SCIENTIFIC DATA

OPEN

Data Descriptor: Viruses of the Nahant Collection, characterization of 251 marine Vibrionaceae viruses

Kathryn M. Kauffman[1,*], Julia M. Brown[2,*,†], Radhey S. Sharma[1,†], David VanInsberghe[1], Joseph Elsherbini[1], Martin Polz[1,**] & Libusha Kelly[2,**]

Received: 16 January 2018
Accepted: 11 April 2018
Published: 3 July 2018

Viruses are highly discriminating in their interactions with host cells and are thought to play a major role in maintaining diversity of environmental microbes. However, large-scale ecological and genomic studies of co-occurring virus-host pairs, required to characterize the mechanistic and genomic foundations of virus-host interactions, are lacking. Here, we present the largest dataset of cultivated and sequenced co-occurring virus-host pairs that captures ecologically representative fine-scale diversity. Using the ubiquitous and ecologically diverse marine Vibrionaceae as a host platform, we isolate and sequence 251 dsDNA viruses and their hosts from three time points within a 93-day time-series study. The virus collection includes representatives of the three *Caudovirales* tailed virus morphotypes, a novel family of nontailed viruses, and the smallest (10,046 bp) and largest (348,911 bp) *Vibrio* virus genomes described. We provide general characterization and annotation of the viruses and describe read-mapping protocols to standardize genome presentation. The rich ecological and genomic contextualization of hosts and viruses make the Nahant Collection a unique platform for high-resolution studies of environmental virus-host infection networks.

Design Type(s)	time series design • biodiversity assessment objective
Measurement Type(s)	genome assembly
Technology Type(s)	DNA sequencing
Factor Type(s)	temporal_instant
Sample Characteristic(s)	dsDNA viruses, no RNA stage • Nahant • coastal sea water

[1]Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA 02141 USA. [2]Department of Systems and Computational Biology, Albert Einstein College of Medicine, Bronx, NY 10461 USA. *These authors contributed equally to this work. **These authors jointly supervised this work. †Present address: Bigelow Laboratory for Ocean Sciences, East Boothbay, ME 04544 USA. (J.M.B.); Bioresources and Environmental Biotechnology Laboratory, Department of Environmental Studies, University of Delhi, Delhi 110007, India (R.S.S.). Correspondence and requests for materials should be addressed to M.P. (email: mpolz@mit.edu) or to L.K. (email: libusha.kelly@einstein.yu.edu).

Background & Summary

Viruses influence the structure, function, ecology, and evolution of microbial communities. They represent the richest reservoir of nucleic acid diversity[1], and the great abundance of viral particles in the environment[1] reflects the expression of these sequences in host cells[2]. However, understanding the structure of virus-host interaction networks in the wild still poses a major challenge as pair-wise interactions between specific viruses and their hosts cannot be predicted without isolate based studies. Thus, though viruses have been predicted to play a major role in maintaining the extensive fine-scale genomic diversity of environmental microbes, it has not been possible to systematically evaluate the mechanistic or genomic foundations of these interactions, nor their ecological and evolutionary consequences.

Here, we present the annotated viral genomes of the Nahant Collection, a large-scale virus-host model system of cultivated and genome-sequenced bacterial and viral isolates, built on the extensively characterized environmental marine Vibrionaceae model system. By capturing large numbers of closely related host and virus strains, the Nahant Collection allows evaluation of the impact of ecologically relevant fine scale diversity on the interactions between bacteria and lytic viruses. This collection of 251 virus genomes and their associated hosts is a resource for interrogating the determinants of host range and the molecular bases of specific virus-host interactions within one of the most richly contextualized environmental microbial model systems[3–7] and time series studies[8] available.

Viruses and hosts were isolated from samples collected at three time points within a 93-consecutive-day study of littoral marine microbial communities, the Nahant Time Series[8]. All viruses were isolated on hosts collected on the same day, and hosts were nearly exclusively *Vibrio*. The *Vibrio* are a well-suited host group for the evaluation of the role of ecology and evolution in structuring virus-host interactions – they are ubiquitous in marine systems, ecologically diverse, and are among the most thoroughly characterized model systems for the study of bacterial populations in the wild[3–6,9,10]. The viruses were recovered using approaches designed to yield representation of diverse viruses, including: isolation from concentrates by plating in agar overlays to allow for more representative recovery of both fast- and slow-growing viruses; use of 2-week incubation times to allow for appearance of plaques by slow plaque-formers; and inclusion of additives to media to improve plaque visualization (glycerol[11]) and mimic environmental substrates (chitin) that might be necessary to induce expression of host receptors.

To standardize assemblies of purified viral isolate genomes we used an approach informed by predicted differences in packaging strategies among viruses, described in greater detail in the methods. This approach suggests that viruses of the Nahant Collection include members with diverse packaged genome types, including cohesive end overhangs, inverted terminal repeats, headful-packaging type terminal redundancy, and Mu-like host ends. To evaluate whether any of the viruses were prophages derived from the host of isolation, rather than environmentally-derived isolates, sequence-based searches between virus and isolation host genomes were performed; only one case of probable prophage purification was identified among the virus strains with sequenced host genomes (Supplementary Table 1).

The collection includes highly diverse dsDNA tailed and non-tailed viruses, including the smallest (10,046 bp) and largest (348,911 bp) described *Vibrio* virus genomes (median 45,072 bp). Using the Virfam[12] *Caudovirales* classifier, we find that the tailed viruses include representatives of all Virfam Types and Clusters previously identified as associated with Proteobacteria, including: Type 1 Clusters 3, 5, 6, 7, 8, 9 (*Siphoviridae* and *Myoviridae*); Type 2 (*Myoviridae*); and Type 3 (*Podoviridae*), as well as 26 viruses not associated to any previously identified Virfam Types or Clusters. Analyzing portal protein phylogeny revealed groups of closely related viruses as well as extensive collection-wide portal protein diversity (Fig. 1). The non-tailed viruses discovered in the collection are a proposed novel family, the *Autolykviridae*, and are discussed in greater detail elsewhere[13]. The overall collection ranges from 37% to 58% GC content, with a median of 43%. Viral genomes in the collection are also notable for their carriage of tRNAs, present in 53 viruses, and the presence of putative CRISPR features, present in 32 viruses (Table 1 (available online only)).

The viruses and hosts of the Nahant Collection, the largest available dataset of sequenced co-occurring cultivated virus-host pairs, are embedded within the rich contextualization of the 93-consecutive-day Nahant Time Series study[8]. The integration of ecological context, sequence-information, and cultivation-based study available for this model system make the Nahant Collection a unique and robust foundation for the study of the role of viruses in the ecology and evolution of their bacterial hosts.

Methods

Environmental sampling

All viruses and their hosts were isolated from water samples collected at three time points within a larger 3-month study[8] of coastal marine microbial communities at Canoe Cove, Nahant, MA, USA in 2010 (42° 25’ 10.6”N, 70° 54’ 24.2”W): August 10 (ordinal day 222, water temperature 13.8 °C), September 18 (261, 16.3 °C), and October 13 (286, 14.2 °C). Bacteria were collected using a size-fractionation approach[3,4] designed to partition co-occurring strains on the basis of differential associations in the water column. Here, as in previous studies of the *Vibrio*[3,4], bacteria were isolated by dilution series plating of material resuspended from 63 µm, 5 µm, 1 µm, and 0.2 µm size-fractions onto vibrio-selective media (MTCBS) for colony growth and serial purification. We purified 3,456 bacterial isolates comprised of 1,152 strains from

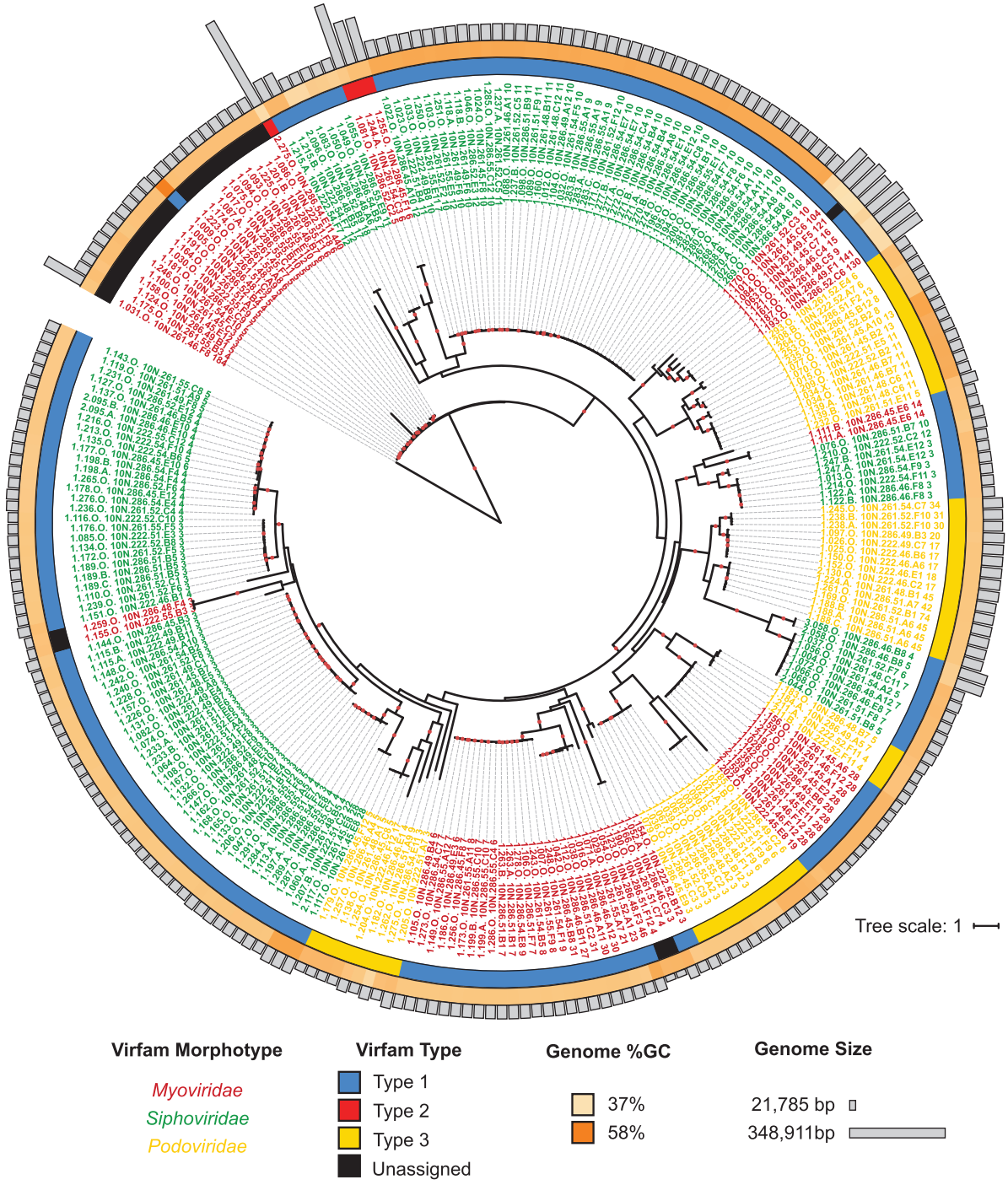

Figure 1. Overview of the diversity of Nahant Collection tailed viruses, organized by portal protein phylogeny. Virfam[12] classifier annotation of the Nahant Collection *Caudovirales* viruses reveals a diverse collection of myo-, sipho-, and podoviruses (indicated by color of leaf lable) representing all Types and Clusters (indicated in first attribute ring, and see Discussion for cluster identifiers) previously known to infect Proteobacteria, as well as many genomes unassignable to previously described groups. All 262 *Caudovirales* genome sequences are presented, including 28 replicate sub-lineage genomes. Genome %GC is provided in the second attribute ring and genome size is indicated by bars. The portal protein tree is unrooted and based on trimmed alignments; red circles indicate aLRT-supports ≥0.9. Associated data provided in Table 1, portal protein sequences in provided in Supplementary Table 3, interactive tree available at http://itol.embl.de/tree/1818971461811915195091 55#.

each of 3 days, evenly distributed over the size-fractions. Samples collected for isolation of viruses were 0.2 μm-filtered to remove bacteria, flocculated by addition of iron chloride, and the flocs collected on 0.2 μm filters and re-dissolved in oxalate for storage at 4˚C (ref. 14). Using this approach, viable viruses in 1000x-fold concentrated seawater could be preserved for later isolation on bacterial isolates derived from the same time and place.

Agar overlay direct plating of concentrates for isolation of viruses

To isolate viruses on co-occurring hosts we used a quantitative agar-overlay approach that allowed for equal representation of both slow- and fast-growing viruses as follows. Viral concentrates equivalent to 15 ml of seawater (15 μl iron-oxalate concentrate) were mixed with host cultures to form agar-overlays within which discrete plaques could form and from which viruses could be isolated[15]. In total 1,334 purified bacterial strains were exposed, comprising >400 strains per isolation day and representing all isolation size-fractions; of these, 295 showed plaques. Agar overlays were performed using 150 μl of host overnight culture, 2 ml of molten top agar (52 °C, 0.4% agar, 5% glycerol, in 2216 marine broth [MB]), and bottom agar containing glycerol and chitin (1% agar, 5% glycerol, 125 ml L^{-1} of chitin supplement [40 g L^{-1} coarsely ground chitin, autoclaved, 0.2 um filtered] in 2216 MB). Glycerol was added to increase the visibility of plaques[11], chitin was added to increase the probability of recovery of viruses dependent on chitin-induced receptors, and low density top agar was used to increase the probability of plaque formation by larger viruses[16]. Agar overlays were wrapped with plastic to reduce desiccation and held at room temperature for 14-16 days. Virus plaques were harvested at the end of the incubation period and archived by filtration of plaque eluates, as described in (ref. 15). Half of each eluate was stored at 4˚C, and half was preserved with glycerol (to a concentration of 25% glycerol) for storage at −20˚C.

Purification of viral strains

To build a diverse and representative collection of virus-host pairs, at least one randomly selected virus was purified from each bacterial strain for which plaques appeared in the agar overlay plating of environmental concentrate. To achieve this, we serially purified viruses recovered from archived material, prepared small-scale lysates to boost viral titer, and then generated high titer stocks by confluent lysis in agar overlays. Purification resulted in genome sequencing of 283 viral strains (from 251 independent plaques) from 246 hosts, described below. Viral and host strain naming conventions are described in Table 2, using examples of virus 1.008.O_10N.286.54.E5 and host 10N.286.54.E5.

Genome sequencing

Viral genomes were prepared from lysates of the host of isolation, as follows. Lysates were concentrated on centrifugal filtration devices (Ultracel 30 K, Amicon Ultra, Millipore, UFC903024), washed with 1:100 2216MB, and concentrates treated with nucleases to digest unencapsidated nucleic acids (18 ml sample brought to 500 μl and amended with DNase I, RNase A, heated for 65 min at 37˚C). Nuclease-treated samples were extracted by addition of 0.1 final volume of SDS mix (0.25 M EDTA; 0.5 M Tris-HCl, pH 9.0; 2.5% sodium dodecyl sulfate), 30 min incubation at 65˚C, addition of 0.125 volumes 8 M potassium acetate, 60 min incubation on ice, addition of 0.5 volumes of phenol-chloroform, and recovery of nucleic acids from aqueous phase by isopropanol and ethanol precipitation. Illumina sequencing libraries of each extract were prepared as follows. Sample DNA (5 μg in 100 μl) was sheared by sonication (6 cycles of 5 min each at an interval of 30 sec on/off on the 'Low Intensity' setting of the Biogenode Bioruptor) to enrich for fragment sizes of ~300 bp. Sequencing constructs were prepared by end repair of sheared DNA, 0.72x/0.21x dSPRI size selection to enrich for ~300 bp sized fragments, ligation of Illumina adapters and unique pairs of forward and reverse barcodes for each sample, SPRI bead clean-up, nick translation, and final SPRI bead clean-up[17]. Constructs were enriched by PCR using paired-end (PE) primers following qPCR-based normalization of template concentrations. Enrichment PCRs were prepared in eight replicate 25 μl volumes, with the recipe: 1 μl Illumina construct template, 5 μl 5x Phusion polymerase buffer, 0.5 μl 10 mM dNTPs, 0.25 μl 40 μm IGA-PCR-PE-F primer, 0.25 μl 40 μm IGA-PCR-PE-R primer, 0.25 μl Phusion polymerase, 17.75 μl PCR-grade water. PCR thermocycling conditions were as follows: initial denaturation at 98 °C for 20 sec; batch dependent number of cycles of 98 °C for 15 sec, 60 °C for 20 sec, 72 °C for 20 sec; final annealing at 72 °C for 5 min; hold at 10 °C. For each sample 8 replicate enrichment PCR reactions were pooled and purified by 0.8x SPRI bead clean-up. Each sample was then checked by Bioanalyzer (2100 expert High Sensitivity DNA Assay) to confirm the presence of a unimodal distribution of fragments with a peak between 350-500 bp. Sequencing of viral genomes was distributed over 4 paired-end sequencing runs as follows: 1 lane on the Illumina HiSeq2000 (18 viral genomes; 100 +100 nt paired-end reads; average of 5.1 million reads per genome), 3 lanes on the Illumina MiSeq (92-96 genomes per lane; 150+150 nt paired-end reads; average of 54 K, 208 K, 210 K reads per genome for each lane). Raw paired-end Illumina reads were imported and demultiplexed using CLC Genomics Workbench v.6.5.1 (https://www.qiagenbioinformatics.com/). Sequencing and assembly of genomes of bacterial hosts is described elsewhere[13].

Genome assembly and curation

Differences in packaging strategies among viruses yield distinctive and characteristic distributions of packaged physical genomes in progeny virions[18]. Common examples of such strategies include

production of virions with genomes that have: variable termini comprised of host DNA (Mu-like viruses); 5′ or 3′ single strand terminal overhangs (cos-viruses); or different start sequences and terminal redundancies ranging from 10 s to 10,000 s of bases (pac-viruses). To inform final curation of genome sequences, we first performed initial assemblies to group similar genomes and allow identification of the packaging-associated large subunit terminase gene (TerL) where possible. We then evaluated read mapping profiles within groups, considering terminase-predicted packaging strategy, to define final genome start sites. We next used an iterative approach, as described below, to standardize genome assemblies with conserved gene orders and genomic start positions for related viruses, and to place genomic termini at the contig ends.

Initial assembly and viral genome clustering

Initial assembly and clustering of viral genomes identified groups of related viruses (Supplementary Table 2), but also highlighted the need for systematic measures to standardize genome curation. Initial assembly and clustering were performed as follows: viral genomes were assembled using the *de novo* assembly tool in CLC Genomics WorkBench v.6.5.1 with default parameters following trimming of reads (default parameters except: quality score = 0.01, ambiguous nucleotides = 0). Open reading frames (ORFs) were identified using Prodigal[19] with default parameters, and reciprocal best BLAST hits with ≥75% coverage of the longer sequence and e-value of $\leq 10^{-5}$ were clustered using OrthoMCL[20]. Viral genomes were clustered into genome groups on the basis of shared protein clusters using the FT algorithm of the ClustnSee[21] plug-in in Cytoscape[22]. Preliminary curation of individual groups to assess synteny between closely related and replicate viruses (see Technical Validation) using LAST[23] indicated that assemblies began at different locations, suggesting that virus genome characteristics were confounding consistency in contig start and end sites.

Final assembly and curation

To systematically address the inconsistency in contigs produced by assemblies of closely related viruses, assemblies were repeated and curated based on read mapping patterns and terminase similarities, as described below.

Viral genomes were re-assembled using the command clc_assembler from CLC Assembly Cell (version 4.4.2, https://www.qiagenbioinformatics.com/), using default assembly parameters and an insert size setting of 100 to 300 bp; 154 out of 285 assemblies resulted in one contiguous sequence (contig). For virus assemblies producing more than one contig, the highest coverage contig was extracted and considered the target viral genome contig; lower coverage contigs were considered contamination from host genome or prophages.

Viral genome open reading frames (ORFs) were identified using Prodigal version 2.6.1 with the -p meta flag to identify small ORFs[24], and virus terminase protein sequences were identified by UBLAST[25] search with a cutoff evalue < 0.001 against a database of terminases from public viral genomes with previously described or predicted physical genomic termini[18,25]. Terminase identity was initially assessed via UBLAST as described above, and then verified via OrthoMCL clustering of terminase ORFs with the same dataset to gauge the fidelity of the BLAST results. To evaluate read coverage patterns in relation to terminases, original reads were mapped back to the contigs using the clc_mapper command with default parameters and per-base coverage was determined using SAMtools[26] and BEDtools[27]. Consistent with previous findings that different terminases are associated with distinct genome packaging strategies[18], and thus genome termini, exploratory evaluation of read coverage patterns showed that viruses with close identity to different known phage terminases also generally showed different read coverage patterns (Fig. 2a–d).

To standardize final gene order presentations, start and stop positions for each genome were defined manually, considering three criteria: 1) terminase identity, as identified by UBLAST and OrthoMCL clustering; 2) read coverage; and 3) comparison of contigs between viruses within the genome groups identified in the initial assembly. All members of each group were assigned to a common inferred genome packaging strategy category (see Supplementary Table 2 for details and exceptions) on the basis of overall patterns within the group, and rearranged using the approaches described below. Coverage patterns were determined by visual inspection and by a series of custom R scripts. Where possible, finalized virus genomes were quality checked by comparing the synteny of related phage genomes using command line LAST. A total of 283 virus genomes were assembled, including 251 unique viruses, 31 sub-lineages purified in parallel to the primary unique isolate, and 1 technical replicate (Table 1 (available online only)). We note that though we were guided by group-level coverage patterns, our primary aim was standardization rather than inference of true genome topology, which must be defined by individual genome read coverage patterns and complemented by laboratory studies.

Re-arrangement based on specific ORF. The majority of viral genomes (168/283) in the collection were standardized by circularizing the *de novo* assembled contigs and re-linearizing them at the start of the ORF upstream of the terminase. As a whole, these viruses showed coverage patterns consistent with a headful, or 'pac' site, based genome packaging strategy, wherein up to 110% genome-length monomers are sequentially cleaved from a multigenome-length concatemer, beginning from a conserved 'pac' site. Terminase best BLAST matches were dominated by similarity to viruses with headful-like strategies (Sf6,

97 best hits; 933W, 12; and T4, 5), though best hits to short direct terminal repeat (T7, 8) and 5′-cohesive ends (P2, 26) viruses were also identified, along with cases of no similarity to reference virus terminases (20). Read coverage patterns among these viruses were dominated by either a pattern of gradual decreases/shifts (112) consistent with a 'headful' or packaging site ('pac') – based genome packaging strategy, or even coverage (46); though other patterns of coverage including short peaks (cos pattern, 1; short internal peak, 7) and multiple coverage peaks (2) were also observed. Examination of read coverage following TerL-based rearrangement of contigs (Fig. 2e) often showed coverage maxima localized near the start, consistent with previous observations that headful-packaging viruses commonly have a pac site in or near the small subunit of the terminase gene[18], which generally lies upstream of the TerL. Viruses curated using this approach included all the viruses from 7 of the preliminary groups (1, 4, 6, 9, 10, 13, 16), the majority of viruses from group 3, and a single virus from group 7.

Re-arrangement based on peaks or valleys in coverage. The second most commonly applied strategy for standardization (66/283) was circularization of contigs followed by re-linearization by cutting in the middle of a short region of either aberrantly low (36), or high (30), coverage (Fig. 2f,g). As a whole, these viruses showed patterns consistent with the presence of either direct terminal repeats (DTRs) or single-stranded cohesive ('cos') ends associated with their genome termini. Genomes with DTRs may yield sharply defined regions of elevated coverage. 'cos' genomes may yield regions of either high or low coverage, depending on whether they are 3′ or 5′ overhangs, due to low frequency ligation of ends during library preparation, as well as T4 DNA polymerase 3′ to 5′ exonuclease activity (degradation of 3′ overhangs) and 5′ to 3′ polymerase activity (endfill of 5′ overhangs) of unligated ends. Terminase best BLAST matches were dominated by similarity to viruses with cohesive ends (lambda 'cos', 22; HK97 'cos'-3′, 7; P2 'cos'-5′, 3) and DTRs (N4, 8; T7, 13), though best hits to headful viruses (933W, 4; P22, 1), were also identified, along with cases of no similarity to reference virus terminases (8). Read coverage patterns among these viruses were dominated by either distinctive 'cos' (32) or short internal peak (22) patterns, though other patterns of coverage including shifts in coverage (8), multiple coverage peaks (2), even coverage (1), or no pattern (1) were also observed. This approach included all virus genomes in preliminary groups 8, 14, and 18; the majority of viruses in groups 5, 7, 11, and 12; and a minority of viruses in groups 3 and 17.

Scaffolded assembly against reference. If viruses did not follow the patterns described above, but closely related members of the same group (identified as sharing 100% of translated proteins identified via reciprocal UBLAST) did follow a particular pattern, viruses were assembled using closely related strains as a scaffold; this approach was used for genomes in group 5 (3).

Maintenance of original de novo assembly. Singleton viruses with no similar members within the dataset were treated based on closest terminase identity and read coverage pattern, but if no distinct pattern was observed the original assemblies were maintained. This approach was used for 16/283 viruses, including viruses in groups 2 (8), 3 (1), 7 (5), 11 (1), and 12 (1).

Removal of terminal unconserved sequences. A subset of viruses (9/283) were found by BLAST comparison to possess Mu-like terminases, suggesting that they also used a Mu-like replicative transposition headful mechanism that incorporates host DNA upstream and downstream of the site of insertion. Read coverage patterns of initial assemblies of Mu-like viruses exhibited sharp drops in coverage at the termini followed by regions of low coverage (Fig. 2d), these regions of low coverage, representing small pieces of the host genome, were removed in the adjusted assemblies (Fig. 2h). However, closer evaluation of these assemblies revealed several cases of truncated sequences and final Mu-like virus assemblies were performed in CLC Genomics Workbench 8.5.1 as follows. Sequences were trimmed using the NGS Core Tools Trim Sequences tool with trims based on quality scores (limit

Name component	Specific Description
1	A unique identifier for each independent plaque isolated from a given host from the initial exposure of a given host to an environmental virus concentrate.
008	A unique working ID for a host strain.
O	A lineage generated from a single plaque during viral serial purification, for example due to the emergence of multiple plaque morphologies. Options: Single lineage: O; Sub-lineages: A, B, C, etc….
10N	Year & site identifier (2010, Nahant).
286	Ordinal day.
54	A code representing the size-fraction of origin. Options: 0.2um: 45,46,47; 1um: 48,49,50; 5um: 51,52,53; 63um: 54,55,56. Note: Multiple codes within the size-fraction identifier reflect independent water samples for the 63um fraction, and independent water sample fractionation series for the other size classes (water sample A: 45,51,54; sample B: 46, 52, 55; sample C: 47, 53, 56).
E5	Unique storage well identifier.

Table 2. Strain identifier nomenclature, using example virus 1.008.O_10 N.286.54.E5 and host 10N.286.54.E5.

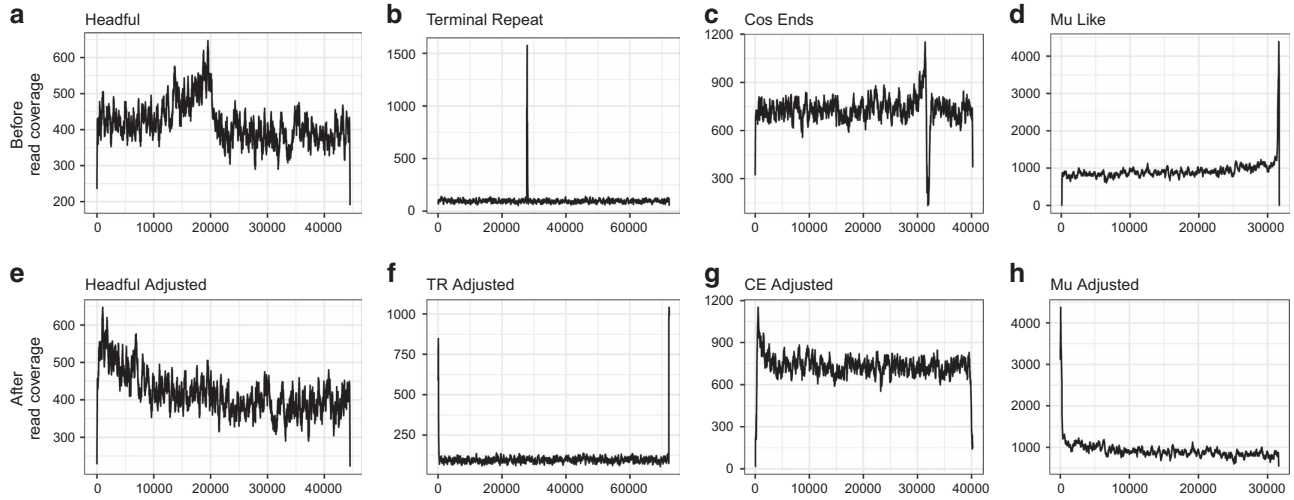

Figure 2. Examples of read recruitment by contig assemblies before and after adjustment for virus genomes with differing read mapping patterns. Coverage mapping onto contigs of viruses with: Headful-like read mapping before (**a**) and after contig adjustment (**e**); Terminal repeat-like read mapping before (**b**) and after (**f**) contig adjustment; Single-stranded cos-end-like read mapping before (**c**) and after (**g**) contig adjustment; Mu-like read mapping before (**d**) and after (**h**) contig adjustment. Note that, as indicated in the methods, though read mapping patterns were evaluated for each virus, final adjustment strategy for each virus (Supplementary Table 2) was not determined solely based on read mapping pattern and the majority of virus contigs were defined as starting one open reading frame upstream of the large subunit of the terminase (TerL) regardless of the read mapping pattern.

0.0001), number of allowable ambiguous nucleotides (max 0), and discard of reads <50 bases. All remaining read pairs and orphans were assembled using the De Novo Assembly tool with a word size of 64 and otherwise default options. The largest contig was extracted from the assembly for each virus and all genomes were aligned using the Geneious 6.0.6 Map to Reference tool to standardize orientation. Genome termini were defined based on the beginning and end of conserved regions at the left and right genome ends, respectively. This yielded 9 independently isolated genomes that were all 100% nucleotide identical and with a length of 31,617 bases, with the exception of virus 1.159.O, which contained a single SNP that was present in both the new and previous assembly versions. Open reading frames for these genomes were called with Prodigal 2.6.3 using the -p meta flag and otherwise default parameters.

Iterative assembly. A subset of the viruses (21/283), described elsewhere as a new family[13], had distinctively short (~10 kb) genomes and did not contain predicted terminases. BLAST comparison of ORFs from these genomes showed similarity to the protein-primed DNA polymerase of viruses of the *Tectiviridae*, which have linear genomes and inverted terminal repeats (ITR), and thus these viruses were also evaluated for ITRs. Final assemblies for this group were performed iteratively, as follows. Following initial assembly, second and third assembly iterations were performed using an increased word size of 64, and the largest contig from the previous assembly was included as one of the "reads" for the successive round of assembly. The longest contigs from each of the three assemblies were then compared and the longest contig that also exhibited ITRs was used, when none of the contigs contained ITRs the longest assembled contig was determined to be the final assembly.

Annotation

Viral genomes were annotated using multiple approaches and tools, as described below. Genomes are available through Genbank (Data Citation 1).

Virfam classification of viral Types and Clusters and morphotypes. Viral proteins were annotated using the Virfam[12] classifier, which identifies multiple genes of the head-neck-tail modules of viral genomes and assigns viral genomes to morphotypes within Types and Clusters on the basis of previous characterization of diverse tailed viruses. Annotation was performed individually per genome by submission to the Virfam webserver (http://biodev.cea.fr/virfam/). Output reports for all Virfam annotation runs are available through figshare (Data Citation 2).

Genome content annotation. Phage proteomes were compared to KEGG[28], COG[29], eggnog[30], Pfam[31], ACLAME[32], CAMERA Viral Proteins (CVP)[33] and the OM-RGC collection of sequences[34] via

Phage	Query	Query Length	Subject	Subject Length	Query Mismatch (bp)	Subject Mismatch (bp)	ANI	SNP (bp)	inDEL (bp)	Start Variation (bp)	Length Variation (bp)	MAFFT Alignment # of diffs
1.021	A	43,743	B	43,743	1	1	0.9999771	0	0	0	0	0
1.021	A	43,743	C	43,743	2	2	0.9999543	1	0	0	0	0
1.021	B	43,743	C	43,743	2	2	0.9999543	0	0	0	0	0
1.107	B	10,447	C	10,447	1	0	0.9999521	0	0	0	0	0
1.107	B	10,447	A	10,447	3	2	0.9997607	2	0	0	0	2
1.107	C	10,447	A	10,447	3	2	0.9997607	2	0	0	0	2
1.111	A	40,209	B	40,209	10	9	0.9997637	0	0	9	0	0
1.115	B	37,416	A	37,416	2	3	0.9999332	1	0	1	0	0
1.118	B	60,458	A	60,458	5	5	0.9999173	4	0	0	0	4
1.122	A	44,523	B	44,523	3	4	0.9999214	2	0	1	0	2
1.139	B	44,094	A	43,893	6	1	0.9999204	0	201	48	201	201
1.188	A	72,305	B	72,305	2	3	0.9999654	0	0	2	0	0
1.188	A	72,305	C	72,305	2	3	0.9999654	0	0	2	0	0
1.188	B	72,305	C	72,305	1	1	0.9999862	0	0	0	0	0
1.189	B	36,855	O	36,855	4	3	0.9999050	2	0	1	0	2
1.189	B	36,855	C	36,855	4	4	0.9998915	3	0	0	0	3
1.189	O	36,855	C	36,855	2	3	0.9999322	1	0	1	0	1
1.198	B	44,343	A	44,472	2	43	0.9994933	1	129	0	129	130
1.199	B	48,312	A	48,312	4	4	0.9999172	3	0	0	0	3
1.211	A	37,169	B	37,169	1	1	0.9999731	0	0	0	0	0
1.215	B	80,834	A	80,834	1	2	0.9999814	0	0	1	0	0
1.233	B	36,823	A	36,823	2	1	0.9999593	0	0	1	0	0
1.237	B	60,160	A	60,097	64	3	0.9994429	0	63	1	63	64
1.238	A	70,467	B	70,494	19	20	0.9997233	0	27	19	27	27
1.247	B	43,896	A	43,896	3	2	0.9999430	1	0	1	0	1
1.249	B	10,611	A	10,611	1	0	0.9999529	0	0	0	0	0
1.263	B	49,640	A	49,640	11	12	0.9997683	2	0	9	0	0
1.268	A	59,297	B	59,297	3	3	0.9999494	2	0	0	0	2
1.270	B	59,297	A	59,297	3	2	0.9999578	1	0	1	0	1
1.271	B	59,297	A	59,297	3	3	0.9999494	2	0	0	0	2
1.277	A	59,419	B	59,297	3	21	0.9997978	1	122	1	122	123
1.283	A	59,530	C	59,530	3	2	0.9999580	1	0	0	0	1
1.283	A	59,530	B	59,530	1	1	0.9999832	0	0	0	0	0
1.283	C	59,530	B	59,530	2	3	0.9999580	1	0	0	0	1
2.095	B	44,649	A	44,649	3	4	0.9999216	2	0	1	0	2
2.159	A	31,617	B	31,617	0	0	1.0000000	0	0	0	0	0

Table 3. Replicate virus genome comparisons.

UBLAST[24]. Annotations were determined as the best hit (maximum bit score) to a non-hypothetical protein from EggNOG, KEGG, COG, ACLAME or Pfam (minimum alignment of 75%, minimum percent identity of 35%). Best hits to remaining databases as well as CVP and OM-RGC are reported as notes within the final annotations. Annotations were combined with annotations identified using InterProScan version 5.17-56.0 using the iprlookup, goterms, and pathways options. InterProScan is a program from EMBL-EBI that uses the InterPro database for annotations. The InterPro database contains by default 13 databases, which are listed here: https://github.com/ebi-pf-team/interproscan/wiki/HowToRun#included-analyses. For these annotations, two optional databases were included: TMHMM for predicted transmembrane proteins and SignalP for predicted signal peptide cleavage sites. tRNA sequences were identified using tRNAscan-SE version 1.23 (ref. 35) using the general tRNA model (-G). CRISPR-like elements were identified using CRT[36].

Portal protein phylogeny

Portal proteins were identified directly using the Virfam classifier as described above, which provides a portal prediction, as well as by using HMM- and blastp-based searches of all Nahant Collection virus proteins, as follows. The portal protein for the representative virus of each Virfam cluster was downloaded through the Virfam page (http://biodev.cea.fr/virfam/), and an HMM generated by performing 3 iterations of Jackhmmer[37] https://www.ebi.ac.uk/Tools/hmmer/search/jackhmmer. Searches of the Nahant Collection virus proteins with this collection of HMMs using the hmmsearch[38] tool (hmmer version 3.1b2) identified putative portal proteins in 241 genomes, these 241 together with 6 portal proteins identified directly through the Virfam web page, were used to search the Nahant Collection with blastp[39], identifying putative portal proteins in 262 of the 263 *Caudovirales* (e value < 0.0001). In the 2 cases where the predicted portal proteins differed across the two methods (Supplementary Table 3), HHpred as implemented in the MPI bioinformatics Toolkit[40] (https://toolkit.tuebingen.mpg.de/#/tools/hhpred) was used to evaluate both predictions and the protein with the longer sequence similarity to a portal protein was selected. The portal protein in virus 1.031.O could only be predicted using the Virfam approach and this protein was included, though HHpred and Phyre2 (ref. 41) based structural similarity based searches did not indicate similarity to known portal proteins. Sequence alignment, trimming, and tree-building were performed using the eggnog41 workflow in the ETE3 (ref. 42) version 3.0.0b36 tree building tool, the tree was visualized using iTOL[43], and the figure prepared using Adobe Illustrator.

Data Records

All virus and host-associated sequences and annotations associated with this work have been deposited to the Nahant Collection NCBI BioProject (Data Citation 1), specific accession numbers for each strain are provided in Supplementary Table 1. Viral genome annotation reports generated by the Virfam tool have been deposited with figshare (Data Citation 2).

Technical Validation

Given the known abundance of prophages in bacterial genomes we evaluated whether any viruses in the collection represented induced prophages from the host of isolation. Using megaBLAST in Geneious 6.1.8 we searched all virus genomes against all sequenced hosts, we identified only a single case of a high query cover and high identity match. The virus 1.202.O (32,014 bp) shared a 30,051 bp 100% identity match with its host 10N.222.45.E8; this match region occurred within a larger host contig of 120,557 bp, suggesting that the failure to achieve a full match with the remaining 1,963 bp region of the virus genome contig was not due to incomplete assembly of the associated host region. Full genomes for the host of isolation were not available for 29 viruses and thus this prophage derivation could not be assessed for these strains, information on host sequence availability is provided in Supplementary Table 1.

This dataset contained sets of virus pairs and triplets that served as biological replicates for assembly optimization. Such sets derive from instances of independent purification of viral sub-lineages from a parent plaque due to the occurrence of variable plaque morphology. Though they exhibited sites of polymorphism at the nucleotide level, ranging from 0 to 4 SNPs, and indels of up to 201 bp (Table 3), members of these sets consistently showed identical gene content and are expected to have the same genomic structure and gene order. Methods for rearrangement to maintain synteny were developed around such sets and were verified via alignment of similar/duplicate genomes before and after rearrangement (Supplementary Figures 1–4).

References

1. Suttle, C. A. Marine viruses–major players in the global ecosystem. *Nat. Rev. Microbiol.* **5,** 801–812 (2007).
2. Forterre, P. The virocell concept and environmental microbiology. *ISME J.* **7,** 233–236 (2013).
3. Hunt, D. E. *et al.* Resource Partitioning and Sympatric Differentiation among Closely Related Bacterioplankton. *Science* **320,** 1081–1085 (2008).
4. Szabo, G. *et al.* Reproducibility of Vibrionaceae population structure in coastal bacterioplankton. *ISME J.* **7,** 509–519 (2013).
5. Shapiro, B. J. *et al.* Population Genomics of Early Events in the Ecological Differentiation of Bacteria. *Science* **336,** 48–51 (2012).
6. Hehemann, J.-H. *et al.* Adaptive radiation by waves of gene transfer leads to fine-scale resource partitioning in marine microbes. *Nat. Commun* **7,** ncomms12860 (2016).
7. Preheim, S. P. *et al.* Metapopulation structure of Vibrionaceae among coastal marine invertebrates. *Environ. Microbiol.* **13,** 265–275 (2011).
8. Martin-Platero, A. M. *et al.* High resolution time series reveals cohesive but short-lived communities in coastal plankton. *Nat. Commun* **9,** ncomms266 (2018).
9. Takemura, A. F., Chien, D. M. & Polz, M. F. Associations and dynamics of Vibrionaceae in the environment, from the genus to the population level. *Front Microbiol* **5,** 38 (2014).
10. Takemura, A. F. *et al.* Natural resource landscapes of a marine bacterium reveal distinct fitness-determining genes across the genome. *Environ. Microbiol.* **19,** 2422–2433 (2017).
11. Santos, S. B. *et al.* The use of antibiotics to improve phage detection and enumeration by the double-layer agar technique. *BMC Microbiol.* **9,** 148 (2009).
12. Lopes, A., Tavares, P., Petit, M.-A., Guérois, R. & Zinn-Justin, S. Automated classification of tailed bacteriophages according to their neck organization. *BMC Genomics* **15** (2014).
13. Kauffman, K. M. *et al.* A major lineage of nontailed dsDNA viruses as unrecognized killers of marine bacteria. *Nature* **554,** 118–122 (2018).

14. John, S. G. *et al.* A simple and efficient method for concentration of ocean viruses by chemical flocculation. *Environ. Microbiol. Rep* **3,** 195–202 (2011).
15. Kauffman, K. M., Polz, M. F. Streamlining standard bacteriophage methods for higher throughput. *MethodsX* **5,** 159–172 (2018).
16. Serwer, P., Hayes, S. J., Thomas, J. A. & Hardies, S. C. Propagating the missing bacteriophages: a large bacteriophage in a new class. *Virol J* **4,** 21 (2007).
17. Rodrigue, S. *et al.* Unlocking Short Read Sequencing for Metagenomics. *PLoS ONE* **5,** e11840 (2010).
18. Casjens, S. R. & Gilcrease, E. B. Determining DNA packaging strategy by analysis of the termini of the chromosomes in tailed-bacteriophage virions. *Methods Mol. Biol.* **502,** 91–111 (2009).
19. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11,** 119 (2010).
20. Li, L., Stoeckert, C. J. & Roos, D. S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* **13,** 2178–2189 (2003).
21. Spinelli, L. *et al.* Clust&See: A Cytoscape plugin for the identification, visualization and manipulation of network clusters. *BioSystems* **113,** 91–95 (2013).
22. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13,** 2498–2504 (2003).
23. Kielbasa, S. M., Wan, R., Sato, K., Horton, P. & Frith, M. C. Adaptive seeds tame genomic sequence comparison. *Genome Res.* **21,** 487–493 (2011).
24. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26,** 2460–2461 (2010).
25. Casjens, S. R. *et al.* The Generalized Transducing Salmonella Bacteriophage ES18: Complete Genome Sequence and DNA Packaging Strategy. *J. Bacteriol.* **187,** 1091–1104 (2005).
26. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
27. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–842 (2010).
28. Kanehisa, M. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* **28,** 27–30 (2000).
29. Tatusov, R. L. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.* **28,** 33–36 (2000).
30. Powell, S. *et al.* eggNOG v3.0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges. *Nucleic Acids Res.* **40,** D284–D289 (2012).
31. Bateman, A. The Pfam protein families database. *Nucleic Acids Res.* **32,** D138–D141 (2004).
32. Leplae, R., Hebrant, A., Wodak, S. & Toussaint, A. ACLAME: A CLAssification of Mobile genetic Elements. *Nucleic Acids Res.* **32,** D45–D49 (2004).
33. Sun, S. *et al.* Community cyberinfrastructure for Advanced Microbial Ecology Research and Analysis: the CAMERA resource. *Nucleic Acids Res.* **39,** D546–D551 (2011).
34. Sunagawa, S. *et al.* Structure and function of the global ocean microbiome. *Science* **348,** 1261359–1261359 (2015).
35. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25,** 955–964 (1997).
36. Bland, C. *et al.* CRISPR Recognition Tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinformatics* **8,** 209 (2007).
37. Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* **39,** W29–W37 (2011).
38. Eddy, S. R. Accelerated Profile HMM Searches. *PLOS Comput Biol* **7,** e1002195 (2011).
39. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410 (1990).
40. Alva, V., Nam, S.-Z., Söding, J. & Lupas, A. N. The MPI bioinformatics Toolkit as an integrative platform for advanced protein sequence and structure analysis. *Nucleic Acids Res.* **44,** W410–W415 (2016).
41. Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N. & Sternberg, M. J. E. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protoc.* **10,** 845–858 (2015).
42. Huerta-Cepas, J., Serra, F. & Bork, P. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Mol. Biol. Evol.* **33,** 1635–1638 (2016).
43. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics* **23,** 127–128 (2007).

Data Citations

1. *GenBank* MG592390-MG592672 (2016).
2. Kauffman, K. M. *et al. figshare* https://dx.doi.org/10.6084/m9.figshare.c.4028239 (2018).

Acknowledgements

We thank members of the Polz lab for assistance with sampling and strain isolation, especially Michael Cutler, Tara Soni, Alison Takemura, Gitta Szabo, Hong Xue, Otto Cordero, Nisha Vahora, Aidong Ruan, and Hans Wildschutte, as well as Robert Ratzlaff (ODU). We also thank Simon Labrie (U. Laval) for guidance in viral genome extractions and sequencing library preparation protocols; Raphael Guerois (CEA - IBITECS) for assistance with VirFam; and Will Chang (Einstein) for assistance with genome curation. This work was supported by grants from the National Science Foundation OCE 1435993 and 1435868 to M.P. and L.K., respectively, and the WHOI Ocean Ventures Fund to K.M.K.

Author Contributions

K.K. and M.P. conceived the sampling study, K.K., J.B., and L.K. conceived the genome assembly strategies. K.K. and J.B. wrote the manuscript with contributions from M.P. and L.K. K.K. performed the sampling, laboratory work, initial viral genome assemblies and genome grouping, and Virfam analyses. R.S. contributed to viral genome extraction and library preparation. D.V. contributed to initial genome grouping analyses. J.B. performed the final viral genome coverage analyses and assemblies, and viral genome annotations. J.E. contributed to viral genome annotation. All authors read, revised, and approved the manuscript.

Additional information

Table 1 is only available in the online version of this paper.

Supplementary information accompanies this paper at http://www.nature.com/sdata

Competing interests: The authors declare no competing interests.

How to cite this article: Kauffman, K. M. *et al.* Viruses of the Nahant Collection, characterization of 251 marine Vibrionaceae viruses. *Sci. Data* 5:180114 doi: 10.1038/sdata.2018.114 (2018).

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.