsDNA viral origin. The acquisition of the complete genome sequences helped classify the viral lineages and provided important clues about their hosts and metabolisms. The genome-based clustering of the metagenome-derived viral genomes together with previously reported ones suggests that 600 of the 617 gOTUs represent new genera of prokaryotic viruses. Additionally, they contain greater genome richness than the reference genomes of cultured prokaryotic viruses that have so far been sequenced. Our analyses also predicted the relationships among the EVGs and the major groups of marine prokaryotes, for which no viruses have been isolated (i.e., MGII and SAR86). Given the lack of isolation of viruses, the physiological features of the sequenced EVGs are unclear. However, some of the newly identified EVGs carried functionally related AMGs, such as those encoding proteins related to Fe-S clusters (16 genes) and to carbon assimilation/cell wall biogenesis enzymes (8 genes). These AMGs may function to coordinate the supply/recycling of Fe-S clusters and to enhance host adaptation during the lysogenic cycle. Previous studies also revealed that cyanophages carry multiple functionally linked photosynthesis and lipopolysaccharide synthesis genes for their efficient replication (79, 87, 88). Therefore, viral survival strategies in marine viruses involving many functionally related AMGs appear to target not only the biosynthesis of molecular building blocks (e.g., nucleotides) but also diverse metabolic and cellular processes.

## MATERIALS AND METHODS

**Sample preparation and sequencing.** Seawater samples (9 × 4 liters) were collected at a 5-m depth at the entrance of Osaka Bay (34°19′28″N, 135°7′15″E), Japan, every 3 h for 24 h on 25 and 26 August 2014. Seawater was filtered through a 142-mm-diameter (3.0-$\mu$m-pore-size) polycarbonate membrane (Millipore, Billerica, MA) and then through a 142-mm-diameter (0.22-$\mu$m-pore-size) Durapore polyvinylidene fluoride membrane (Millipore). The filtrates were stored at 4°C prior to treatments. The viruses in the filtrate were concentrated by $FeCl_3$ precipitation (89) and purified using DNase and a CsCl density centrifugation step (90). The DNA was then extracted as previously described (91). Libraries were prepared using a Nextera XT DNA sample preparation kit (Illumina, San Diego, CA) according to the manufacturer's protocol, except that we used 0.25 ng viral DNA. Samples were sequenced with a MiSeq sequencing system and MiSeq V2 (2 × 150 bp; five samples) or V3 (2 × 300 bp; four samples) reagent kits (Illumina, San Diego, CA).

**Genome assembly and error estimation.** Nine OBVs were individually assembled using the following four assemblers: SPAdes, metaSPAdes (http://bioinf.spbau.ru/spades), IDBA-UD, and Ray Meta. SPAdes 3.1.1 was used with default k-mer lengths as well as the accompanying BayesHammer (92) and MismatchCorrector. The metaSPAdes 3.7.0 program was used with default k-mer lengths and BayesHammer. The IDBA-UD 1.1.1 program was used with fixed multiple k-mer lengths (24 to 124, increased by 10 for 2 × 300 bp reads; 24 to 84, increased by 10 for 2 × 150 bp reads) and the option of a preread correction with a minimum contig length of 300 bp. Ray Meta 2.3.1 was used with a fixed k-mer length ($k = 41$). Additionally, we used scaffolds of these assemblies, which we called contigs for simplicity. The REAPR 1.0.18 program was used to assess the quality of the assemblies. This program reports four types of errors categorized as short insertion/deletion errors (i.e., types 1 and 3) or scaffolding errors (i.e., types 2 and 4).

Nine sets of OBV reads were also coassembled by SPAdes with the same settings as described above. We determined that a contig was circular (i.e., complete) if its 5′ and 3′ terminal regions were nearly identical (i.e., >94% and ≥50 bp). We identified 40 circular contigs (>10 kb) satisfying this condition. A coassembly involving the merged paired-end reads generated by FLASH was also prepared (93). We included the merged and remaining unmerged reads for the assembly. With this second coassembly, we detected 34 circular contigs (>10 kb), of which 6 were not detected in the first coassembly. We incorporated these 6 contigs in our data set, and we ultimately obtained 934 OBV contigs (>10 kb), including 46 circular ones. Forty-three TOV samples were similarly analyzed, except that the sequence assemblies were prepared sample by sample and only with raw reads (i.e., not from merged paired-end reads). Code for circular contig detection is downloadable at ftp://ftp.genome.jp/pub/db/community/EVG2017.

**Gene prediction and annotation.** Gene predictions were completed using MetaGeneMark (94). Homology searches were conducted using BLASTp against the NCBI-nr database (E value, <1e−5), RPS-BLAST against the COG database (as of April 2015; E value, <1e−4), and HMMER against the Pfam (as of May 2015; E value, <1e−4) and TIGRFAMs (release 15; E value, <1e−4) databases. For predictions of tailed-virus hallmark genes and integrase genes, we used HHsearch (E value, <1e−9) against the Pfam database after constructing query hidden Markov models (HMMs) using jackhmmer (part of the HMMER package) with default settings (95, 96). We also used PSI-BLAST to identify homologs of specific genes.

**Discrimination of viral and prokaryotic contigs and PCR assays.** We used a newly developed method (see Text S1 in the supplemental material) and VirSorter (97) to distinguish between viral and prokaryotic contigs. We discarded all contigs predicted to be of prokaryotic origin by either or both methods. Finally, 879 of the 934 OBV contigs (including 46 circular ones) and 1,554 of the 1,618 TOV circular contigs were considered to originate from viruses.

We conducted PCR assays for 21 weakly supported regions in four randomly selected OBV circular contigs (i.e., OBV_N00005, OBV_N00020, OBV_N00021, and OBV_N00023; see Fig. S2A in the supplemental material). Primer sequences are provided in Table S1H in the supplemental material.

**Genomic colinearity.** Colinearity was evaluated on the basis of the percentage of OBV-EVG genes that had orthologous relationships with the most closely related genome (i.e., $B_g$ in Fig. S2B). If ≥60% of the OBV-EVG genes had orthologs in the closest relative, we considered the OBV-EVG to exhibit nearly complete genomic colinearity. Eighteen OBV-EVGs (39%) were observed to exhibit complete or nearly complete colinearity with other viral genomes. Additionally, we identified colinear genomic regions using MCScanX (98) and calculated the percentage of OBV-EVG genes in these regions (i.e., $C_g$ in Fig. S2B).

**Quality control of reads.** We used raw reads for the above assemblies, but the reads underwent a quality-control screening before being back-mapped to contigs with the following procedure: (i) duplicated reads were removed using FastUniq (99); (ii) paired-end reads were merged with FLASH, and the merged and unmerged reads were kept; (iii) reads were removed if the percentage of high-quality nucleotide positions (i.e., quality score >30) was <80%; and (iv) reads were removed if the sum of the lengths of ambiguous nucleotide positions and low-complexity regions detected by DUST was >40% of the total length. If one of the paired-end reads was removed in step iii or step iv, the mate was retained as a single read.

**Detection of single nucleotide polymorphisms and calculation of nucleotide diversity.** To detect SNPs and assess nucleotide diversity, we mapped quality-controlled reads on contigs using the Bowtie 2 program. To minimize the inclusion of sequencing errors among the mapped nucleotides, we considered only high-quality nucleotides (i.e., quality score, >30). Nucleotide diversity was defined as previously described (100) and was calculated using equation 1 of a published method (101). The SNPs were detected for positions with ≥5× sequence coverage using the following six criteria: (i) at least one read, (ii) at least two reads, (iii) more than 10% coverage, (iv) more than 20% coverage, (v) more than 10% coverage or at least two reads, and (vi) more than 10% coverage and at least two reads. These criteria were applied to the second-most-frequent nucleotide at each position.

**Redundancy of obtained environmental viral genomes.** To detect redundancies among TOV-EVGs and OBV-EVGs, an all-against-all BLASTn search was conducted. We merged high-scoring segment pairs (HSPs) for each resulting pair, and if the merged HSPs covered ≥80% of the shorter EVG, with ≥95% average identity, the EVGs were considered redundant. Nonredundant EVGs were obtained by single-linkage clustering of these redundant pairs.

**Viral genomes.** We first compiled 46 OBV-EVGs, 1,554 TOV-EVGs, and 247 EVGs from three projects, including 192 complete contigs (33), 54 circular consensus genomes (29), and a complete viral genome obtained from samples from single amplified genomes (SAG) (39). The RVGs were retrieved from RefSeq (release 75; March 2016), EBI Genomes Pages (May 2015), and CAMERA. We selected dsDNA viral genomes that were larger than 10 kb. We then removed the genomes of eukaryotic viruses identified using the GenomeNet Virus–Host Database (85). Thirty-six EVGs (i.e., 1 OBV, 32 TOVs, and 3 others) were most similar to eukaryotic viral genomes among RVGs and were removed from the proteomic tree and gOTU analyses, which were used to compare the levels of diversity of the RVGs and EVGs of prokaryotic viruses.

**Proteomic tree.** We constructed a proteomic tree as previously described (102). Briefly, the all-against-all distance matrix of the EVG/RVG data set was calculated on the basis of the normalized bit score of tBLASTx ($S_G$), and the proteomic tree was built with BIONJ using the distance matrix. The proteomic tree, gene annotations, and genome alignment views are accessible at http://www.genome.ad.jp/viptree/EVG2017.

**Genus-level operational taxonomic units.** The genus-level threshold value for gOTU clustering was estimated from a subset of the RVGs used in this study (i.e., 345 prokaryotic dsDNA viruses), each of which was assigned to a viral genus (i.e., 82 genera in total). We constructed gOTUs with different $S_G$ cutoffs (intervals of 0.01) and evaluated how closely the resulting gOTUs corresponded to the genus-level viral classifications using the adjusted Rand index (103).

**Host predictions according to proteomic similarities.** We attempted to predict host taxonomic groups for EVGs on the basis of viral genomic similarities measured with $S_G$. We estimated the precision of our prediction method on the basis of RVGs (i.e., 1,285 prokaryotic dsDNA viruses), each of which was linked to a uniquely assigned host taxonomic group according to the Virus-Host Database. Regarding host taxonomic groups, *Cyanobacteria* (phylum) and *Enterobacteriaceae* (family) were regarded as individual host taxonomic groups because closely related viruses are known to infect hosts of different genera belonging to these host groups. The remaining viral hosts were grouped at the genus level. For each RVG, the best $S_G$ values for the members of the same host group, and for the members outside the host group, were recorded (i.e., 2,570 $S_G$ scores in total). A precision curve was generated using sliding $S_G$ cutoff values (Fig. S6). When the $S_G$ cutoff value was >0.3889 or >0.2937, the viral pairs were predicted to infect hosts in the same group at >95% or >90% precision, respectively.

**Photosynthetic gene identification.** To detect photosynthetic genes in the EVG/RVG data set, we used PSI-BLAST (E value, 1e−6; inclusion_ethresh, 1e−6; num_iterations, 3) and the query sequences listed in Table S1I.

**Phylogenetic trees.** Multiple sequences were aligned using the MAFFT program (version 7.245) (104), with the FFT-NS-2 mode and a maximum of 1,000 iterations (--retree 2, --maxiterate 1000). Conserved positions in the alignments were selected with the trimAl program (version 1.3) (105). Maximum likelihood trees with 100 bootstrap replicates were calculated with RAxML (version 8.2.4) (106) using the fast bootstrapping mode, and models were selected by the use of ProteinModelSelection.pl (i.e., LGF for DNA polymerase B and LG for chaperonins, IscU, and ATC).

**Recruitment of Pacific Ocean virome sequences.** Reads (3.68 M sequences) and proteins (2.78 M sequences) of 16 photic POV samples were downloaded from iMicrobe (http://data.imicrobe.us). These sequences were mapped on EVGs and RVGs using BLASTn (for reads; E value, $<$1e$-$3) and tBLASTn (for proteins; E value, $<$1e$-$3) if the alignment revealed $\geq$60% identity and covered $\geq$80% of the query sequence.

**Accession number(s).** Read and assembled sequences obtained from OBV were deposited at DNA Data Bank of Japan (DDBJ) under accession numbers DRR053207 to DRR053215 and SAMD00045684 to SAMD00045692. The sequence data for the OBV project are accessible under DDBJ BioProject accession number PRJDB4437. Sequences and additional data are available at ftp://ftp.genome.jp/pub/db/community/EVG2017.

## SUPPLEMENTAL MATERIAL

Supplemental material for this article may be found at https://doi.org/10.1128/mSphere.00359-16.

**TEXT S1,** DOCX file, 0.05 MB.
**FIG S1,** PDF file, 1.7 MB.
**FIG S2,** PDF file, 1.5 MB.
**FIG S3,** PDF file, 0.1 MB.
**FIG S4,** PDF file, 0.3 MB.
**FIG S5,** PDF file, 0.1 MB.
**FIG S6,** PDF file, 0.1 MB.
**FIG S7,** PDF file, 0.3 MB.
**FIG S8,** PDF file, 0.3 MB.
**TABLE S1,** PDF file, 1.2 MB.

## ACKNOWLEDGMENTS

We thank the *Tara* Oceans consortium, people, and sponsors who supported the *Tara* Oceans expedition (http://www.embl.de/tara-oceans/) for making the data accessible. Computational work was completed at the Supercomputer System, Institute for Chemical Research, Kyoto University.

This work was supported by the Canon Foundation (no. 203143100025), JSPS/KAKENHI (no. 26430184 and 16KT0020), Scientific Research on Innovative Areas from the Ministry of Education, Culture, Science, Sports and Technology (MEXT) of Japan (no. 16H06429, 16K21723, and 16H06437), and the Collaborative Research Program of the Institute for Chemical Research, Kyoto University (no. 2016-28). P.H. was supported by the OCEANOMICS "Investissements d'Avenir" program of the French Government (no. ANR-11-BTBR-0008). M.B.S. was supported by Gordon and Betty Moore Foundation grants (no. 3790 and GBMF2631), and S.R. was partially supported by the University of Arizona Technology and Research Initiative Fund through a grant from the Water, Environmental, and Energy Solutions Initiative and the Ecosystem Genomics Institute to M.B.S.

This is contribution number 51 of the *Tara* Oceans Expedition 2009–2012.

## REFERENCES

1. Bergh O, Børsheim KY, Bratbak G, Heldal M. 1989. High abundance of viruses found in aquatic environments. Nature 340:467–468. https://doi.org/10.1038/340467a0.
2. Falkowski PG, Fenchel T, Delong EF. 2008. The microbial engines that drive Earth's biogeochemical cycles. Science 320:1034–1039. https://doi.org/10.1126/science.1153213.
3. Proctor LM, Fuhrman JA. 1990. Viral mortality of marine bacteria and cyanobacteria. Nature 343:60–62. https://doi.org/10.1038/343060a0.
4. Mann NH, Cook A, Millard A, Bailey S, Clokie M. 2003. Marine ecosystems: bacterial photosynthesis genes in a virus. Nature 424:741. https://doi.org/10.1038/424741a.
5. Brüssow H, Canchaya C, Hardt WD. 2004. Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion. Microbiol Mol Biol Rev 68:560–602. https://doi.org/10.1128/MMBR.68.3.560-602.2004.
6. Sullivan MB, Lindell D, Lee JA, Thompson LR, Bielawski JP, Chisholm SW. 2006. Prevalence and evolution of core photosystem II genes in marine cyanobacterial viruses and their hosts. PLoS Biol 4:e234. https://doi.org/10.1371/journal.pbio.0040234.
7. Hatfull GF. 2008. Bacteriophage genomics. Curr Opin Microbiol 11:447–453. https://doi.org/10.1016/j.mib.2008.09.004.
8. Rohwer F. 2003. Global phage diversity. Cell 113:141. https://doi.org/10.1016/S0092-8674(03)00276-9.
9. Brum JR, Ignacio-Espinoza JC, Roux S, Doulcier G, Acinas SG, Alberti A, Chaffron S, Cruaud C, de Vargas C, Gasol JM, Gorsky G, Gregory AC, Guidi L, Hingamp P, Iudicone D, Not F, Ogata H, Pesant S, Poulos BT,

Schwenck SM, Speich S, Dimier C, Kandels-Lewis S, Picheral M, Searson S; Tara Oceans Coordinators, Bork P, Bowler C, Sunagawa S, Wincker P, Karsenti E, Sullivan MB. 2015. Ocean plankton. Patterns and ecological drivers of ocean viral communities. Science 348:1261498. https://doi.org/10.1126/science.1261498.
10. Zhao Y, Temperton B, Thrash JC, Schwalbach MS, Vergin KL, Landry ZC, Ellisman M, Deerinck T, Sullivan MB, Giovannoni SJ. 2013. Abundant SAR11 viruses in the ocean. Nature 494:357–360. https://doi.org/10.1038/nature11921.
11. Kang I, Oh HM, Kang D, Cho JC. 2013. Genome of a SAR116 bacteriophage shows the prevalence of this phage type in the oceans. Proc Natl Acad Sci U S A 110:12343–12348. https://doi.org/10.1073/pnas.1219930110.
12. Roux S, Hallam SJ, Woyke T, Sullivan MB. 2015. Viral dark matter and virus-host interactions resolved from publicly available microbial genomes. Elife 4:e08490. https://doi.org/10.7554/eLife.08490.
13. Hingamp P, Grimsley N, Acinas SG, Clerissi C, Subirana L, Poulain J, Ferrera I, Sarmento H, Villar E, Lima-Mendez G, Faust K, Sunagawa S, Claverie JM, Moreau H, Desdevises Y, Bork P, Raes J, de Vargas C, Karsenti E, Kandels-Lewis S, Jaillon O, Not F, Pesant S, Wincker P, Ogata H. 2013. Exploring nucleo-cytoplasmic large DNA viruses in Tara oceans microbial metagenomes. ISME J 7:1678–1695. https://doi.org/10.1038/ismej.2013.59.
14. Hurwitz BL, Hallam SJ, Sullivan MB. 2013. Metabolic reprogramming by viruses in the sunlit and dark ocean. Genome Biol 14:R123. https://doi.org/10.1186/gb-2013-14-11-r123.
15. Hurwitz BL, Brum JR, Sullivan MB. 2015. Depth-stratified functional and taxonomic niche specialization in the 'core' and "flexible" Pacific Ocean virome. ISME J 9:472–484. https://doi.org/10.1038/ismej.2014.143.
16. Roux S, Brum JR, Dutilh BE, Sunagawa S, Duhaime MB, Loy A, Poulos BT, Solonenko N, Lara E, Poulain J, Pesant S, Kandels-Lewis S, Dimier C, Picheral M, Searson S, Cruaud C, Alberti A, Duarte CM, Gasol JM, Vaqué D; Tara Oceans Coordinators, Bork P, Acinas SG, Wincker P, Sullivan MB. 2016. Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. Nature 537:689–693. https://doi.org/10.1038/nature19366.
17. Culley AI, Lang AS, Suttle CA. 2006. Metagenomic analysis of coastal RNA virus communities. Science 312:1795–1798. https://doi.org/10.1126/science.1127404.
18. Culley AI, Mueller JA, Belcaid M, Wood-Charlson EM, Poisson G, Steward GF. 2014. The characterization of RNA viruses in tropical seawater using targeted PCR and metagenomics. mBio 5:e01210-14. https://doi.org/10.1128/mBio.01210-14.
19. López-Bueno A, Tamames J, Velázquez D, Moya A, Quesada A, Alcamí A. 2009. High diversity of the viral community from an Antarctic lake. Science 326:858–861. https://doi.org/10.1126/science.1179287.
20. Rosario K, Duffy S, Breitbart M. 2009. Diverse circovirus-like genome architectures revealed by environmental metagenomics. J Gen Virol 90:2418–2424. https://doi.org/10.1099/vir.0.012955-0.
21. 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. https://doi.org/10.1038/ismej.2010.188.
22. Diemer GS, Stedman KM. 2012. A novel virus genome discovered in an extreme environment suggests recombination between unrelated groups of RNA and DNA viruses. Biol Direct 7:13. https://doi.org/10.1186/1745-6150-7-13.
23. Roux S, Krupovic M, Poulet A, Debroas D, Enault F. 2012. Evolution and diversity of the Microviridae viral family through a collection of 81 new complete genomes assembled from virome reads. PLoS One 7:e40418. https://doi.org/10.1371/journal.pone.0040418.
24. Labonté JM, Suttle CA. 2013. Previously unknown and highly divergent ssDNA viruses populate the oceans. ISME J 7:2169–2177. https://doi.org/10.1038/ismej.2013.110.
25. McDaniel LD, Rosario K, Breitbart M, Paul JH. 2014. Comparative metagenomics: natural populations of induced prophages demonstrate highly unique, lower diversity viral sequences. Environ Microbiol 16:570–585. https://doi.org/10.1111/1462-2920.12184.
26. Zawar-Reza P, Argüello-Astorga GR, Kraberger S, Julian L, Stainton D, Broady PA, Varsani A. 2014. Diverse small circular single-stranded DNA viruses identified in a freshwater pond on the McMurdo Ice Shelf (Antarctica). Infect Genet Evol 26:132–138. https://doi.org/10.1016/j.meegid.2014.05.018.
27. Emerson JB, Thomas BC, Andrade K, Allen EE, Heidelberg KB, Banfield JF. 2012. Dynamic viral populations in hypersaline systems as revealed by metagenomic assembly. Appl Environ Microbiol 78:6309–6320. https://doi.org/10.1128/AEM.01212-12.
28. Anantharaman K, Duhaime MB, Breier JA, Wendt KA, Toner BM, Dick GJ. 2014. Sulfur oxidation genes in diverse deep-sea viruses. Science 344:757–760. https://doi.org/10.1126/science.1252229.
29. Bellas CM, Anesio AM, Barker G. 2015. Analysis of virus genomes from glacial environments reveals novel virus groups with unusual host interactions. Front Microbiol 6:656. https://doi.org/10.3389/fmicb.2015.00656.
30. Duhaime MB, Sullivan MB. 2012. Ocean viruses: rigorously evaluating the metagenomic sample-to-sequence pipeline. Virology 434:181–186. https://doi.org/10.1016/j.virol.2012.09.036.
31. Santos F, Meyerdierks A, Peña A, Rosselló-Mora R, Amann R, Antón J. 2007. Metagenomic approach to the study of halophages: the environmental halophage 1. Environ Microbiol 9:1711–1723. https://doi.org/10.1111/j.1462-2920.2007.01289.x.
32. Garcia-Heredia I, Martin-Cuadrado AB, Mojica FJ, Santos F, Mira A, Antón J, Rodriguez-Valera F. 2012. Reconstructing viral genomes from the environment using fosmid clones: the case of haloviruses. PLoS One 7:e33802. https://doi.org/10.1371/journal.pone.0033802.
33. Mizuno CM, Rodriguez-Valera F, Kimes NE, Ghai R. 2013. Expanding the marine virosphere using metagenomics. PLoS Genet 9:e1003987. https://doi.org/10.1371/journal.pgen.1003987.
34. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 19:455–477. https://doi.org/10.1089/cmb.2012.0021.
35. Peng Y, Leung HCM, Yiu SM, Chin FYL. 2012. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics 28:1420–1428. https://doi.org/10.1093/bioinformatics/bts174.
36. Boisvert S, Raymond F, Godzaridis E, Laviolette F, Corbeil J. 2012. Ray Meta: scalable de novo metagenome assembly and profiling. Genome Biol 13:R122. https://doi.org/10.1186/gb-2012-13-12-r122.
37. Hunt M, Kikuchi T, Sanders M, Newbold C, Berriman M, Otto TD. 2013. REAPR: a universal tool for genome assembly evaluation. Genome Biol 14:R47. https://doi.org/10.1186/gb-2013-14-5-r47.
38. Bao Y, Chetvernin V, Tatusova T. 2014. Improvements to pairwise sequence comparison (PASC): a genome-based web tool for virus classification. Arch Virol 159:3293–3304. https://doi.org/10.1007/s00705-014-2197-x.
39. Labonté JM, Swan BK, Poulos B, Luo H, Koren S, Hallam SJ, Sullivan MB, Woyke T, Wommack KE, Stepanauskas R. 2015. Single-cell genomics-based analysis of virus-host interactions in marine surface bacterioplankton. ISME J 9:2386–2399. https://doi.org/10.1038/ismej.2015.48.
40. Rohwer F, Edwards R. 2002. The phage proteomic tree: a genome-based taxonomy for phage. J Bacteriol 184:4529–4535. https://doi.org/10.1128/JB.184.16.4529-4535.2002.
41. Paez-Espino D, Eloe-Fadrosh EA, Pavlopoulos GA, Thomas AD, Huntemann M, Mikhailova N, Rubin E, Ivanova NN, Kyrpides NC. 2016. Uncovering Earth's virome. Nature 536:425–430. https://doi.org/10.1038/nature19094.
42. Ackermann HW, Prangishvili D. 2012. Prokaryote viruses studied by electron microscopy. Arch Virol 157:1843–1849. https://doi.org/10.1007/s00705-012-1383-y.
43. Krupovic M, Spang A, Gribaldo S, Forterre P, Schleper C. 2011. A thaumarchaeal provirus testifies for an ancient association of tailed viruses with archaea. Biochem Soc Trans 39:82–88. https://doi.org/10.1042/BST0390082.
44. Hildenbrand ZL, Bernal RA. 2012. Chaperonin-mediated folding of viral proteins. Adv Exp Med Biol 726:307–324. https://doi.org/10.1007/978-1-4614-0980-9_13.
45. Iverson V, Morris RM, Frazar CD, Berthiaume CT, Morales RL, Armbrust EV. 2012. Untangling genomes from metagenomes: revealing an uncultured class of marine Euryarchaeota. Science 335:587–590. https://doi.org/10.1126/science.1212665.
46. Martin-Cuadrado AB, Garcia-Heredia I, Moltó AG, López-Úbeda R, Kimes N, López-García P, Moreira D, Rodriguez-Valera F. 2015. A new class of marine Euryarchaeota group II from the Mediterranean deep chlorophyll maximum. ISME J 9:1619–1634. https://doi.org/10.1038/ismej.2014.249.
47. Fuhrman JA, Davis AA. 1997. Widespread Archaea and novel Bacteria

from the deep sea as shown by 16S rRNA gene sequences. Mar Ecol Prog Ser 150:275–285. https://doi.org/10.3354/meps150275.
48. Könneke M, Bernhard AE, de la Torre JR, Walker CB, Waterbury JB, Stahl DA. 2005. Isolation of an autotrophic ammonia-oxidizing marine archaeon. Nature 437:543–546. https://doi.org/10.1038/nature03911.
49. Massana R, DeLong EF, Pedrós-Alió C. 2000. A few cosmopolitan phylotypes dominate planktonic archaeal assemblages in widely different oceanic provinces. Appl Environ Microbiol 66:1777–1787. https://doi.org/10.1128/AEM.66.5.1777-1787.2000.
50. DeLong EF, Preston CM, Mincer T, Rich V, Hallam SJ, Frigaard NU, Martinez A, Sullivan MB, Edwards R, Brito BR, Chisholm SW, Karl DM. 2006. Community genomics among stratified microbial assemblages in the ocean's interior. Science 311:496–503. https://doi.org/10.1126/science.1120250.
51. Needham DM, Fuhrman JA. 2016. Pronounced daily succession of phytoplankton, archaea and bacteria following a spring bloom. Nat Microbiol 1:16005. https://doi.org/10.1038/nmicrobiol.2016.5.
52. Barras F, Loiseau L, Py B. 2005. How Escherichia coli and Saccharomyces cerevisiae build Fe/S proteins. Adv Microb Physiol 50:41–101. https://doi.org/10.1016/S0065-2911(05)50002-X.
53. Sharon I, Battchikova N, Aro EM, Giglione C, Meinnel T, Glaser F, Pinter RY, Breitbart M, Rohwer F, Béjà O. 2011. Comparative metagenomics of microbial traits within oceanic viral communities. ISME J 5:1178–1190. https://doi.org/10.1038/ismej.2011.2.
54. Vinella D, Brochier-Armanet C, Loiseau L, Talla E, Barras F. 2009. Iron-sulfur (Fe/S) protein biogenesis: phylogenomic and genetic studies of A-type carriers. PLoS Genet 5:e1000497. https://doi.org/10.1371/journal.pgen.1000497.
55. Lill R, Dutkiewicz R, Elsässer HP, Hausmann A, Netz DJ, Pierik AJ, Stehling O, Urzica E, Mühlenhoff U. 2006. Mechanisms of iron-sulfur protein maturation in mitochondria, cytosol and nucleus of eukaryotes. Biochim Biophys Acta 1763:652–667. https://doi.org/10.1016/j.bbamcr.2006.05.011.
56. Shepard EM, Boyd ES, Broderick JB, Peters JW. 2011. Biosynthesis of complex iron-sulfur enzymes. Curr Opin Chem Biol 15:319–327. https://doi.org/10.1016/j.cbpa.2011.02.012.
57. Dupont CL, Rusch DB, Yooseph S, Lombardo MJ, Richter RA, Valas R, Novotny M, Yee-Greenbaum J, Selengut JD, Haft DH, Halpern AL, Lasken RS, Nealson K, Friedman R, Venter JC. 2012. Genomic insights to SAR86, an abundant and uncultivated marine bacterial lineage. ISME J 6:1186–1199. https://doi.org/10.1038/ismej.2011.189.
58. Grell TA, Goldman PJ, Drennan CL. 2015. SPASM and twitch domains in S-adenosylmethionine (SAM) radical enzymes. J Biol Chem 290:3964–3971. https://doi.org/10.1074/jbc.R114.581249.
59. White MF, Dillingham MS. 2012. Iron-sulphur clusters in nucleic acid processing enzymes. Curr Opin Struct Biol 22:94–100. https://doi.org/10.1016/j.sbi.2011.11.004.
60. Hooton SP, Connerton IF. 2014. Campylobacter jejuni acquire new host-derived CRISPR spacers when in association with bacteriophages harboring a CRISPR-like Cas4 protein. Front Microbiol 5:744. https://doi.org/10.3389/fmicb.2014.00744.
61. Roche B, Aussel L, Ezraty B, Mandin P, Py B, Barras F. 2013. Iron/sulfur proteins biogenesis in prokaryotes: formation, regulation and diversity. Biochim Biophys Acta 1827:455–469. https://doi.org/10.1016/j.bbabio.2012.12.010.
62. Fernández C, Ferrández A, Miñambres B, Díaz E, García JL. 2006. Genetic characterization of the phenylacetyl-coenzyme A oxygenase from the aerobic phenylacetic acid degradation pathway of Escherichia coli. Appl Environ Microbiol 72:7422–7426. https://doi.org/10.1128/AEM.01550-06.
63. Chénard C, Chan AM, Vincent WF, Suttle CA. 2015. Polar freshwater cyanophage S-EIV1 represents a new widespread evolutionary lineage of phages. ISME J 9:2046–2058. https://doi.org/10.1038/ismej.2015.24.
64. Hahnke RL, Bennke CM, Fuchs BM, Mann AJ, Rhiel E, Teeling H, Amann R, Harder J. 2015. Dilution cultivation of marine heterotrophic bacteria abundant after a spring phytoplankton bloom in the North Sea. Environ Microbiol 17:3515–3526. https://doi.org/10.1111/1462-2920.12479.
65. Teeling H, Fuchs BM, Becher D, Klockow C, Gardebrecht A, Bennke CM, Kassabgy M, Huang S, Mann AJ, Waldmann J, Weber M, Klindworth A, Otto A, Lange J, Bernhardt J, Reinsch C, Hecker M, Peplies J, Bockelmann FD, Callies U, Gerdts G, Wichels A, Wiltshire KH, Glöckner FO, Schweder T, Amann R. 2012. Substrate-controlled succession of marine bacterioplankton populations induced by a phytoplankton bloom. Science 336:608–611. https://doi.org/10.1126/science.1218344.
66. Borriss M, Lombardot T, Glöckner FO, Becher D, Albrecht D, Schweder T. 2007. Genome and proteome characterization of the psychrophilic Flavobacterium bacteriophage 11b. Extremophiles 11:95–104. https://doi.org/10.1007/s00792-006-0014-5.
67. Kang I, Kang D, Cho JC. 2012. Complete genome sequence of Croceibacter bacteriophage P2559S. J Virol 86:8912–8913. https://doi.org/10.1128/JVI.01396-12.
68. Kang I, Jang H, Cho JC. 2012. Complete genome sequences of two Persicivirga bacteriophages, P12024S and P12024L. J Virol 86:8907–8908. https://doi.org/10.1128/JVI.01327-12.
69. Holmfeldt K, Solonenko N, Shah M, Corrier K, Riemann L, Verberkmoes NC, Sullivan MB. 2013. Twelve previously unknown phage genera are ubiquitous in global oceans. Proc Natl Acad Sci U S A 110:12798–12803. https://doi.org/10.1073/pnas.1305956110.
70. Senčilo A, Luhtanen AM, Saarijärvi M, Bamford DH, Roine E. 2015. Cold-active bacteriophages from the Baltic Sea ice have diverse genomes and virus-host interactions. Environ Microbiol 17:3628–3641. https://doi.org/10.1111/1462-2920.12611.
71. Kang I, Jang H, Cho JC. 2015. Complete genome sequences of bacteriophages P12002L and P12002S, two lytic phages that infect a marine Polaribacter strain. Stand Genomic Sci 10:82. https://doi.org/10.1186/s40793-015-0076-z.
72. Pinhassi J, Bowman JP, Nedashkovskaya OI, Lekunberri I, Gomez-Consarnau L, Pedrós-Alió C. 2006. Leeuwenhoekiella blandensis sp. nov., a genome-sequenced marine member of the family Flavobacteriaceae. Int J Syst Evol Microbiol 56:1489–1493. https://doi.org/10.1099/ijs.0.64232-0.
73. Gómez-Consarnau L, González JM, Coll-Lladó M, Gourdon P, Pascher T, Neutze R, Pedrós-Alió C, Pinhassi J. 2007. Light stimulates growth of proteorhodopsin-containing marine Flavobacteria. Nature 445:210–213. https://doi.org/10.1038/nature05381.
74. Cozzone AJ. 1998. Regulation of acetate metabolism by protein phosphorylation in enteric bacteria. Annu Rev Microbiol 52:127–164. https://doi.org/10.1146/annurev.micro.52.1.127.
75. Dunn MF, Ramírez-Trujillo JA, Hernández-Lucas I. 2009. Major roles of isocitrate lyase and malate synthase in bacterial and fungal pathogenesis. Microbiology 155:3166–3175. https://doi.org/10.1099/mic.0.030858-0.
76. Vimr ER, Steenbergen SM. 2009. Early molecular-recognition events in the synthesis and export of group 2 capsular polysaccharides. Microbiology 155:9–15. https://doi.org/10.1099/mic.0.023564-0.
77. Smyth KM, Marchant A. 2013. Conservation of the 2-keto-3-deoxymanno-octulosonic acid (Kdo) biosynthesis pathway between plants and bacteria. Carbohydr Res 380:70–75. https://doi.org/10.1016/j.carres.2013.07.006.
78. Williamson SJ, McLaughlin MR, Paul JH. 2001. Interaction of the PhiHSIC virus with its host: lysogeny or pseudolysogeny? Appl Environ Microbiol 67:1682–1688. https://doi.org/10.1128/AEM.67.4.1682-1688.2001.
79. Sullivan MB, Coleman ML, Weigele P, Rohwer F, Chisholm SW. 2005. Three Prochlorococcus cyanophage genomes: signature features and ecological interpretations. PLoS Biol 3:e144. https://doi.org/10.1371/journal.pbio.0030144.
80. Dziewit L, Jazurek M, Drewniak L, Baj J, Bartosik D. 2007. The SXT conjugative element and linear prophage N15 encode toxin-antitoxin-stabilizing systems homologous to the tad-ata module of the Paracoccus aminophilus plasmid pAMI2. J Bacteriol 189:1983–1997. https://doi.org/10.1128/JB.01610-06.
81. Thomsen LE, Chadfield MS, Bispham J, Wallis TS, Olsen JE, Ingmer H. 2003. Reduced amounts of LPS affect both stress tolerance and virulence of Salmonella enterica serovar Dublin. FEMS Microbiol Lett 228:225–231. https://doi.org/10.1016/S0378-1097(03)00762-6.
82. Oh HM, Kwon KK, Kang I, Kang SG, Lee JH, Kim SJ, Cho JC. 2010. Complete genome sequence of "Candidatus Puniceispirillum marinum" IMCC1322, a representative of the SAR116 clade in the Alphaproteobacteria. J Bacteriol 192:3240–3241. https://doi.org/10.1128/JB.00347-10.
83. Grote J, Thrash JC, Huggett MJ, Landry ZC, Carini P, Giovannoni SJ, Rappé MS. 2012. Streamlining and core genome conservation among highly divergent members of the SAR11 clade. mBio 3:e00252-12. https://doi.org/10.1128/mBio.00252-12.
84. Cardinale DJ, Duffy S. 2011. Single-stranded genomic architecture constrains optimal codon usage. Bacteriophage 1:219–224. https://doi.org/10.4161/bact.1.4.18496.
85. Mihara T, Nishimura Y, Shimizu Y, Nishiyama H, Yoshikawa G, Uehara H,

Hingamp P, Goto S, Ogata H. 2016. Linking virus genomes with host taxonomy. Viruses 8:66. https://doi.org/10.3390/v8030066.
86. 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:e57355. https://doi.org/10.1371/journal.pone.0057355.
87. Breitbart M, Thompson LR, Suttle CA, Sullivan MB. 2007. Exploring the vast diversity of marine viruses. Oceanography 20:135–139. https://doi.org/10.5670/oceanog.2007.58.
88. Thompson LR, Zeng Q, Kelly L, Huang KH, Singer AU, Stubbe J, Chisholm SW. 2011. Phage auxiliary metabolic genes and the redirection of cyanobacterial host carbon metabolism. Proc Natl Acad Sci U S A 108:E757–E764. https://doi.org/10.1073/pnas.1102164108.
89. John SG, Mendez CB, Deng L, Poulos B, Kauffman AK, Kern S, Brum J, Polz MF, Boyle EA, Sullivan MB. 2011. A simple and efficient method for concentration of ocean viruses by chemical flocculation. Environ Microbiol Rep 3:195–202. https://doi.org/10.1111/j.1758-2229.2010.00208.x.
90. 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 15:1428–1440. https://doi.org/10.1111/j.1462-2920.2012.02836.x.
91. Kimura S, Yoshida T, Hosoda N, Honda T, Kuno S, Kamiji R, Hashimoto R, Sako Y. 2012. Diurnal infection patterns and impact of Microcystis cyanophages in a Japanese pond. Appl Environ Microbiol 78:5805–5811. https://doi.org/10.1128/AEM.00571-12.
92. Nikolenko SI, Korobeynikov AI, Alekseyev MA. 2013. BayesHammer: Bayesian clustering for error correction in single-cell sequencing. BMC Genomics 14(Suppl 1):S7. https://doi.org/10.1186/1471-2164-14-S1-S7.
93. Magoč T, Salzberg SL. 2011. FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics 27:2957–2963. https://doi.org/10.1093/bioinformatics/btr507.
94. Zhu W, Lomsadze A, Borodovsky M. 2010. Ab initio gene identification in metagenomic sequences. Nucleic Acids Res 38:e132. https://doi.org/10.1093/nar/gkq275.
95. Söding J. 2005. Protein homology detection by HMM-HMM comparison. Bioinformatics 21:951–960. https://doi.org/10.1093/bioinformatics/bti125.
96. Eddy SR. 2011. Accelerated profile HMM searches. PLoS Comput Biol 7:e1002195. https://doi.org/10.1371/journal.pcbi.1002195.
97. Roux S, Enault F, Hurwitz BL, Sullivan MB. 2015. VirSorter: mining viral signal from microbial genomic data. PeerJ 3:e985. https://doi.org/10.7717/peerj.985.
98. Wang Y, Tang H, Debarry JD, Tan X, Li J, Wang X, Lee TH, Jin H, Marler B, Guo H, Kissinger JC, Paterson AH. 2012. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res 40:e49. https://doi.org/10.1093/nar/gkr1293.
99. Xu H, Luo X, Qian J, Pang X, Song J, Qian G, Chen J, Chen S. 2012. FastUniq: a fast de novo duplicates removal tool for paired short reads. PLoS One 7:e52249. https://doi.org/10.1371/journal.pone.0052249.
100. Nei M, Li WH. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc Natl Acad Sci U S A 76:5269–5273. https://doi.org/10.1073/pnas.76.10.5269.
101. Shao W, Kearney MF, Boltz VF, Spindler JE, Mellors JW, Maldarelli F, Coffin JM. 2014. PAPNC, a novel method to calculate nucleotide diversity from large scale next generation sequencing data. J Virol Methods 203:73–80. https://doi.org/10.1016/j.jviromet.2014.03.008.
102. Bhunchoth A, Blanc-Mathieu R, Mihara T, Nishimura Y, Askora A, Phironrit N, Leksomboon C, Chatchawankanphanich O, Kawasaki T, Nakano M, Fujie M, Ogata H, Yamada T. 2016. Two Asian jumbo phages, $\phi$RSL2 and $\phi$RSF1, infect Ralstonia solanacearum and show common features of $\phi$KZ-related phages. Virology 494:56–66. https://doi.org/10.1016/j.virol.2016.03.028.
103. Hubert L, Arabie P. 1985. Comparing partitions. J Classif 2:193–218. https://doi.org/10.1007/BF01908075.
104. Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30:772–780. https://doi.org/10.1093/molbev/mst010.
105. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25:1972–1973. https://doi.org/10.1093/bioinformatics/btp348.
106. Stamatakis A, Hoover P, Rougemont J. 2008. A rapid bootstrap algorithm for the RAxML Web servers. Syst Biol 57:758–771. https://doi.org/10.1080/10635150802429642.
107. Chao A, Gotelli NJ, Hsieh TC, Sander EL, Ma KH, Colwell RK, Ellison AM. 2014. Rarefaction and extrapolation with Hill numbers: a framework for sampling and estimation in species diversity studies. Ecol Monogr 84:45–67. https://doi.org/10.1890/13-0133.1.