Genomic diversity and biogeographic distributions of a novel lineage of bacteriophages that infect marine OM43 bacteria

Mingyu Yang,[1] Sen Du,[1] Zefeng Zhang,[1] Qian Xia,[1] He Liu,[1] Fang Qin,[1] Zuqing Wu,[1] Hanqi Ying,[1] Yin Wu,[1] Jiabing Shao,[1] Yanlin Zhao[1,2]

AUTHOR AFFILIATIONS See affiliation list on p. 15.

ABSTRACT The marine methylotrophic OM43 clade is considered an important bacterial group in coastal microbial communities. OM43 bacteria, which are closely related to phytoplankton blooms, have small cell sizes and streamlined genomes. Bacteriophages profoundly shape the evolutionary trajectories, population dynamics, and physiology of microbes. The prevalence and diversity of several phages that infect OM43 bacteria have been reported. In this study, we isolated and sequenced two novel OM43 phages, MEP401 and MEP402. These phages share 90% of their open reading frames (ORFs) and are distinct from other known phage isolates. Furthermore, a total of 99 metagenomic viral genomes (MVGs) closely related to MEP401 and MEP402 were identified. Phylogenomic analyses suggest that MEP401, MEP402, and these identified MVGs belong to a novel subfamily in the family *Zobellviridae* and that they can be separated into two groups. Group I MVGs show conserved whole-genome synteny with MEP401, while group II MVGs possess the MEP401-type DNA replication module and a distinct type of morphogenesis and packaging module, suggesting that genomic recombination occurred between phages. Most members in these two groups were predicted to infect OM43 bacteria. Metagenomic read-mapping analysis revealed that the phages in these two groups are globally ubiquitous and display distinct biogeographic distributions, with some phages being predominant in cold regions, some exclusively detected in estuarine stations, and others displaying wider distributions. This study expands our knowledge of the diversity and ecology of a novel phage lineage that infects OM43 bacteria by describing their genomic diversity and global distribution patterns.

IMPORTANCE OM43 phages that infect marine OM43 bacteria are important for host mortality, community structure, and physiological functions. In this study, two OM43 phages were isolated and characterized. Metagenomic viral genome (MVG) retrieval using these two OM43 phages as baits led to the identification of two phage groups of a new subfamily in the family *Zobellviridae*. We found that group I MVGs share similar genomic content and arrangement with MEP401 and MEP402, whereas group II MVGs only possess the MEP401-type DNA replication module. Metagenomic mapping analysis suggests that members in these two groups are globally ubiquitous with distinct distribution patterns. This study provides important insights into the genomic diversity and biogeography of the OM43 phages in the global ocean.

Viruses are widely recognized as the most abundant biological entities in the ocean and play pivotal roles in the marine microbial loop and biogeochemical cycles. It has been estimated that viral shunts produce about 10 billion tons of carbon per day by releasing organic matter, which is considered fundamental to nutrient cycling,

promoting productivity of the oceans (1–3). As the major top-down forces of bacteria, bacteriophages outnumber bacteria by an order of magnitude (4, 5). Phages constitute the vast majority of marine viral communities (4–6) and influence nutrient cycling, host community structure, and evolution.

Marine viral community structure and diversity have been uncovered using culture-independent viral diversity surveys, such as metagenomics (7–12) and single-cell genomics (13–15). In the recent decade, enormous amounts of metagenomic data have been generated to reveal a previously unrecognized viral diversity in various marine environments (7–12). Meanwhile, a substantial number of novel viral genomic fragments have been assembled from marine metagenomic and viromic reads, providing valuable viral reference genomes without isolation efforts (8–12, 16, 17). In comparison, culture-dependent viral isolation lags far behind metagenomic studies because of the lack of laboratory marine bacterial cultures and difficulties in isolating viruses. Recently, considerable efforts have been made to isolate bacteriophages that infect ecologically important and abundant marine bacterioplankton. For example, SAR11 phages (pelagiphages), *Roseobacter* RCA phages, and SAR116 phages have been isolated and shown to have significant genomic diversity and to dominate the marine viral communities (18–24). Isolation of these phages has greatly facilitated the annotation and host prediction of many related metagenomic viral sequences. In addition, phage isolation not only provides infectious data but also provides experimental model systems for the investigation of phage-host interactions and phage ecological impact. Phages have pervasively mosaic genomes, resulting from complex evolutionary processes over a long period of time (25, 26). Lateral gene transfer via genetic recombination and fast mutations are the major forces that promote the genomic evolution of phages (27–29). Analysis of the phage genomes can help to understand phage evolution.

As an important group of methylotrophs, the OM43 clade in the family *Methylophilaceae* of *Betaproteobacteria* occurs commonly in marine environments and plays an important role in C1 metabolism and carbon cycling (30–37). The OM43 clade was first discovered by sequencing ribosomal RNA genes from the western Atlantic coast (30), and it was speculated to be linked to phytoplankton populations and primary productivity (33). The OM43 clade, which has coincided with phytoplankton blooms in surface water (36), can account for up to 5% of the microbial community in coastal ecosystems (30). Members of the OM43 clade are small and genomically streamlined (~1.3 Mbp) (34, 35, 37). To date, there are two known OM43 subgroups (37), the H-RS cluster and the HTCC2181 cluster. OM43 bacteria are difficult to culture in laboratories owing to their sensitivity to slight biochemical variations in seawater (34).

To date, only three OM43 phage isolates have been reported, phage Venkman (EXVC282S), isolated from the western English Channel on OM43 strain H5P1 (23), phage MEP301, isolated from the Bohai Sea on OM43 strain HTCC2181 (38), and phage Melnitz (EXVC044M), isolated from the western English Channel and Sargasso Sea on OM43 strain H5P1 (39). These three OM43 phages were classified into three distinct viral groups. Venkman and MEP301 both have an icosahedral head and a short tail, belonging to two distinct viral groups (23, 38). Melnitz has a myovirus morphotype and is most closely related to pelagiphage Mosig and HTVC008M (39). Metagenomic analysis showed that most OM43 phages were correlated with temperature, having a higher relative abundance in cold waters (38, 39).

To broaden the understanding of marine OM43 phages, a novel OM43 strain, FZCC0133, was used as a host to isolate OM43 phages. FZCC0133 belongs to a novel OM43 subgroup, separated from the H-RS cluster and HTCC2181 cluster. Two phages that infect FZCC0133 were isolated and designated as MEP401 and MEP402. A metagenomic mining was performed to identify metagenomic viral genomes (MVGs) related to MEP401 and MEP402. Genomic analyses revealed that MEP401, MEP402, and all the recovered MVGs represent a novel phage subfamily in the family *Zobellviridae* that can be separated into two genus-level groups. A metagenomic recruitment analysis was also performed to illustrate the global distribution patterns of these phages.

RESULTS

Host strain

The OM43 strain FZCC0133 was isolated in 2017, from the coastal water of Pingtan Island, Fujian, China (25°26'N, 119°47'E), using the dilution-to-extinction method. 16S rRNA gene analysis showed that FZCC0133 is more closely related to OM43 strain H5P1 (98.90% 16S rRNA gene sequence identity) than H-RS and HTCC2181 (96.72% and 97.64% 16S rRNA gene sequence identity, respectively). Phylogenetic analysis based on 16S rRNA gene sequences showed that FZCC0133 falls within the OM43 clade and is distinct from the known H-RS and HTCC2181 clusters. This result suggests that FZCC0133 and H5P1 belong to a new cluster within the OM43 clade (Fig. S1).

Isolation and biological features of MEP401 and MEP402

The OM43 phage MEP401 was isolated from the coastal surface water of Yantai, Bohai Sea, China (37°28'N, 121°29'E), whereas MEP402 was isolated from the coastal surface water of Osaka Bay, Japan (34°27'N, 135°21'E). The transmission electron microscopy (TEM) images showed that the two phages have an icosahedral capsid of 62 ± 2 nm in diameter with an obscure short tail (Fig. 1A). The morphological characteristics of these two phages suggest that they belong to the order *Caudoviricetes*.

Infections of FZCC0133 cell cultures by both MEP401 and MEP402 caused sharp declines in host numbers and produced phage particles (Fig. 1B). The latent periods of MEP401 and MEP402 are both approximately 17–20 hours.
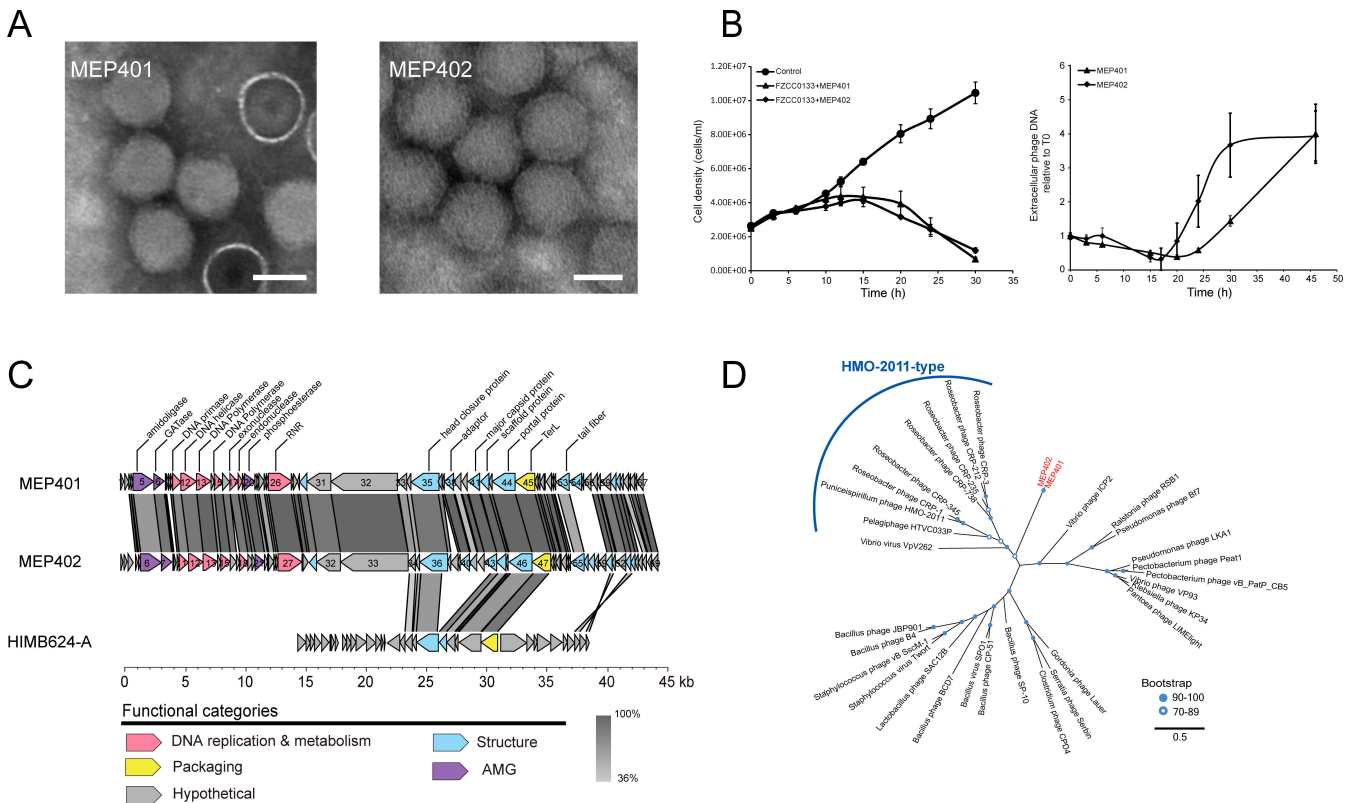

FIG 1 General biological and genomic characteristics of MEP401 and MEP402. (A) Transmission electron microscopy images of MEP401 and MEP402. (B) Left panel: Growth curves of host FZCC0133 infected with the two OM43 phages. Right panel: Growth curves of the two OM43 phages on exponential cultures of FZCC0133. (C) Genomic alignment and comparison of MEP401, MEP402, and HIMB624-A. Predicted open reading frames (ORFs) are represented by arrows, and the direction of transcription is indicated by the direction of the arrow. The numbers inside the arrows indicate ORF numbers. ORFs annotated with known functions are marked using distinct colors according to their functions. (D) Maximum-likelihood phylogenetic tree of DNA polymerase constructed with conserved polymerase domains. MEP401 and MEP402 are shown in red.

TABLE 1 General features of two OM43 phages isolated in this study

Phage	Original host	Source water	Depth	Latitude	Longitude	Genome size (bp)	No. of ORFs	% G + C	Accession number
MEP401	FZCC0133	Yantai, Bohai Sea	Surface	N37°28'	E121°29'	42,987	67	34.4	OP830906
MEP402	FZCC0133	Osaka Bay, Japan	Surface	N34°27'	E135°21'	43,738	69	34.3	OP830907

General genome characteristics of MEP401 and MEP402

Genome sequencing revealed that MEP401 and MEP402 have dsDNA genomes of 42,987 bp and 43,738 bp in size, respectively (Table 1). The G + C content of both phage genomes is approximately 34%, which is similar to that of their host FZCC0133 (33.1%). Their G + C content is slightly higher than that of the first OM43 phage Venkman (31.9%) (23) but lower than those of the previously reported OM43 phages MEP301 (44.4%) (38) and Melnitz (37.6%) (39). MEP401 and MEP402 encode 67 and 69 open reading frames (ORFs), respectively. According to the direction of transcription, the two genomes can be divided into two reverse direction genome units (Fig. 1C). ORFs 1–28 in MEP401 and ORFs 1–29 in MEP402 have a reverse transcription direction compared with that of the remaining ORFs in both genomes. A total of 62 ORFs are shared between these two phages with high homology (43%–100% amino acid identity) (Fig. 1C). Genome comparison revealed that MEP401 and MEP402 share high sequence similarity, with 92.19% average nucleotide identity (ANI) and 85.15% average amino acid sequence identity (AAI). Since 95% ANI threshold is used as one of the standards for phage species level demarcation (40, 41), MEP401 and MEP402 can be considered as distinct phage species. Further genomic analysis showed that they do not display substantial genome similarity with any cultured phages and therefore belong to a novel phage group. Approximately 31% of the identified ORFs in both genomes could be assigned putative biological functions based on the sequence similarities and conserved domains. Genes responsible for phage DNA replication and metabolism, structure and DNA packaging, and cell lysis were identified in MEP401 and MEP402.

In the DNA replication region, both genomes contain genes encoding DNA helicase, DNA primase, DNA polymerase (DNAP), endonuclease, exonuclease, ribonucleoside-triphosphate reductase (RNR), and a few proteins with unknown functions (Fig. 1C). DNAPs in MEP401 and MEP402 are encoded by two ORFs (ORF13 and ORF15 in both genomes) (Fig. 1C), which are segmented by a small unknown protein (ORF14 in both genomes). ORF13 in MEP401 and MEP402 contain the 3′–5′ exonuclease domain (PF01612) and the *N*-terminal DNA polymerase A domain (PF00476), whereas ORF15 in both phages contain the *C*-terminal DNA polymerase A domain (PF00476). In many phage genomes, group I introns are inserted into phage DNAP genes (42–44). Thus, it was possible that DNAP genes in MEP401 and MEP402 both phages were interrupted by an intron.

The HMO-2011-type phages and *Vibrio* phage ICP2 appear to be the closest relatives of MEP401 and MEP402 based on sequence analysis. The DNAPs of MEP401 and MEP402 share 24.8%–31.2% amino acid identities with those in HMO-2011-type phages. However, DNAPs in MEP401 and MEP402 do not possess the unusual domain structure of typical HMO-2011-type DNAPs (19, 24) and only share a few genes with HMO-2011-type phages. These results suggest that MEP401 and MEP402 are distinct from HMO-2011-type phages. The DNAPs of MEP401 and MEP402 share 28.0% and 28.2% amino acid identities with that in phage ICP2, respectively. In addition, both MEP401 and MEP402 possess 13 other genes that are homologous to those in ICP2, suggesting that they may have evolutionary relatedness with ICP2. Phylogenetic analysis of the DNA polymerase family A domain confirmed that MEP401 and MEP402 are distinct from other known phages, representing a novel phage group (Fig. 1D). The DNA primases of MEP401 and MEP402 share weak homology with those in *Ralstonia* phage RSB1 (NC_011201) (31.4% amino acid identity) and *Burkholderia* phage Bp-AMP4 (HG796221) (30.5% amino acid identity). The DNA helicases of MEP401 and MEP402 share sequence similarity with that in *Burkholderia* phage JG068 (NC_022916) with approximately 40% amino acid identity. The DNA endonucleases of MEP401 and MEP402 are similar to those in

Cyanophage KBS-S-1A (JF974297) (~43% amino acid identity). Their DNA exonucleases share approximately 38% amino acid identity with that of *Roseobacter* phage CRP-5 (MK613347).

In their structural and packaging regions, MEP401 and MEP402 have genes predicted to be involved in phage morphogenesis and packaging, including genes encoding major capsid protein (MCP), scaffold protein, tail protein, portal protein, and terminase large subunit (TerL). Some of the structural genes in MEP401 and MEP402 share similarities with those in other phages. For example, their MCPs are homologous to that in *Vibrio* virus VPMCC5 (OL676770) with 53.4% amino acid identity. Their scaffold proteins share 38.2% amino acid identity with that in *Salinivibrio* phage CW02 (NC_019540). The TerLs of MEP401 and MEP402 share approximately 54.0% amino acid identity with that of *Lentibacter* phage vB_LenP_ICBM2 (NC_048671). Their portal proteins are similar to that in *Citrobacter* phage CVT22 (51.2% amino acid identity).

When searching against the NCBI database, we noticed that MEP401 and MEP402 share high similarity with an unfinished OM43 phage HIM624-A (AFB70783.1), which infects *Methylophilacea* HIMB624. The partial genome of HIM624-A is 23,493 bp in size, encoding 35 ORFs. A total of 14 ORFs in HIM624-A have homologs in MEP401 and MEP402, with 22%–83% amino acid identity (Fig. 1C). The partial genome of HIM624-A only covers the structural and packaging module and does not contain any replication-related genes. This analysis suggests that HIMB624-A may belong to the MEP401-type phage group.

Identification of MVGs related to MEP401 and MEP402 from the metagenomic data sets

To expand our knowledge of the genomic diversity of phages related to MEP401 and MEP402, a search was performed to retrieve MVGs from environmental metagenomes. After the workflow for MVGs retrieval, a total of 99 MVGs (>50% genome completeness) from various oceanic stations were extracted for further analysis. Most of these MVGs originated from surface waters (0–200 m) of polar and estuarine regions (Table S1).

A phylogenetic tree based on the viral proteome was constructed by ViPTree. MEP401, MEP402, and the retrieved MVGs were located adjacent to phages of the family *Zobellviridae* in the ViPTree, suggesting their evolutionary relatedness with members of this family. Whole genome-based VICTOR phylogeny analysis showed that MEP401, MEP402, and all retrieved MVGs belong to the family *Zobellviridae* but form a distinct subfamily. The subfamily can be further divided into two groups, with 64 MVGs clustering with MEP401 and MEP402 (group I), and the remaining 35 MVGs forming a separate group (group II) (Fig. 2). All group I MVGs were classified into the same genus with MEP401 and MEP402 using VICTOR (Fig. 2). Host prediction using the RaFAH tool also assigned OM43 bacteria as potential hosts of most group I MVGs (Table S1). These group I MVGs range in size from 20.4 to 47.7 kb (50.94%–100% genome completeness), encoding 27–73 ORFs, and their G + C content ranges from 31.1% to 38.6%. The 35 MVGs in group II were assigned into a different genus by VICTOR analysis (Fig. 2). MVGs in this group have a genome size ranging from 18.1 to 44.7 kb (50.84%–100% genome completeness), with a G + C content ranging from 32.3% to 46.9%. Host prediction using the RaFAH tool suggested that most group II phages also infect OM43 bacteria (Table S1).

Conserved genomic structure and gene content variation

Pangenome analysis identified 328 orthologous protein groups (containing two or more members), of which 91 protein groups could be assigned biological functions (Table S2). Genomic comparison revealed that genome synteny was well conserved across all group I genomes, with few genomic rearrangements (Fig. 3). The functional module structure of all group I MVGs is similar to that of MEP401 and MEP402, and they have significant sequence similarity in the DNA replication and metabolism module and in the structural and DNA packaging module. All group I MVGs were also referred to as MEP401-type MVGs. Split DNA polymerases were also observed in most group I MVGs
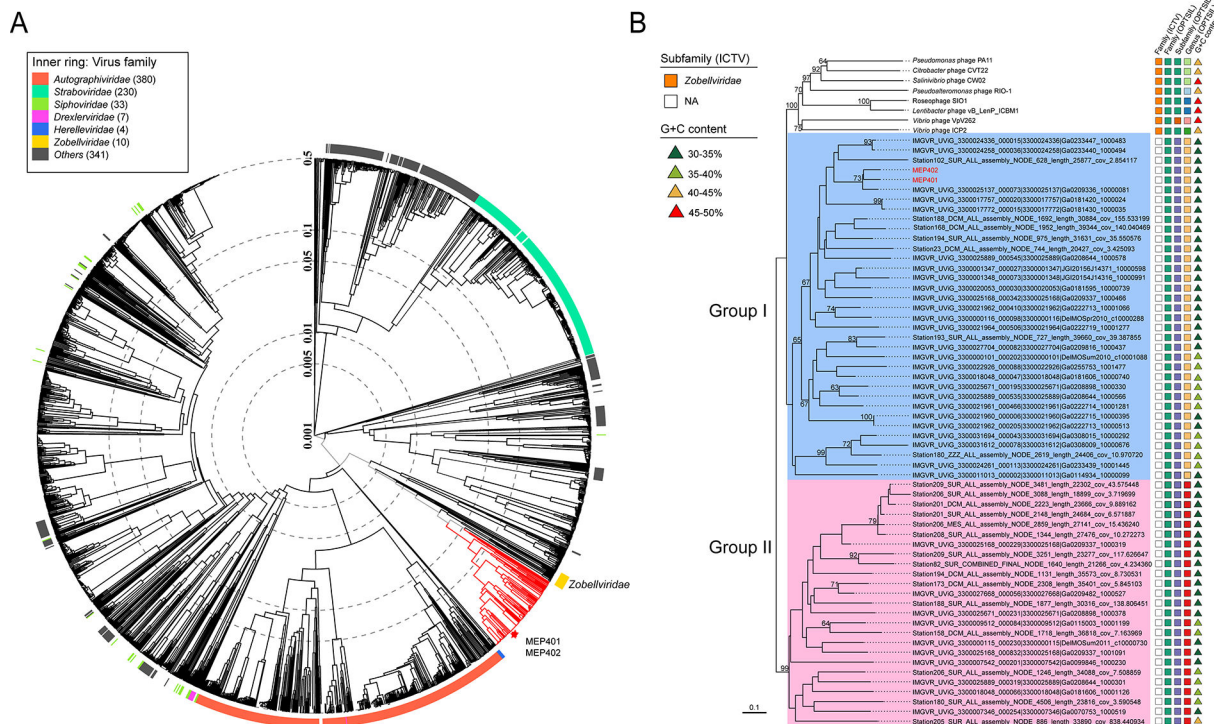
FIG 2 (A) Viral proteomic tree for MEP401, MEP402, and all retrieved MVGs. The tree was generated using ViPTree against all prokaryotic dsDNA viruses. Only the genomes related to the analyzed phages are shown. Red branches indicate MEP401, MEP402, and all retrieved MVGs. MEP401 and MEP402 are indicated with red asterisks. (B) Whole-genome-based phylogenetic tree of MEP401-related phages at the amino acid level. The tree was constructed by VICTOR. All phages were clusters at 95% ANI, and the longest genomes in each cluster were used for VICTOR analysis. The predicted OPTSIL taxon is shown, and the G + C content of each genome is indicated.

(Fig. 3). Core genome analysis based on the complete group I genomes identified 13 core genes that were shared by all the complete group I genomes (Fig. 3). These core genes predominantly encode proteins related to phage DNA replication, phage development, and DNA packaging, suggesting that group I phages employ similar overall replication and propagation processes.

Phylogenomic analysis based on the concatenated sequences of 10 selective core genes was performed to resolve the evolutionary relationship among group I phages. The tree topology was similar to that of the VICTOR analysis. The phylogenomic tree showed that group I phages are diverse and can be separated into two well-supported subgroups (I-A and I-B) (Fig. 4). Subgroup I-A contains MEP401, MEP402, and 56 MVGs from various oceanic regions, while subgroup I-B only contains six MVGs. Subgroups I-A and I-B belong to the same genus according to the VICTOR analysis (Fig. 2B).

Genomic comparisons revealed that group II MVGs also have a DNA replication module similar to those in group I phages (Fig. 5A). In contrast, only a few small genes in the morphogenesis and packaging modules of group II MVGs have homologs in group I genomes. Their MCPs and TerLs are very distantly related to those in group I MVGs but more closely related to those in Cyanophage PP (30%–36.5% amino acid identity) and *Pseudomonas* phage hairong (37%–47.8% amino acid identity). The conserved DNAP and TerL sequences were used for phylogenetic analyses. DNAP phylogeny demonstrated that all group I and II phages cluster together and that DNAPs from group II did not show a good separation from those in group I (Fig. 5B), suggesting that these two groups share a common ancestor. However, the TerL phylogeny showed that group II phages are distinct from group I phages (Fig. 5C), suggesting that they may have arisen from recombination between two different phage ancestors.
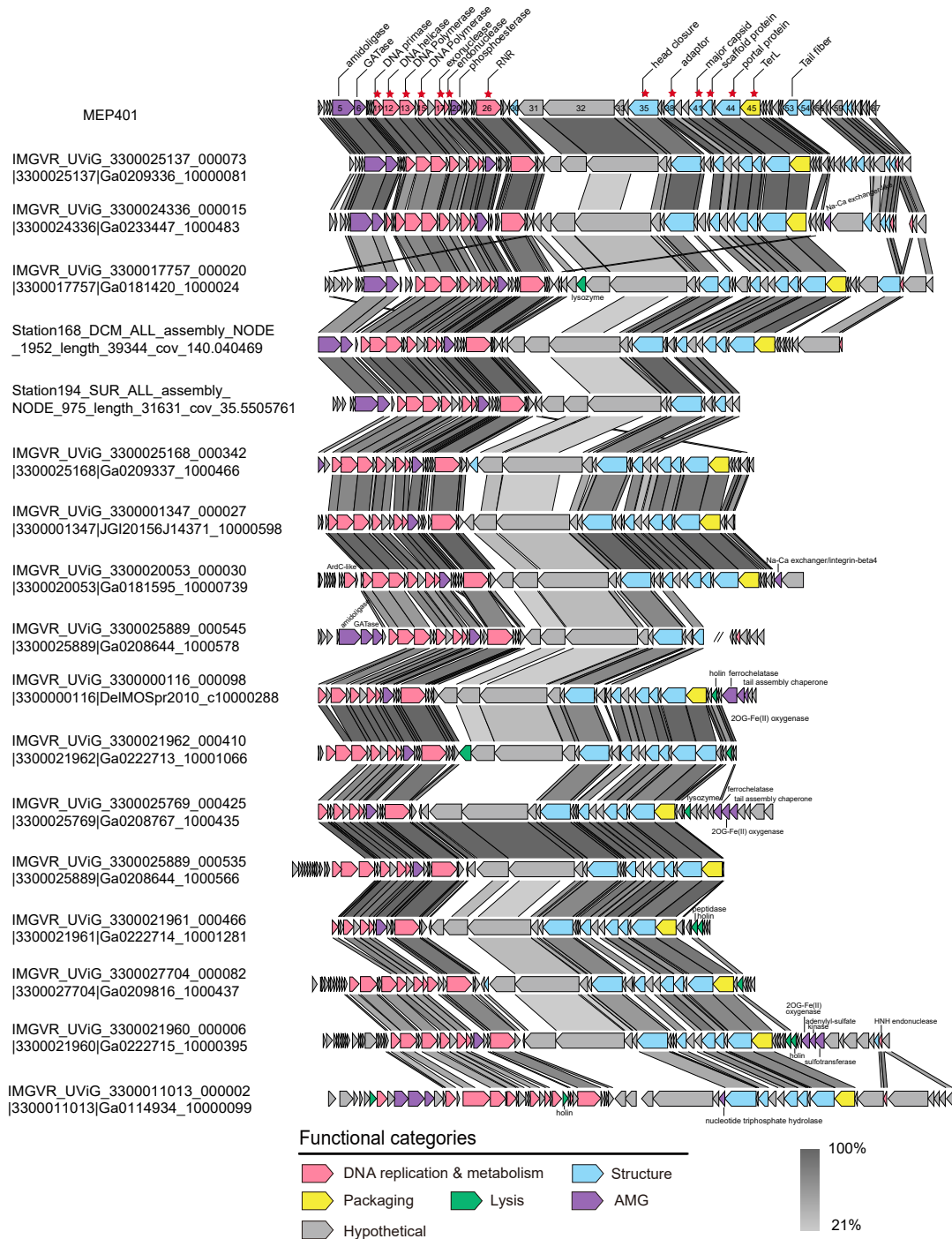
FIG 3 Alignment and comparison of genomes of representative group I phages showing the genomic synteny of this group. Open reading frames (ORFs) are represented by arrows, and the direction of transcription is indicated by the arrow directions. The numbers inside the arrows indicate ORF numbers. ORFs annotated with known functions are marked using distinct colors according to their functions. Core genes are indicated with red asterisks. The color of the shading connecting homologous genes indicates the level of amino acid identity between the genes. RNR, ribonucleoside-triphosphate reductase; TerL, terminase large subunit; GATase, glutamine amidotransferase.

Metabolic capacity

Auxiliary metabolic genes (AMGs) are phage-encoded metabolic genes that are highly similar to their host homologs. Several AMGs that are potentially involved in diverse

metabolic processes have been identified in group I and II genomes. In the DNA replication module, 73 MVGs were found to possess calcineurin-like metallophosphoesterase genes (MPEs, PF00149), sharing homology with those in *Pseudomonas* phage VCM (43.2%–47.8% amino acid identity). This gene was also identified in OM43 phage. Genes encoding amidoligase (PF12224) and glutamine amidotransferase (GATase) (PF13522) were identified in MEP401, MEP402, and over 30 MVGs. Six MVGs from groups I and II contain genes encoding the 2-Oxoglutarate and Fe(II)-dependent oxygenase (2OG-Fe(II) oxygenase) superfamily (PF13640), sharing 29.2%–38.3% amino acid identity with that in *Synechococcus* phage ACG-2014f. Genes encoding the Fe-S cluster assembly scaffold protein (*sufA*, PF01521) were identified in two group I MVGs. These genes have been previously identified in HMO-2011-type MVGs (24). Two MVGs encode the sulfotransferase family protein (Sulfotransfer_3 domain, PF13469), sharing similarity with that in *Synechococcus* phage S-SRM01 (33.2% amino acid identity).
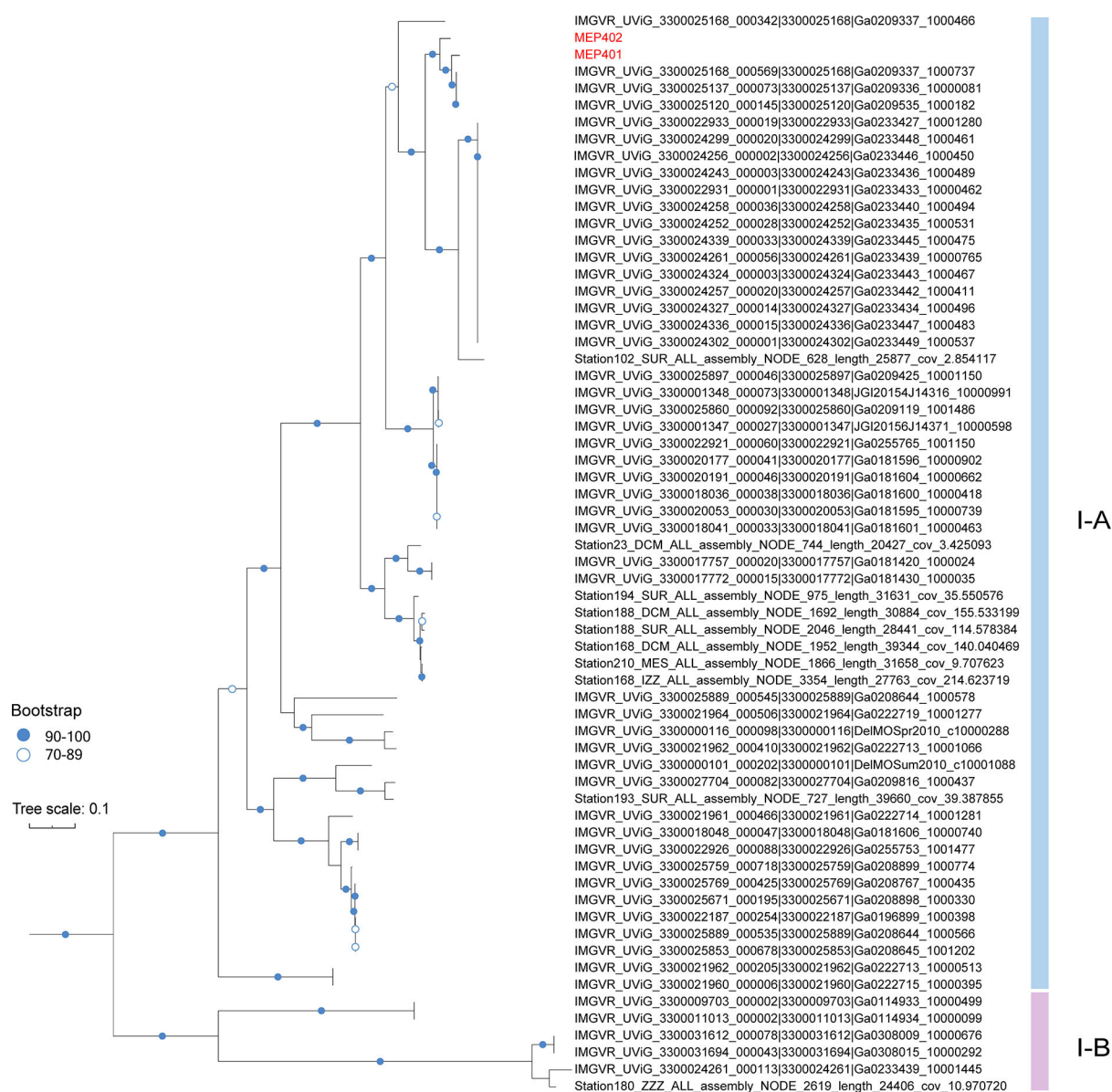

FIG 4 A maximum-likelihood tree was constructed using the concatenated sequences of 10 core genes in group I genomes. The scale bar indicates the number of amino acid substitutions per site.

Distribution in the global ocean

To demonstrate the biogeographical patterns of phages in these two groups, their presence and RPKM (reads per kilobase pair of genomes per million reads) at various oceanic stations were analyzed by mapping the reads from 147 viromes to each phage genome (≥95% nucleotide identity). Generally, these phages were exclusively detected in the surface and mesopelagic waters (0–1,000 m), with varying RPKM values (Fig. 6). Among all analyzed phages, many were more prevalent and abundant in colder regions (Fig. 6). These oceanic regions also exhibit higher chlorophyll values and lower salinities. The linear regression analysis showed significant negative correlation between the RPKM values of some phages and temperature ($P < 0.05$) (Table S3). In addition, these phages were mainly found in oceanic regions and were rarely found in estuarine stations. We found that MEP401, MEP402, and several MVGs closely related to MEP401 and MEP402 exhibited a relatively wider distribution compared with other MVGs. They were detected not only in cold regions but also in Chesapeake Bay and Delaware Bay stations with high temperature and lower salinity (<20 psu). Furthermore, we found that most MVGs that were not correlated with temperature originated from various estuarine stations and were exclusively distributed in estuarine environments (Fig. 6).

DISCUSSION

The marine OM43 clade is one of the most abundant bacterial groups in the ocean. Bacteria in this group have diverse metabolic profiles and play important roles in the metabolism of C1 compounds. Phages that infect marine OM43 bacteria are poorly characterized. In this study, two novel OM43 phages, MEP401 and MEP402, infecting OM43 strain FZCC0133 were isolated. Genomes of these two OM43 phages share high sequence identity with each other but are distinct from other known phages.

Genes related to DNA replication and metabolism, phage structure, and DNA packaging in the two OM43 phages were found sharing sequence identity with some known phages, most of which are *Zobellviridae* phages, implying that MEP401 and MEP402 are evolutionary related to *Zobellviridae* phages. The DNAP genes are responsible for phage genome replication and play an important role in shaping the evolutionarily history and fitness of the phages (45–47). DNAPs of both MEP401-like phages split into two genes with a small unknown ORF inserted. Homing endonucleases-encoding introns were commonly found in many phage genomes (42–44, 48, 49). The introns interrupt various phage genes and can function as self-splicing ribozymes (42–44, 48, 49). It is possible that DNAP genes in MEP401 and MEP402 were also interrupted by an intron. However, in our case, the two inserted ORFs do not show homology to any known gene products. Whether these small ORFs are self-splicing introns is unknown. Their functions remain to be further explored.

Metagenomic mining identified a total of 99 MVGs related to MEP401 and MEP402. Phylogenomic analyses revealed that MEP401, MEP402, and all retrieved MVGs represent a novel phage subfamily in the family *Zobellviridae*. They are distinct from other known members of this family and can further be separated into two genus-level groups (group I and group II). Genome comparison analysis revealed that group I and group II members have similar DNA replication modules, but their morphogenesis and packaging modules are distinct. These findings suggest that phages in group I and group II have a close evolutionary relationship and have undergone genome recombination events. Genetic recombination between phage populations can produce novel combinations of genomic modules and is the main driving force of phage evolution (26–29, 50–53). The genomic characteristics and phylogenetic analyses of group I and group II phages suggest that genome recombination is instrumental in the genome evolution and diversification of OM43 phages.

AMGs are a class of phage-encoded metabolic that have potential roles in regulating host metabolism (3). A set of AMGs were identified from group I and II genomes, including genes encoding MPEs, amidoligase, GATase, 2OG-Fe(II) oxygenase, SufA, and

Sulfotransferse. MPEs exhibit hydrolase activity against a wide range of phosphorylated substrates and are common in bacterial and archaeal genomes (54–56). Phage-encoded
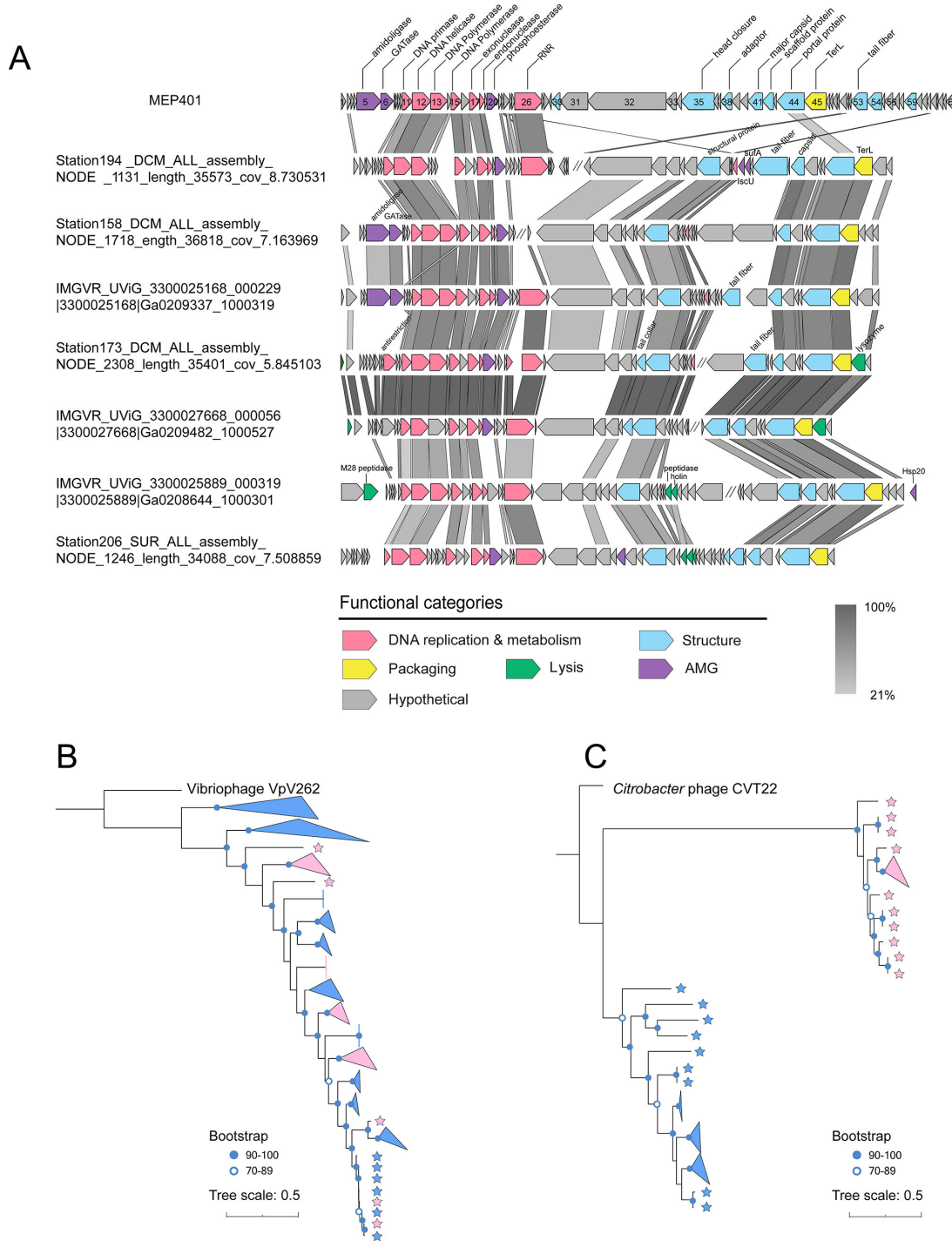

FIG 5 (A) Genomic organization and comparison of group II genomes showing the combination of the MEP401-type DNA replication module and a distinct type of morphogenesis and packaging module. Open reading frames (ORFs) are represented by arrows, and the direction of transcription is indicated by the arrow directions. The numbers inside the arrows indicate ORF numbers. ORFs annotated with known functions are marked using distinct colors according to their functions. The color of the shading connecting homologous genes indicates the level of amino acid identity between the genes. RNR, ribonucleoside-triphosphate reductase; TerL, terminase large subunit; IsuC, Fe-S cluster assembly scaffold protein; SufA, Fe-S cluster assembly scaffold protein. (B) Maximum-likelihood phylogenetic tree of all group I and group II phages based on DNA polymerase sequences. (C) Phylogenetic tree constructed based on the large terminase (TerL) sequences. Group I and group II phages are indicated by blue and pink asterisks, respectively.

MPEs have been identified in many phage genomes (56, 57). A previous study has suggested that *Mycobacterium* virus D29-encoded MPEs can negatively regulate the growth of bacteriophages and bacteria (56). The function of phage-encoded MPEs in OM43 phages requires further investigation. Amidoligase and GATase have been suggested to be involved in modifying the cell wall and thus preventing superinfection by other phages in *Escherichia* phage phiEco32 (58) and suggested to be involved in the synthesis of secondary metabolites and the interactions between phage and hosts in mycobacteriophage Marvin (59). The prevalence of genes encoding these two proteins suggests that they may play a vital role in these phages; however, their specific functions remain to be explored. The 2OG-Fe(II) oxygenase superfamily is typically involved in the oxidation of organic substrate using a dioxygen molecule, which is mainly responsible for protein modification, nucleic acid repair and/or modification, and fatty acid metabolism (60, 61). Fe-S cluster participates in a wide variety of cellular biological processes (62). SufA is a scaffold protein for Fe-S cluster assembly. The discovery of SufA genes in two group I MVGs suggests that these phages may play important roles in Fe-S cluster biogenesis and function. Sulfotransferases are responsible for transferring a sulfate group from 3′-phosphoadenylyl sulfate to a wide range of substrates and therefore have many functions (63). However, the biological function of sulfotransferases in phage genomes remains unclear.

Read-mapping analysis indicated that many phages in these two groups were more prevalent and abundant in cold regions and showed patterns that were similar to some previously reported OM43 phages (23, 38), implying that their hosts were more prevalent in cold waters. Currently, the OM43 lineage has two identified ecotypic clusters, with

FIG 6 Heatmap displaying the RPKM of group I and group II phage in different marine viromic data sets. Normalized relative abundance is depicted as reads recruited per kilobase of contig per million reads (RPKM). DCM, deep chlorophyll maximum; MESO, mesopelagic; SRF, surface; MIX, bottom of mixed layer; DB, Delware Bay; CB, Chesapeake Bay; NA, not available.

the H-RS cluster more abundant in low-chlorophyll a and/or warm water, while the HTCC2181 cluster is more abundant in lower temperature but higher-productivity waters (37). Furthermore, in a previous study, OM43 phages displayed two distinct distribution patterns, similar to the distributions of two known OM43 clusters (37). In this study, we found several MVGs have a distribution pattern that is distinct from the distributions of previously reported OM43 phages. These MVGs were exclusively prevalent in estuarine stations. Considering that the distribution of phages is broadly correlated with the distribution of their hosts, it is possible that these MVGs infect an OM43 cluster that is specifically adapted to estuarine environments.

Conclusion

In this study, two new phages (MEP401 and MEP402) that infect the marine OM43 strain FZCC0133 were isolated and sequenced. In addition, 99 MVGs related to MEP401 and MEP402 were identified using a metagenomics-based mining analysis. Through comparative genomic and phylogenomic analyses, these phages were shown to represent a novel phage subfamily that can be separated into two phage groups with distinct types of morphogenesis and packaging modules. These results further support the idea that genetic recombination plays an important role in the generation of phage genetic diversity. Furthermore, metagenomic mapping analysis revealed the ubiquity and distinct distribution patterns among the members of these two phage groups. Overall, our study provides novel insights into the diversity and biogeography of phages infecting marine OM43 bacteria and establishes models as valuable tools for evaluating the ecological roles of viruses in marine environments.

MATERIALS AND METHODS

Cultivation, purification, and phylogenetic analysis of OM43 strain FZCC0133

The OM43 strain FZCC0133 was isolated in May 2017 from the coastal waters of Pingtan Island in China (N25°26′, E119°47′) using the dilution-to-extinction method with low-nutrient medium (64). FZCC0133 was grown in a sterilized natural seawater–based medium with 100 μM methanol, 1 mM NH_4Cl, 100 μM KH_2PO_4, 1 μM $FeCl_3$, and a vitamin mixture (65). The FZCC0133 cultures were incubated at 23°C in the dark without shaking. The 16S rRNA gene of FZCC0133 was amplified using PCR with the universal primers 16S-27F and 16S-1492R (66). The 16S rRNA gene sequence of the strain was obtained by Sanger sequencing and assembled using Chro-masPro (Technelysium Pty. Ltd., Tewantin, QLD, Australia). A phylogenetic tree based on 16S rRNA gene sequences was constructed using IQ-TREE v1.6.12 (67) with 1,000 bootstrap replicates.

Source waters and OM43 phage isolation

The water samples used to isolate OM43 phages were collected from two different coastal stations: Yantai coast, Bohai Sea, China (37°28'N, 121°29'E), and Osaka Bay, Japan (34°27′N, 135°21′E). The seawater samples were filtered through 0.1-μm-pore-size filters (Pall Life Sciences) to remove nonviral components and stored at 4°C prior to use. Details of phage isolation have been previously described (18, 20, 22). Briefly, filtered seawater samples were inoculated with exponential phase FZCC0133 culture. A Guava EasyCyte flow cytometer (Merck Millipore, Billerica, MA) was used to monitor cell growth. For cultures that displayed a significant cell decrease, the presence of phage particles was confirmed using epifluorescence microscopy (68) Purified phage clones were obtained using the dilution-to-extinction method (18, 20), and phage purity was verified using genome sequencing.

Transmission electron microscopy and growth experiments

The morphology of the isolated OM43 phages was observed using TEM. The OM43 phages lysate was filtered through a 0.1-μm filter, concentrated using Amicon Ultra

centrifugal filters (30 kDa; Merck Millipore), and centrifuged by ultracentrifugation (Beckman Coulter, USA) at 50,000 × g for 2 hours. A drop of the concentrated phage sample was placed on a copper TEM grid and subsequently dried in the air. The grid was stained for 2 minutes with 2% uranyl acetate and then observed under a Hitachi TEM at a voltage of 80 kV. The growth experiments were performed as described previously (18). For the growth curves of hosts, exponentially growing cultures of FZCC0133 were inoculated with the phages at a phage-to-host ratio of ~3. Cultures without the addition of phages were set as controls. Cell counts were determined using a flow cytometer. For the growth curves of phages, exponentially growing cultures of FZCC0133 were inoculated with the phages at a phage-to-host ratio of ~0.1. After the addition of the phages, an aliquot of the cell suspension was collected for phage enumeration from each culture every 2–6 hours for 30 hours, and the relative abundance of phage particles was quantified by quantitative PCR (qPCR). Two pairs of specific qPCR primers (Table S4) were designed using Primer-BLAST tool, targeting the DNAP sequences of the two OM43 phages. Phage lysates from different time points were used as templates for qPCR. Reaction was performed using the Taq Pro Universal SYBR qPCR Master Mix kit (Vazyme, Nanjing, China) with three biological replicates and three technical repeats. qPCR conditions were as follows: initial denaturation at 95°C for 3 minutes, followed by 40 cycles of 95°C for 5 seconds, 60°C for 15 seconds.

Phage concentration, DNA extraction, and genome sequencing

Each phage lysate (150 mL) was filtered through 0.1-µm filters to remove cell debris and then concentrated to approximately 300 µL using Amicon Ultra Centrifugal Filters (30 kDa; Merck Millipore) and Nanosep Centrifugal Devices (30 kDa; Pall Life Sciences). Phage genomic DNA was extracted using a Blood & Tissue Kit (Qiagen, Hilden, Germany). Whole-genome sequencing of MEP401 and MEP402 were conducted using the Illumina HiSeq 2500 platform (paired-end technology 2 × 150 bp). Quality-filtering, trimming, and *de novo* assembly were performed using the CLC Genomic Workbench v11.0.1 (Qiagen, Hilden, Germany) with default settings. Gap closing of the phage genomes was performed using Sanger sequencing of PCR products covering the gap areas.

Metagenomic retrieval of MVGs related to MEP401 and MEP402

For our analyses, marine MVGs reconstructed from the IMG/VR v3 database (17), Global Ocean Viromes (GOV and GOV 2.0) (11, 12), MedDCM fosmid library (8), Station ALOHA assembly free virus genomes (10), ALOHA 2.0 viromic database (9), 78 marine viromes from Metavir (69, 70), and 34 viral fosmids obtained from the oxic surface and oxygen-starved basin waters of Saanich Inlet (71) were downloaded. DNAP genes in MEP401 and MEP402 were used as baits to retrieve the related MVGs. Profile hidden Markov models (HMMs) were constructed using DNAP protein sequences using hmmbuild with default parameters (72). The HMM profiles were used to query the downloaded MVGs (≥10 kb) using the hmmSearch program (e-value ≤ 10^{-3} and score ≥50). Only matches with ≥25% aminio acid identity were considered. Of these MVGs, those containing HMO-2011-type DNAP (24) were removed. This analysis procedure retrieved 5,256 MVGs containing MEP401 DNAP homologs. The whole-genome phylogenetic analysis of these 5,265 MVGs, MEP401, MEP402, and bacterial virus genomes downloaded from NCBI-RefSeq (v96) was then performed using the GL-UVAB workflow (73) with the Dice coefficient under default settings. Taxonomic classification of phages at the subfamily level was performed according to the recommended minimum node depth of 0.0056 and the number of representatives ≥3. A total of 2,915 MVGs classified into the same phage family with MEP401 and MEP402 were used for further analysis. Protein clustering networks for taxonomic assignment were performed using vConTACT 2.0 (74) with default settings. Viral clusters were identified using ClusterONE (75) with the default parameters defined in vConTACT 2.0. The network was visualized using an edge-weighted spring embedded model in Cytoscape v3.8.0. Here, only MVGs that have genome-genome similarity score of ≥1 were considered. CheckV v0.8.1 was used for

the completeness and quality estimation of these MVGs (76). MVGs with a genome completeness ≥50% were used for comparative genomic and ecological analyses.

Genome annotation and comparative genomic analysis

The GeneMark online server (77) and Prodigal (78) were used to predict ORFs from all phage genomes. Translated ORFs were analyzed and annotated using BLASTP against the NCBI nonredundant and NCBI Refseq databases (e-value ≤10^{-3}; ≥25% amino acid identity; ≥50% alignment length). The ORFs were searched against the Pfam (79) database using the HMMER web server (80) for recognizable conserved PFAM domains. For structure and function prediction, we also used the Conserved Domain Search Service of the NCBI (81) and HHpred servers (82). The phage genomes were compared and visualized using Easyfig v2.2.2 (83). OrthoFinder v2.5.2 (84) was used to identify groups of orthologous genes based on sequence similarity (BLASTP option: e-value ≤10^{-3}; ≥25% identity; ≥50% alignment length). tRNA scan-SE was used to identify the tRNA genes (85). The ANI between genomes was calculated using fastANI v1.3 (86), and the AAI values were obtained using CompareM v.0.0.1 software (https://github.com/dparks1134/CompareM).

Phylogenomic analyses

A viral proteomic tree was constructed created using ViPTree (87). Whole-genome phylogeny based on amino acid sequences was built using VICTOR (88) with the Genome-BLAST Distance Phylogeny (GBDP) method under recommended settings for prokaryotic viruses. The analyzed genomes were compared using CD-hit (89) with a nucleotide identity of ≥95% and ≥80% of the short genome (-c 0.95 -aS 0.8), and only the longest MVGs within a species cluster were retained for VICTOR analysis. Genome-based classification at the genus and family level was performed using the OPTSIL program (90). We conducted phylogenomic analyses to evaluate the evolutionary relationships between group I phages. Ten core genes were selected for phylogenomic analysis (DNA helicase, DNAP, capsid, portal, DNA primase, exonuclease, endonuclease, adaptor protein, scaffold, and TerL). Core gene sequences were aligned using MAFFT-7.455 (91) and edited using trimALv1.4.1 (92). The alignments were concatenated and a phylogenetic tree was constructed using IQ-TREE v1.6.12 (67) with 1,000 bootstrap replicates.

Phylogenetic trees of the DNAP and TerL sequences were also constructed to reveal the evolutionary relationships between the two phage groups. Amino acid sequence alignment and editing were performed using MAFFT-7.455 (91) and trimAl v1.4.1 (92), respectively. ProtTest3.4.2 was used to evaluate the optimal model and run with IQ-TREE v1.6.12. All phylogenetic trees were visualized using Interactive Tree Of Life ITOL v.5 (93).

Recruitment of metagenomic reads and statistical analysis

The relative abundance of MEP401, MEP402, and related MVGs were estimated using a viromic read-mapping analysis. A total of 147 data sets were used for viromic read-mapping analysis. The analyzed genomes were compared using CD-hit (89) with a nucleotide identity of ≥95% and ≥80% of the short genome (-c 0.95 -aS 0.8), and only the longest MVGs within a species cluster were retained for recruitment analysis. Viromic reads were mapped against the nonredundant set of analyzed phage genomes using coverm with a nucleotide identity of ≥95%, alignment length of >50 bp (-p bwa-mem--min-read-percent-identity 95--min-read-aligned-length 50). The relative abundances of these phages were normalized by mapped reads per kilobase pair of genomes per million reads (RPKM). Phage genomes for which <40% of the genomes were covered by recruited viromic reads in a given data set were regarded as absent and were assigned an RPKM value of 0 (23). A heatmap of the RPKM of phages was generated using the pheatmap package in R. Linear regression analysis generated using R was used to test

the relationship between environmental parameters and the relative abundance of these phages.

Host prediction

Potential hosts of MEP401-related MVGs were predicted using the RaFAH tool with default settings (94). The training and validating random forest model for RaFAH was built using 4,269 host-known phages, MEP401 and MEP402.

AUTHOR AFFILIATIONS

[1]Fujian Provincial Key Laboratory of Agroecological Processing and Safety Monitoring, College of Life Sciences, Fujian Agriculture and Forestry University, Fuzhou, China
[2]Key Laboratory of Marine Biotechnology of Fujian Province, Institute of Oceanology, Fujian Agriculture and Forestry University, Fuzhou, China

AUTHOR ORCIDs

Mingyu Yang http://orcid.org/0009-0009-5521-5272

DATA AVAILABILITY

The 16S rRNA sequence of FZCC0133 has been deposited in the GenBank database under the accession number OQ306544. The genome sequences of MEP401 and MEP402 have been deposited in the GenBank database under the accession numbers OP830906 and OP830907.

ADDITIONAL FILES

The following material is available online.

Supplemental Material

Supplemental material (Spectrum04942-22-s0001.pdf). Fig. S1 and Tables S1, S3, and S4.
Table S2 (Spectrum04942-22-s0002.xlsx). Protein orthologous groups.

REFERENCES

1. Suttle CA. 2007. Marine viruses — major players in the global ecosystem. Nat Rev Microbiol 5:801–812. https://doi.org/10.1038/nrmicro1750
2. Brussaard CPD, Wilhelm SW, Thingstad F, Weinbauer MG, Bratbak G, Heldal M, Kimmance SA, Middelboe M, Nagasaki K, Paul JH, Schroeder DC, Suttle CA, Vaqué D, Wommack KE. 2008. Global-scale processes with a nanoscale drive: the role of marine viruses. ISME J 2:575–578. https://doi.org/10.1038/ismej.2008.31
3. Breitbart M. 2012. Marine viruses: truth or dare. Ann Rev Mar Sci 4:425–448. https://doi.org/10.1146/annurev-marine-120709-142805
4. Fuhrman JA. 1999. Marine viruses and their biogeochemical and ecological effects. Nature 399:541–548. https://doi.org/10.1038/21119
5. Cobián Güemes AG, Youle M, Cantú VA, Felts B, Nulton J, Rohwer F. 2016. Viruses as winners in the game of life. Annu Rev Virol 3:197–214. https://doi.org/10.1146/annurev-virology-100114-054952
6. Wommack KE, Colwell RR. 2000. Virioplankton: viruses in aquatic ecosystems. Microbiol Mol Biol Rev 64:69–114. https://doi.org/10.1128/MMBR.64.1.69-114.2000
7. 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
8. 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
9. Luo E, Eppley JM, Romano AE, Mende DR, DeLong EF. 2020. Double-stranded DNA virioplankton dynamics and reproductive strategies in the oligotrophic open ocean water column. ISME J 14:1304–1315. https://doi.org/10.1038/s41396-020-0604-8

10. Beaulaurier J, Luo E, Eppley JM, Uyl PD, Dai X, Burger A, Turner DJ, Pendelton M, Juul S, Harrington E, DeLong EF. 2020. Assembly-free single-molecule sequencing recovers complete virus genomes from natural microbial communities. Genome Res 30:437–446. https://doi.org/10.1101/gr.251686.119

11. Tara Oceans Coordinators, 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, 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

12. Gregory AC, Zayed AA, Conceição-Neto N, Temperton B, Bolduc B, Alberti A, Ardyna M, Arkhipova K, Carmichael M, Cruaud C, Dimier C, Domínguez-Huerta G, Ferland J, Kandels S, Liu Y, Marec C, Pesant S, Picheral M, Pisarev S, Poulain J, Tremblay J-É, Vik D, Tara Oceans Coordinators, Babin M, Bowler C, Culley AI, de Vargas C, Dutilh BE, Iudicone D, Karp-Boss L, Roux S, Sunagawa S, Wincker P, Sullivan MB. 2019. Marine DNA viral macro- and microdiversity from pole to pole. Cell 177:1109–1123. https://doi.org/10.1016/j.cell.2019.03.040

13. Roux S, Hawley AK, Torres Beltran M, Scofield M, Schwientek P, Stepanauskas R, Woyke T, Hallam SJ, Sullivan MB. 2014. Ecology and evolution of viruses infecting uncultivated SUP05 bacteria as revealed by single-cell- and meta-genomics. Elife 3:e03125. https://doi.org/10.7554/eLife.03125

14. 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

15. Martinez-Hernandez F, Fornas O, Lluesma Gomez M, Bolduc B, de la Cruz Peña MJ, Martínez JM, Anton J, Gasol JM, Rosselli R, Rodriguez-Valera F, Sullivan MB, Acinas SG, Martinez-Garcia M. 2017. Single-virus genomics reveals hidden cosmopolitan and abundant viruses. Nat Commun 8:15892. https://doi.org/10.1038/ncomms15892

16. Roux S, Adriaenssens EM, Dutilh BE, Koonin EV, Kropinski AM, Krupovic M, Kuhn JH, Lavigne R, Brister JR, Varsani A, Amid C, Aziz RK, Bordenstein SR, Bork P, Breitbart M, Cochrane GR, Daly RA, Desnues C, Duhaime MB, Emerson JB, Enault F, Fuhrman JA, Hingamp P, Hugenholtz P, Hurwitz BL, Ivanova NN, Labonté JM, Lee K-B, Malmstrom RR, Martinez-Garcia M, Mizrachi IK, Ogata H, Páez-Espino D, Petit M-A, Putonti C, Rattei T, Reyes A, Rodriguez-Valera F, Rosario K, Schriml L, Schulz F, Steward GF, Sullivan MB, Sunagawa S, Suttle CA, Temperton B, Tringe SG, Thurber RV, Webster NS, Whiteson KL, Wilhelm SW, Wommack KE, Woyke T, Wrighton KC, Yilmaz P, Yoshida T, Young MJ, Yutin N, Allen LZ, Kyrpides NC, Eloe-Fadrosh EA. 2019. Minimum information about an uncultivated virus genome (MIUViG). Nat Biotechnol 37:29–37. https://doi.org/10.1038/nbt.4306

17. Roux S, Páez-Espino D, Chen I-MA, Palaniappan K, Ratner A, Chu K, Reddy TBK, Nayfach S, Schulz F, Call L, Neches RY, Woyke T, Ivanova NN, Eloe-Fadrosh EA, Kyrpides NC. 2021. IMG/VR v3: an integrated ecological and evolutionary framework for interrogating genomes of uncultivated viruses. Nucleic Acids Res 49:D764–D775. https://doi.org/10.1093/nar/gkaa946

18. 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

19. Kang I, Oh H-M, Kang D, Cho J-C. 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

20. Zhang Z, Chen F, Chu X, Zhang H, Luo H, Qin F, Zhai Z, Yang M, Sun J, Zhao Y, Rappe MS. 2019. Diverse, abundant, and novel viruses infecting the marine *Roseobacter* RCA lineage . mSystems 4:e00494-419. https://doi.org/10.1128/mSystems.00494-19

21. Zhao Y, Qin F, Zhang R, Giovannoni SJ, Zhang Z, Sun J, Du S, Rensing C. 2019. Pelagiphages in the *Podoviridae* family integrate into host genomes. Environ Microbiol 21:1989–2001. https://doi.org/10.1111/1462-2920.14487

22. Zhang Z, Qin F, Chen F, Chu X, Luo H, Zhang R, Du S, Tian Z, Zhao Y. 2021. Culturing novel and abundant pelagiphages in the ocean. Environ Microbiol 23:1145–1161. https://doi.org/10.1111/1462-2920.15272

23. Buchholz HH, Michelsen ML, Bolaños LM, Browne E, Allen MJ, Temperton B. 2021. Efficient dilution-to-extinction isolation of novel virus–host model systems for fastidious heterotrophic bacteria. ISME J 15:1585–1598. https://doi.org/10.1038/s41396-020-00872-z

24. Qin F, Du S, Zhang Z, Ying H, Wu Y, Zhao G, Yang M, Zhao Y. 2022. Newly identified HMO-2011-type phages reveal genomic diversity and biogeographic distributions of this marine viral group. ISME J 16:1363–1375. https://doi.org/10.1038/s41396-021-01183-7

25. Hendrix RW, Smith MC, Burns RN, Ford ME, Hatfull GF. 1999. Evolutionary relationships among diverse bacteriophages and prophages: all the world's a phage. Proc Natl Acad Sci U S A 96:2192–2197. https://doi.org/10.1073/pnas.96.5.2192

26. Hatfull GF, Hendrix RW. 2011. Bacteriophages and their genomes. Curr Opin Virol 1:298–303. https://doi.org/10.1016/j.coviro.2011.06.009

27. Pedulla ML, Ford ME, Houtz JM, Karthikeyan T, Wadsworth C, Lewis JA, Jacobs-Sera D, Falbo J, Gross J, Pannunzio NR, Brucker W, Kumar V, Kandasamy J, Keenan L, Bardarov S, Kriakov J, Lawrence JG, Jacobs WR, Hendrix RW, Hatfull GF. 2003. Origins of highly mosaic mycobacteriophage genomes. Cell 113:171–182. https://doi.org/10.1016/s0092-8674(03)00233-2

28. Lucchini S, Desiere F, Brüssow H. 1999. Comparative genomics of *Streptococcus thermophilus* phage species supports a modular evolution theory. J Virol 73:8647–8656. https://doi.org/10.1128/JVI.73.10.8647-8656.1999

29. Kupczok A, Neve H, Huang KD, Hoeppner MP, Heller KJ, Franz C, Dagan T. 2018. Rates of mutation and recombination in *Siphoviridae* phage genome evolution over three decades. Mol Biol Evol 35:1147–1159. https://doi.org/10.1093/molbev/msy027

30. Rappé MS, Kemp PF, Giovannoni SJ. 1997. Phylogenetic diversity of marine coastal picoplankton 16S rRNA genes cloned from the continental shelf off Cape Hatteras, North Carolina. Limnol Oceanogr 42:811–826. https://doi.org/10.4319/lo.1997.42.5.0811

31. Rappé MS, Vergin K, Giovannoni SJ. 2000. Phylogenetic comparisons of a coastal bacterioplankton community with its counterparts in open ocean and freshwater systems. FEMS Microbiol Ecol 33:219–232. https://doi.org/10.1111/j.1574-6941.2000.tb00744.x

32. Suzuki MT, Preston CM, Béjà O, de la Torre JR, Steward GF, DeLong EF. 2004. Phylogenetic screening of ribosomal RNA gene-containing clones in bacterial artificial chromosome (BAC) libraries from different depths in monterey bay. Microb Ecol 48:473–488. https://doi.org/10.1007/s00248-004-0213-5

33. Morris RM, Longnecker K, Giovannoni SJ. 2006. Pirellula and OM43 are among the dominant lineages identified in an Oregon coast diatom bloom. Environ Microbiol 8:1361–1370. https://doi.org/10.1111/j.1462-2920.2006.01029.x

34. Giovannoni SJ, Hayakawa DH, Tripp HJ, Stingl U, Givan SA, Cho J-C, Oh H-M, Kitner JB, Vergin KL, Rappé MS. 2008. The small genome of an abundant coastal ocean methylotroph. Environ Microbiol 10:1771–1782. https://doi.org/10.1111/j.1462-2920.2008.01598.x

35. Huggett MJ, Hayakawa DH, Rappé MS. 2012. Genome sequence of strain HIMB624, a cultured representative from the OM43 clade of marine betaproteobacteria. Stand Genomic Sci 6:11–20. https://doi.org/10.4056/sigs.2305090

36. Halsey KH, Carter AE, Giovannoni SJ. 2012. Synergistic metabolism of a broad range of C1 compounds in the marine methylotrophic bacterium HTCC2181. Environ Microbiol 14:630–640. https://doi.org/10.1111/j.1462-2920.2011.02605.x

37. Jimenez-Infante F, Ngugi DK, Vinu M, Alam I, Kamau AA, Blom J, Bajic VB, Stingl U. 2016. Comprehensive Genomic analyses of the OM43 clade, including a novel species from the Red sea, indicate ecotype differentiation among marine methylotrophs. Appl Environ Microbiol 82:1215–1226. https://doi.org/10.1128/AEM.02852-15

38. Yang M, Xia Q, Du S, Zhang Z, Qin F, Zhao Y. 2021. Genomic characterization and distribution pattern of a novel marine OM43 phage. Front Microbiol 12:651326. https://doi.org/10.3389/fmicb.2021.651326

39. Buchholz HH, Bolaños LM, Bell AG, Michelsen ML, Allen MJ, Temperton B. 2022. A novel and ubiquitous marine methylophage provides insights into viral-host coevolution and possible host-range expansion in streamlined marine heterotrophic bacteria. Appl Environ Microbiol 88:e0025522. https://doi.org/10.1128/aem.00255-22

40. Goris J, Konstantinidis KT, Klappenbach JA, Coenye T, Vandamme P, Tiedje JM. 2007. DNA–DNA hybridization values and their relationship to whole-genome sequence similarities. Int J Syst Evol Microbiol 57:81–91. https://doi.org/10.1099/ijs.0.64483-0

41. Konstantinidis KT, Tiedje JM. 2005. Towards a genome-based taxonomy for prokaryotes. J Bacteriol 187:6258–6264. https://doi.org/10.1128/JB.187.18.6258-6264.2005

42. Goodrich-Blair H, Shub DA. 1994. The DNA polymerase genes of several HMU-bacteriophages have similar group I introns with highly divergent open reading frames. Nucleic Acids Res 22:3715–3721. https://doi.org/10.1093/nar/22.18.3715

43. Landthaler M, Shub DA. 2003. The nicking homing endonuclease I-BasI is encoded by a group I intron in the DNA polymerase gene of the *Bacillus thuringiensis* phage Bastille. Nucleic Acids Res 31:3071–3077. https://doi.org/10.1093/nar/gkg433

44. Lee C-N, Lin J-W, Weng S-F, Tseng Y-H. 2009. Genomic characterization of the intron-containing T7-like phage phiL7 of *Xanthomonas campestris*. Appl Environ Microbiol 75:7828–7837. https://doi.org/10.1128/AEM.01214-09

45. Nasko DJ, Chopyk J, Sakowski EG, Ferrell BD, Polson SW, Wommack KE. 2018. Family A DNA polymerase phylogeny uncovers diversity and replication gene organization in the virioplankton. Front Microbiol 9:3053. https://doi.org/10.3389/fmicb.2018.03053

46. Wommack KE, Nasko DJ, Chopyk J, Sakowski EG. 2015. Counts and sequences, observations that continue to change our understanding of viruses in nature. J Microbiol 53:181–192. https://doi.org/10.1007/s12275-015-5068-6

47. Doublié S, Tabor S, Long AM, Richardson CC, Ellenberger T. 1998. Crystal structure of a bacteriophage T7 DNA replication complex at 2.2 A resolution. Nature 391:251–258. https://doi.org/10.1038/34593

48. Lambowitz AM, Belfort M. 1993. Introns as mobile genetic elements. Annu Rev Biochem 62:587–622. https://doi.org/10.1146/annurev.bi.62.070193.003103

49. Edgell DR, Chalamcharla VR, Belfort M. 2011. Learning to live together: mutualism between self-splicing introns and their hosts. BMC Biol 9:22. https://doi.org/10.1186/1741-7007-9-22

50. Jäckel C, Hammerl JA, Reetz J, Kropinski AM, Hertwig S. 2015. Campylobacter group II phage CP21 is the prototype of a new subgroup revealing a distinct modular genome organization and host specificity. BMC Genomics 16:629. https://doi.org/10.1186/s12864-015-1837-1

51. Zhan Y, Huang S, Voget S, Simon M, Chen F. 2016. A novel roseobacter phage possesses features of podoviruses, siphoviruses, prophages and gene transfer agents. Sci Rep 6:30372. https://doi.org/10.1038/srep30372

52. Johnson MC, Sena-Velez M, Washburn BK, Platt GN, Lu S, Brewer TE, Lynn JS, Stroupe ME, Jones KM. 2017. Structure, proteome and genome of *Sinorhizobium meliloti* phage ΦM5: a virus with LUZ24-like morphology and a highly mosaic genome. J Struct Biol 200:343–359. https://doi.org/10.1016/j.jsb.2017.08.005

53. Zhai Z, Zhang Z, Zhao G, Liu X, Qin F, Zhao Y. 2021. Genomic characterization of two novel RCA phages reveals new insights into the diversity and evolution of marine viruses. Microbiol Spectr 9:e0123921. https://doi.org/10.1128/Spectrum.01239-21

54. Koonin EV. 1994. Conserved sequence pattern in a wide variety of phosphoesterases. Protein Sci 3:356–358. https://doi.org/10.1002/pro.5560030218

55. Aravind L, Koonin EV. 1998. Phosphoesterase domains associated with DNA polymerases of diverse origins. Nucleic Acids Res. 26:3746–3752. https://doi.org/10.1093/nar/26.16.3746

56. Dutta S, Bhawsinghka N, Das Gupta SK. 2014. Gp66, a calcineurin family phosphatase encoded by mycobacteriophage D29, is a 2′, 3′ cyclic nucleotide phosphodiesterase that negatively regulates phage growth. FEMS Microbiol Lett 361:84–93. https://doi.org/10.1111/1574-6968.12625

57. Matange N, Podobnik M, Visweswariah SS. 2015. Metallophosphoesterases: structural fidelity with functional promiscuity. Biochem J 467:201–216. https://doi.org/10.1042/BJ20150028

58. Iyer LM, Abhiman S, Maxwell Burroughs A, Aravind L. 2009. Amidoligases with ATP-grasp, glutamine synthetase-like and acetyltransferase-like domains: synthesis of novel metabolites and peptide modifications of proteins. Mol Biosyst 5:1636–1660. https://doi.org/10.1039/b917682a

59. Mageeney C, Pope WH, Harrison M, Moran D, Cross T, Jacobs-Sera D, Hendrix RW, Dunbar D, Hatfull GF. 2012. Mycobacteriophage marvin: A new singleton phage with an unusual genome organization. J Virol 86:4762–4775. https://doi.org/10.1128/JVI.00075-12

60. Aravind L, Koonin EV. 2001. The DNA-repair protein AlkB, EGL-9, and leprecan define new families of 2-oxoglutarate- and iron-dependent dioxygenases. Genome Biol 2:RESEARCH0007. https://doi.org/10.1186/gb-2001-2-3-research0007

61. Zhou J, Bo S, Wang H, Zheng L, Liang P, Zuo Y. 2021. Identification of disease-related 2-oxoglutarate/fe (II)-dependent oxygenase based on reduced amino acid cluster strategy. Front Cell Dev Biol 9:707938. https://doi.org/10.3389/fcell.2021.707938

62. Beinert H, Holm RH, Münck E. 1997. Iron-sulfur clusters: nature's modular, multipurpose structures. Science 277:653–659. https://doi.org/10.1126/science.277.5326.653

63. Chapman E, Best MD, Hanson SR, Wong C-H. 2004. Sulfotransferases: structure, mechanism, biological activity, inhibition, and synthetic utility. Angew Chem Int Ed Engl 43:3526–3548. https://doi.org/10.1002/anie.200300631

64. Connon SA, Giovannoni SJ. 2002. High-throughput methods for culturing microorganisms in very-low-nutrient media yield diverse new marine isolates. Appl Environ Microbiol 68:3878–3885. https://doi.org/10.1128/AEM.68.8.3878-3885.2002

65. Cho J-C, Giovannoni SJ. 2004. Cultivation and growth characteristics of a diverse group of oligotrophic marine Gammaproteobacteria. Appl Environ Microbiol 70:432–440. https://doi.org/10.1128/AEM.70.1.432-440.2004

66. Lane DJ. 1991. 16S/23S rRNA sequencing, p 115–175. In Stackebrandt E, Goodfellow M (ed), Nucleic acid techniques in bacterial systematics. John Wiley & Sons, Chichester, United Kingdom.

67. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R. 2020. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Mol Biol Evol 37:1530–1534. https://doi.org/10.1093/molbev/msaa131

68. Wilhelm SW, Weinbauer MG, Suttle CA. 2010. Enumeration of virus particles in aquatic or sediment samples by Epifluorescence microscopy, p 145–153. In Manual of aquatic viral Ecolog. American Society of Limnology and Oceanography.

69. Coutinho FH, Silveira CB, Gregoracci GB, Thompson CC, Edwards RA, Brussaard CPD, Dutilh BE, Thompson FL. 2017. Marine viruses discovered via metagenomics shed light on viral strategies throughout the oceans. Nat Commun 8:15955. https://doi.org/10.1038/ncomms15955

70. Roux S, Faubladier M, Mahul A, Paulhe N, Bernard A, Debroas D, Enault F. 2011. Metavir: a web server dedicated to virome analysis. Bioinformatics 27:3074–3075. https://doi.org/10.1093/bioinformatics/btr519

71. Chow C-ET, Winget DM, White RA, Hallam SJ, Suttle CA. 2015. Combining genomic sequencing methods to explore viral diversity and reveal potential virus-host interactions. Front Microbiol 6:265. https://doi.org/10.3389/fmicb.2015.00265

72. Eddy SR. 2009. A new generation of Homology search tools based on probabilistic inference. Genome Inform 23:205–211. https://doi.org/10.1142/9781848165632_0019

73. Coutinho FH, Edwards RA, Rodríguez-Valera F. 2019. Charting the diversity of uncultured viruses of archaea and bacteria. BMC Biol 17:109. https://doi.org/10.1186/s12915-019-0723-8

74. Bin Jang H, Bolduc B, Zablocki O, Kuhn JH, Roux S, Adriaenssens EM, Brister JR, Kropinski AM, Krupovic M, Lavigne R, Turner D, Sullivan MB. 2019. Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. Nat Biotechnol 37:632–639. https://doi.org/10.1038/s41587-019-0100-8

75. Nepusz T, Yu H, Paccanaro A. 2012. Detecting overlapping protein complexes in protein-protein interaction networks. Nat Methods 9:471–472. https://doi.org/10.1038/nmeth.1938

76. Nayfach S, Camargo AP, Schulz F, Eloe-Fadrosh E, Roux S, Kyrpides NC. 2021. CheckV assesses the quality and completeness of metagenome-assembled viral genomes. Nat Biotechnol 39:578–585. https://doi.org/10.1038/s41587-020-00774-7

77. Besemer J, Lomsadze A, Borodovsky M. 2001. GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. implications for finding sequence motifs in regulatory regions. Nucleic Acids Res 29:2607–2618. https://doi.org/10.1093/nar/29.12.2607

78. Hyatt D, Chen G-L, Locascio PF, Land ML, Larimer FW, Hauser LJ. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11:119. https://doi.org/10.1186/1471-2105-11-119

79. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, Sonnhammer ELL, Tate J, Punta M. 2014. Pfam: the protein families database. Nucleic Acids Res 42:D222–D230. https://doi.org/10.1093/nar/gkt1223

80. Potter SC, Luciani A, Eddy SR, Park Y, Lopez R, Finn RD. 2018. HMMER web server: 2018 update. Nucleic Acids Res 46:W200–W204. https://doi.org/10.1093/nar/gky448

81. Marchler-Bauer A, Lu S, Anderson JB, Chitsaz F, Derbyshire MK, DeWeese-Scott C, Fong JH, Geer LY, Geer RC, Gonzales NR, Gwadz M, Hurwitz DI, Jackson JD, Ke Z, Lanczycki CJ, Lu F, Marchler GH, Mullokandov M, Omelchenko MV, Robertson CL, Song JS, Thanki N, Yamashita RA, Zhang D, Zhang N, Zheng C, Bryant SH. 2011. CDD: a conserved domain database for the functional annotation of proteins. Nucleic Acids Research 39:D225–D229. https://doi.org/10.1093/nar/gkq1189

82. Söding J, Biegert A, Lupas AN. 2005. The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res 33:W244–W248. https://doi.org/10.1093/nar/gki408

83. Sullivan MJ, Petty NK, Beatson SA. 2011. Easyfig: a genome comparison Visualizer. Bioinformatics 27:1009–1010. https://doi.org/10.1093/bioinformatics/btr039

84. Emms DM, Kelly S. 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol 16:157. https://doi.org/10.1186/s13059-015-0721-2

85. Lowe TM, Eddy SR. 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res 25:955–964. https://doi.org/10.1093/nar/25.5.955

86. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. 2018. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat Commun 9:5114. https://doi.org/10.1038/s41467-018-07641-9

87. Nishimura Y, Yoshida T, Kuronishi M, Uehara H, Ogata H, Goto S. 2017. ViPTree: the viral proteomic tree server. Bioinformatics 33:2379–2380. https://doi.org/10.1093/bioinformatics/btx157

88. Meier-Kolthoff JP, Göker M. 2017. VICTOR: genome-based phylogeny and classification of prokaryotic viruses. Bioinformatics 33:3396–3404. https://doi.org/10.1093/bioinformatics/btx440

89. Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22:1658–1659. https://doi.org/10.1093/bioinformatics/btl158

90. Göker M, García-Blázquez G, Voglmayr H, Tellería MT, Martín MP. 2009. Molecular taxonomy of phytopathogenic fungi: a case study in peronospora. PLoS One 4:e6319. https://doi.org/10.1371/journal.pone.0006319

91. Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast fourier transform. Nucleic Acids Res 30:3059–3066. https://doi.org/10.1093/nar/gkf436

92. 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

93. Letunic I, Bork P. 2016. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Res 44:W242–W245. https://doi.org/10.1093/nar/gkw290

94. Coutinho FH, Zaragoza-Solas A, López-Pérez M, Barylski J, Zielezinski A, Dutilh BE, Edwards R, Rodriguez-Valera F. 2021. RaFAH: host prediction for viruses of bacteria and archaea based on protein content. Patterns (N Y) 2:100274. https://doi.org/10.1016/j.patter.2021.100274