Supplemental material for: Assembly-free single-molecule sequencing recovers complete virus genomes from natural microbial communities

John Beaulaurier, Elaine Luo, John M. Eppley, Paul Den Uyl, Xiaoguang Dai, Andrew Burger, Daniel J Turner, Matthew Pendelton, Sissel Juul, Eoghan Harrington and Edward F. DeLong

Contents:

Supplemental Results

Detection of circularly permuted genomes in uvMED mock viral metagenome

The genome AP013502 is circularly permuted from AP013504 by 1 kb (2.7% of their lengths) and AP013468 is circularly permuted from AP013467 by only 300 bp (0.9% of their lengths). This minor difference resulted in the simulated reads from the similar genomes being merged into the same alignment cluster for polishing. We were, however, able to separately cluster reads simulated from the genomes AP013483 and AP013484, which are circularly permuted from each other by 1.35 kb (4.4% of their length).

Supplemental Methods

Sample collection and DNA purification from virus-enriched tangential flow filtration retentates

To collect and concentrate viral particles, seawater samples were collected at three different depths and prefiltered to remove bacterioplankton cells. The resulting virus-enriched filtrate was concentrated via tangential flow filtration (TFF) before DNA extraction. The 25 m deep sample was collected on the HOT-314 cruise on August 5, 2019 at Station ALOHA (22°45' N, 158° W; http://hahana.soest.hawaii.edu/hot/). Seawater samples from depths of 117 m (22° 1841 N, 157° 05.57 W on April 2, 2018) and 250 m (24.33.5438 N, 160.50.6582 W on April 8,2018) were both collected on the the Falkor Eddy cruise oceanographic expedition conducted in 2018 (http://scope.soest.hawaii.edu/data/falkor2018/falkor2018.html). For each sample, 90 - 110 L of seawater was collected using a Niskin bottle rosette attached to a conductivity-temperature-depth (CTD) package. The seawater was pre-filtered by peristatic pumping through either a 0.22 um filter (Sterivex GV) for the 25 m sample, or a 0.1 uM Supor cartridge filter (Acropak 500, Pall, USA) for the 117 m and 250 m samples. The resulting virus-enriched filtrate was concentrated by tangential flow filtration (TFF) over a 30 kDa filter (Biomax 30 kDa membrane, catalogue #: P3B030D01, Millipore). Subsequently, the retentate was reduced to a volume of ~200 mL, the tangential flow filter was backflushed with 100 mL of permeate to release virus particles trapped in the filter, and the concentrated viruses were recovered in a final retentate volume of ~300 mL. The virus-containing retentates were stored at 4°C until final processing and DNA extraction.

Immediately prior to DNA extraction, the ~300 mL retentates were concentrated using a Millipore Centricon Plus-70 Centrifugal Filter Units (10 KDa) (catalog: UFC701008, MilliporeSigma) following the manufacturers recommended protocol, resulting in a volume of 2-3 mL of virus-enriched retentate. A second round of centrifugal concentration of the resulting concentrated retentate was performed on a 100K Microsep Advanced centrifugal concentration device fitted with an Omega Advance membrane (modified polyethersulfone; catalogue #MCP 100C41, Pall, USA) following the manufacturers recommended protocol, resulting in a final volume of 200 uL of concentrated virus-enriched retentate, which was used for final DNA purifications. Lysis and DNA purification were performed in a single tube using the Qiagen Genomic-tip 20/G protocol following manufacturer's recommendations. First, an RNase A solution (200 ug/mL) was prepared by adding 20 ul of 10 mg/ml RNase A to 1 ml of Buffer B1 (50mM Tris•HCl, pH 8; 50mM EDTA, pH 8; 0.5% Tween-20; 0.5% Triton X-100). Next, 1 mL of the Buffer B1 lysis buffer containing RNase A was added to 200 uL of viral concentrate. Then, 20 uL of a 100 mg/ml stock solution of lysozyme prepared in sterile water and 45 uL of a Proteinase K solution (20 mg/mL) prepared in sterile water were added, followed by gentle mixing and incubation at 37°C for 30 minutes. Next, 350 uL of Buffer B2 lysis buffer (3M guanidine HCl; 20% Tween 20) was added and incubated at 50°C for 30 minutes.

The lysates were loaded onto single Qiagen Genomic-tip 20/G column and purification was performed following manufacturers recommendations (Qiagen, Hildern, Germany). Each Genomic-tip 20/G column was equilibrated with 1 mL of Buffer QBT equilibration buffer (750 mM NaCl; 50 mM MOPS pH 7.0; 15% isopropanol; 0.15% Triton X-100) by gravity flow. Sample lysates were then mixed by inverting several times and carefully pipetted sequentially onto one equilibrated Genomic-tip 20/G column, allowing samples to enter the resin by gravity flow. Next, 1 mL Buffer B1 was combined with 350 uL Buffer B2, and all sample lysate tubes were rinsed carefully with this 1 mL solution, and the rinse solution applied to the same Genomic-tip 20/G column. The Genomic-tip 20/G column was washed by gravity flow by applying 1 mL of Buffer QC wash buffer (1.0 M NaCl; 50 mM MOPS, pH 7; 15% isopropanol) three times, in succession. Finally, the genomic DNA was eluted from the column by two successive applications of 1mL Buffer QF elution buffer (1.25 M NaCl; 50 mM Tris•HCl, pH 8.5; 15% isopropanol), resulting in a purified DNA preparation in a 2 mL final volume.

The column purified DNA was concentrated by isopropanol precipitation as follows: The DNA eluant was split into two 2 mL conical screw cap tubes, 1mL per tube. The DNA was precipitated by adding 0.7 mL of room-temperature isopropanol per each 1 mL of DNA solution, followed by mixing by gentle inversion. After 2 hours at room temperature, the DNA was pelleted by centrifugation at 10,000 x g for 30 min at 4°C. The supernatant was removed, and the DNA precipitate washed by gentle addition of 1 mL of cold 70% ethanol and incubation for 60 seconds, followed by centrifugation at 10,000 x g for 15 min at 4°C. The supernatant was removed, and the DNA pellet air dried for 10 min. The purified DNA was resuspended in a final volume of 12 uL of 1X TE buffer (10 mM Tris•HCl, pH 8.0; I mM EDTA pH 8.0), and allowed to dissolve for a minimum of 10 min at room temperature, before final storage at 4°C. Final DNA quantity and quality was assessed initially by spectrophotometry and agarose gel electrophoreses, and final yields were quantified via Quant-iT Picogreen dsDNA fluorimetric assay (catalogue #P7589, Invitrogen).

Virus-enriched samples from 110 L of 0.22um pre-filtered seawater collected at a depth of 25 m yielded a total of 3.2 µg of purified, high molecular weight DNA. Virus-enriched samples from 90 L of 0.1 µm pre-filtered seawater collected at depths of 117m and 250 m, yielded a total of 5.3 µg and 3.5 µg of purified, high molecular weight DNA, respectively.

Nanopore sequencing methodology

Care was given to minimize shearing when handling DNA. Sequencing was conducted on a GridION X5 with FLO-MIN106 (R 9.4.1) flowcells (Oxford Nanopore Technologies, Ltd.). Read basecalls were generated from the signal traces using Guppy v2.2.2 for the 117 m and 250 m samples, while Guppy v3.0.4 was used for the 25 m sample. The sequencing output from each of the three flowcells ranged from 5.15 - 12.28 gigabases, generating read lengths up to 254 kb (Table 1). For the lambda spike-in experiment, 10 ng of purified lambda DNA (Cat.#N3013S, New England Biolabs, Ipswich, MA) was added to 1.5 ug of virus-enriched DNA purified from tangential flow filtration concentrates from seawater. The lambda-spiked DNA sample was then processed and sequenced using exactly the same methods as described for all other samples.

Preparation of short-read libraries and Illumina sequencing

A total of 60 ng of genomic DNA from each sample was sheared to an average size of 350 bp using a Covaris M220 Focused-ultrasonicator (Covaris, Woburn, MA) with Micro AFA fiber tubes (Covaris, #520166, Woburn, MA). Libraries were sequenced using a 150 bp paired-end NextSeq High Output V2 reagent kit (Illumina, FC-404-2004, San Diego, CA). Illumina short reads were assembled into contigs in two steps. First, low quality sequence was removed using iu-filter-quality-minoche from the illumin-utils package (Eren et al. 2013). Second, remaining reads were assembled into contigs using the "meta-sensitive" mode of MEGAHIT (Li et al. 2015).

Summary of phage discovery bioinformatic pipeline

We developed an assembly-free bioinformatic pipeline to isolate and polish full-length phage genomes from nanopore reads (Fig. 1). An initial filtering step removed all reads that were not bound by DTR sequences, as this common feature of tailed dsDNA viruses (Casjens and Gilcrease 2009) served as a marker of phage genome completeness. The 5-mer count vectors were created for all DTR-containing reads. Such 5-mer frequency features have been shown to be useful for binning and classifying prokaryotic genome sequencing data, using approximations of genome-wide patterns of *k*-mer usage (e.g. from assembled contigs or long reads) (Teeling et al. 2004; Saeed et al. 2012; Laczny et al. 2014; Beaulaurier et al. 2018). Since such whole genome reads are not sampled from different regions of the genome, the only source of within-genome read variation in 5-mer count vectors is due to the sequencing error, rendering the 5-mer count vector a simplified representation of the genome sequence. Next, 5-mer count vectors were reduced from 512 to 2 dimensions using the uniform manifold approximation and projection (UMAP) technique (McInnes et al. 2018) and bins were automatically called in the resulting two-dimensional space using the hdbscan algorithm (McInnes et al. 2017) (Table 1).

Each bin identified by this approach was further refined using an alignment-based clustering approach. Reads in each bin were aligned to each other and clustered based on pairwise read alignment scores that prioritize alignment length and sequence similarity. One read was selected from each alignment cluster to serve as the draft genome reference and the remaining reads in the cluster were used for polishing.

Draft genome deduplication

In order to ensure that each polished sequence represents a unique phage genome, the polished sequences were deduplicated. An all-vs-all alignment of the polished genomes from a given depth sample was done to eliminate redundant sequences. Polished sequences were considered redundant if alignment by NUCmer v3.1 (Kurtz et al. 2004) using "--nosimplify" resulted in a hit with >=98% identity and covering >=98% of both polished sequences. In the case of redundant polished sequences, the sequence with the higher number of polishing reads was retained.

Additional polishing with Illumina short reads

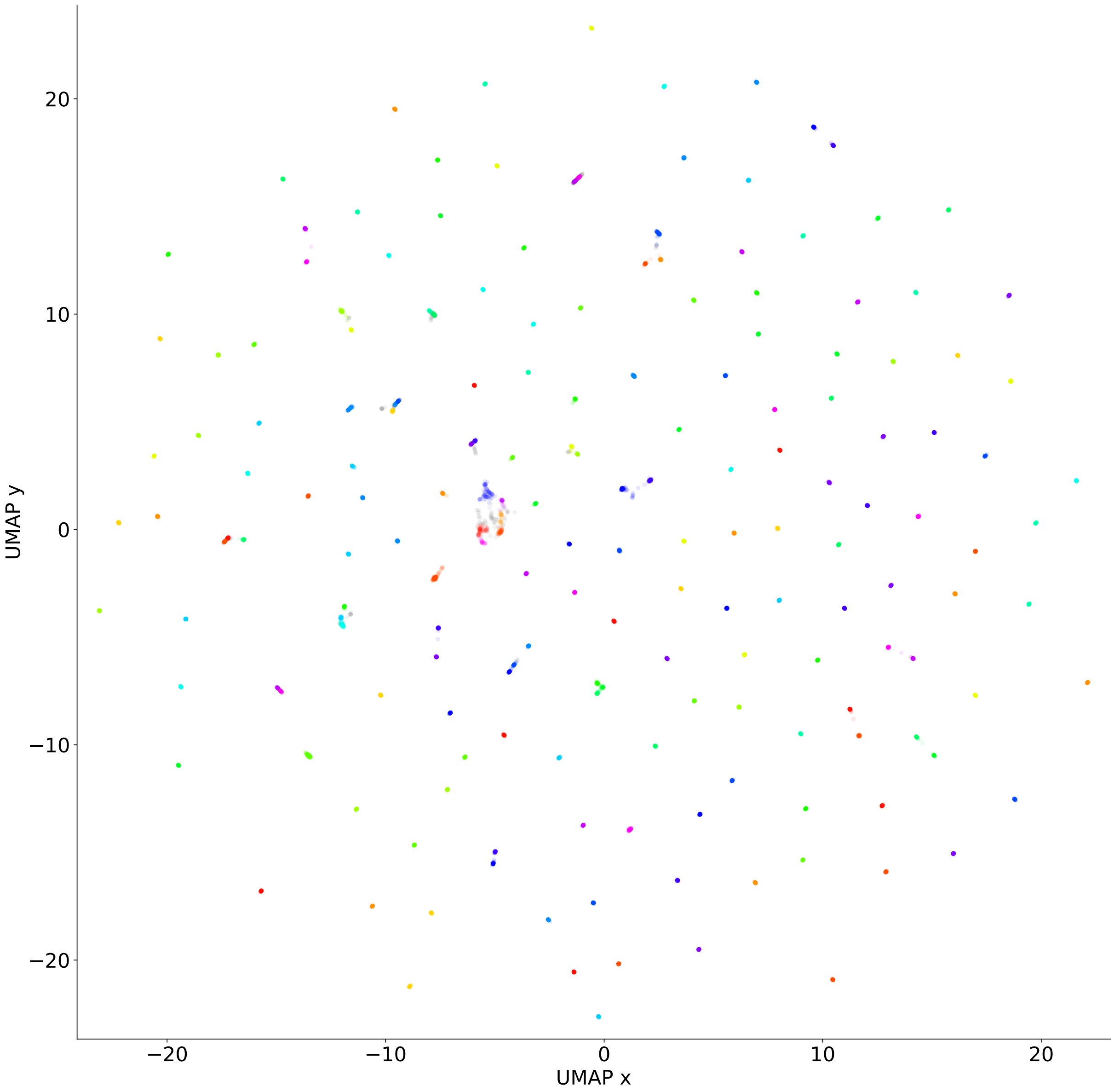
Draft genomes were refined further by mapping short reads with BWA-MEM v0.7.13 (Li 2013), indexing alignments with SAMtools v1.9 (Li 2009), and polishing with Pilon v1.23 (Walker 2014). Naïve alignment of short reads to the putative PICI sequences resulted in uneven mapping due to the repetitive structure and therefore incomplete polishing. To circumvent this, putative PICI sequences were fragmented into 5 kb chunks that were polished separately using Pilon then re-concatenated.

Phage genome relative abundances in short-read assemblies vs. long-read polished genomes

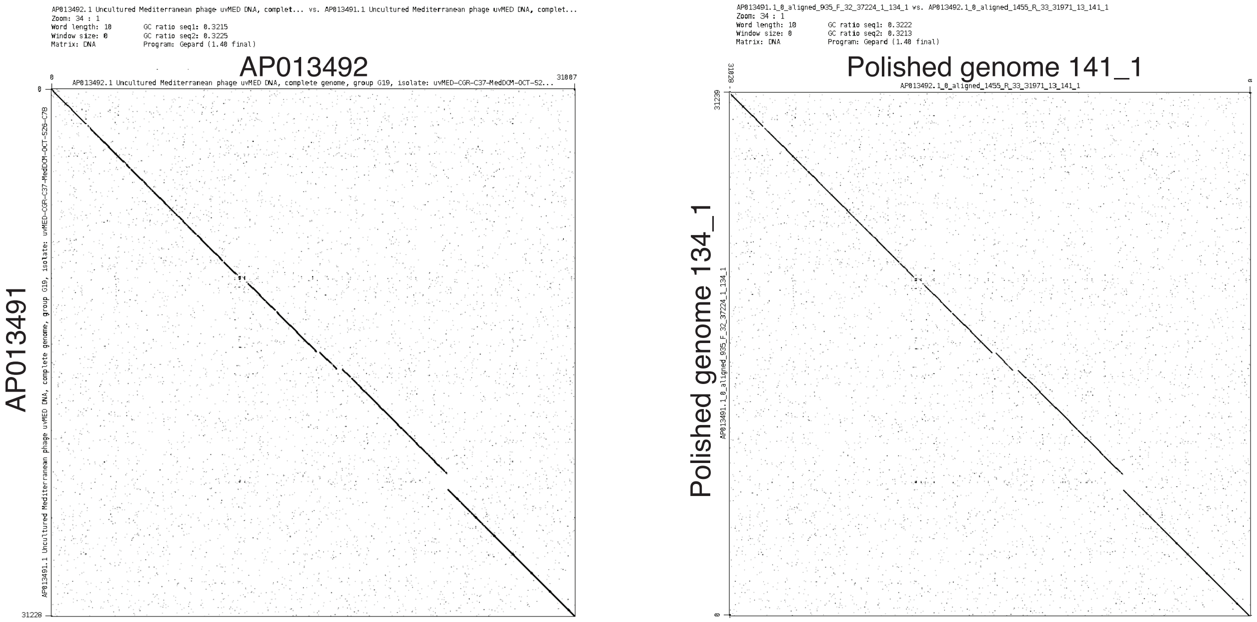
Relative abundances of polished genomes were compared between short- and long-read sequencing, using normalized coverage for short reads, and relative number of reads in each alignment cluster for long reads (Supplemental Fig. S7). BWA-MEM v0.7.15 (Li 2013) and msamtools (Arumugam et al. 2010) was used to map and filter corresponding short reads to polished nanopore genomes at >95% average nucleotide identity (ANI) across >45 bp. SAMtools (Li 2009) was used to calculate nucleotides mapping to polished genome. Coverage was calculated for each genome by dividing nucleotides mapped to the length of the genome, and normalized coverage was calculated by dividing coverage to genome by summed coverage across all genomes in sample. Pearson's correlations were assessed with the stats v3.6.1 package on R (R Core Team 2018).

AFVG large subunit terminase sequence comparisons

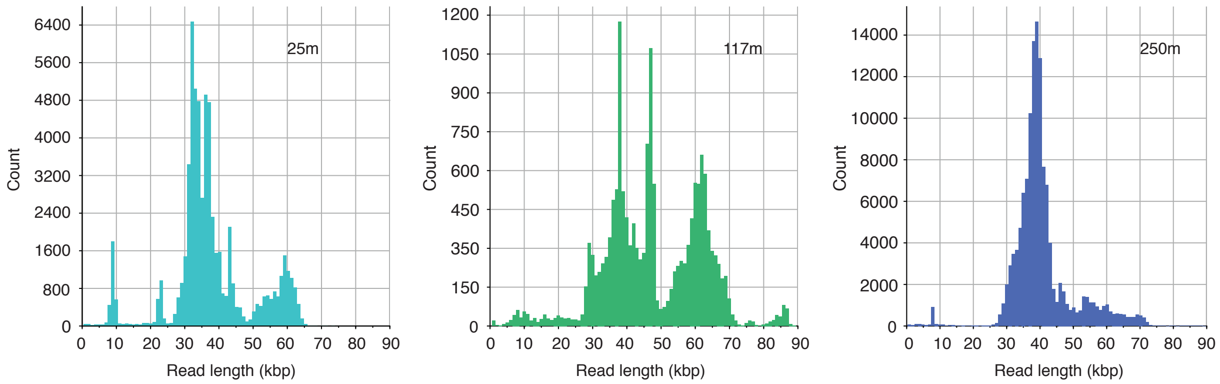
To examine the diversity of AFVG terminase-like protein coding sequences with homology to known large subunit terminases (Supplementary Fig. S14), 1,051 AFVG terminase-like protein coding sequences were compared to large subunit terminase sequences derived from seven reference enterobacteriophages, as well as several marine bacteria and cyanobacteria. A phylogenetic tree was generated with the ete3 standard_fasttree workflow (Sievers, 2011; Morgan, 2010; Huerta-Cepas, 2016) using ete3 default parameters for maximum likelihood phylogenetic tree generation.



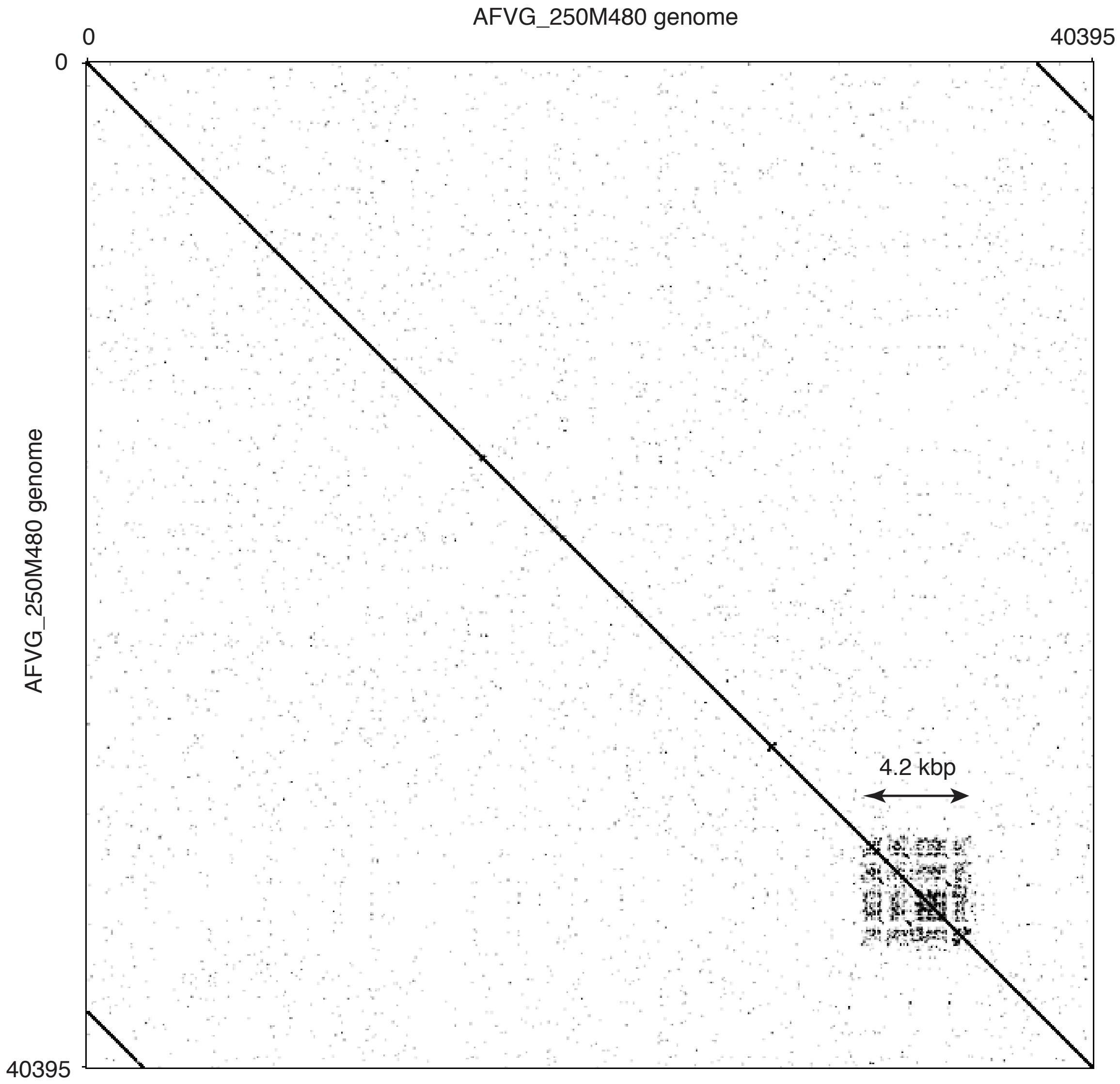
Supplementary Figure S1. Nanopore reads simulated from the uvMED phage genome. Nanopore reads simulated from the uvMED phage genome collection (Mizuno et al. 2016) (see Methods) were represented by their 5-mer counts and dimensionally reduced into a 2D embedding using UMAP (McInnes et al. 2018). Reads in the 2D embedding are colored based on their assignment to the 183 bins called by hdbscan (McInnes et al. 2017). Some bin colors are redundant due to the large number of bins. Reads not assigned to a bin are colored grey.



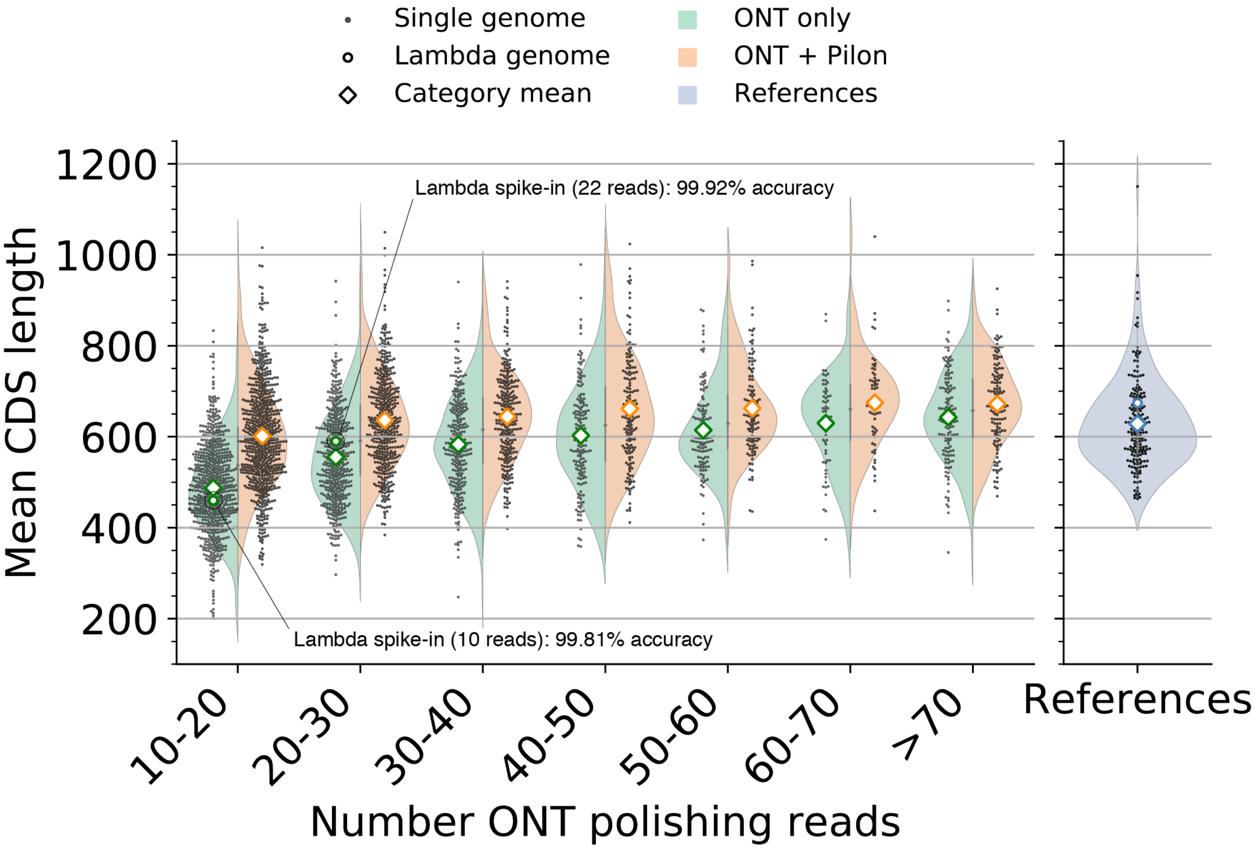
Supplementary Figure S2. Strain-level differences among reference genomes maintained in polished genome reconstructions. The two reference genomes AP013491 and AP013492 from the uvMED collection share a high degree of sequence similarity across their full lengths. These similarities are conserved in the two polished genomes corresponding to these two references produced by the phage discovery pipeline.



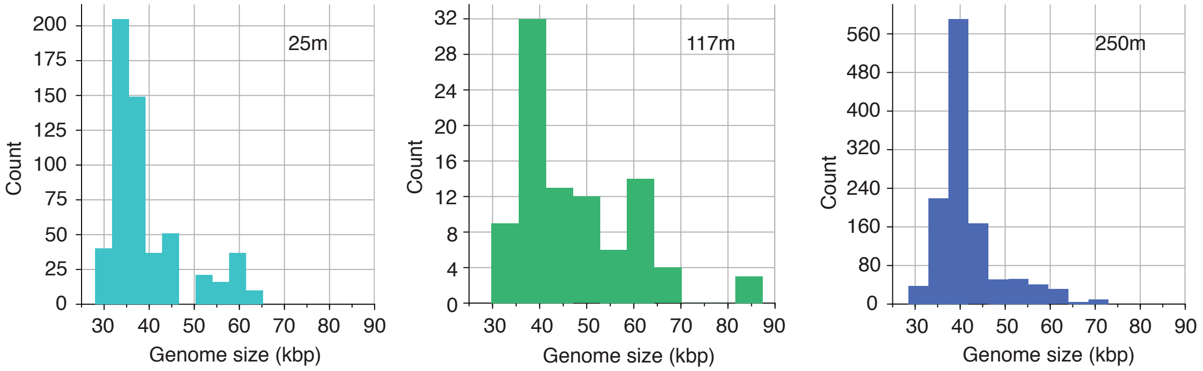
Supplementary Figure S3. Read length distributions of DTR-containing reads in each sample. After filtering for only reads that contain direct terminal repeats (DTRs), each sample depth revealed an enrichment for reads between 30-50 kb. Reads of 55+ kb were more prevalent in the 25 m and 117 m samples than in the 250 m sample.



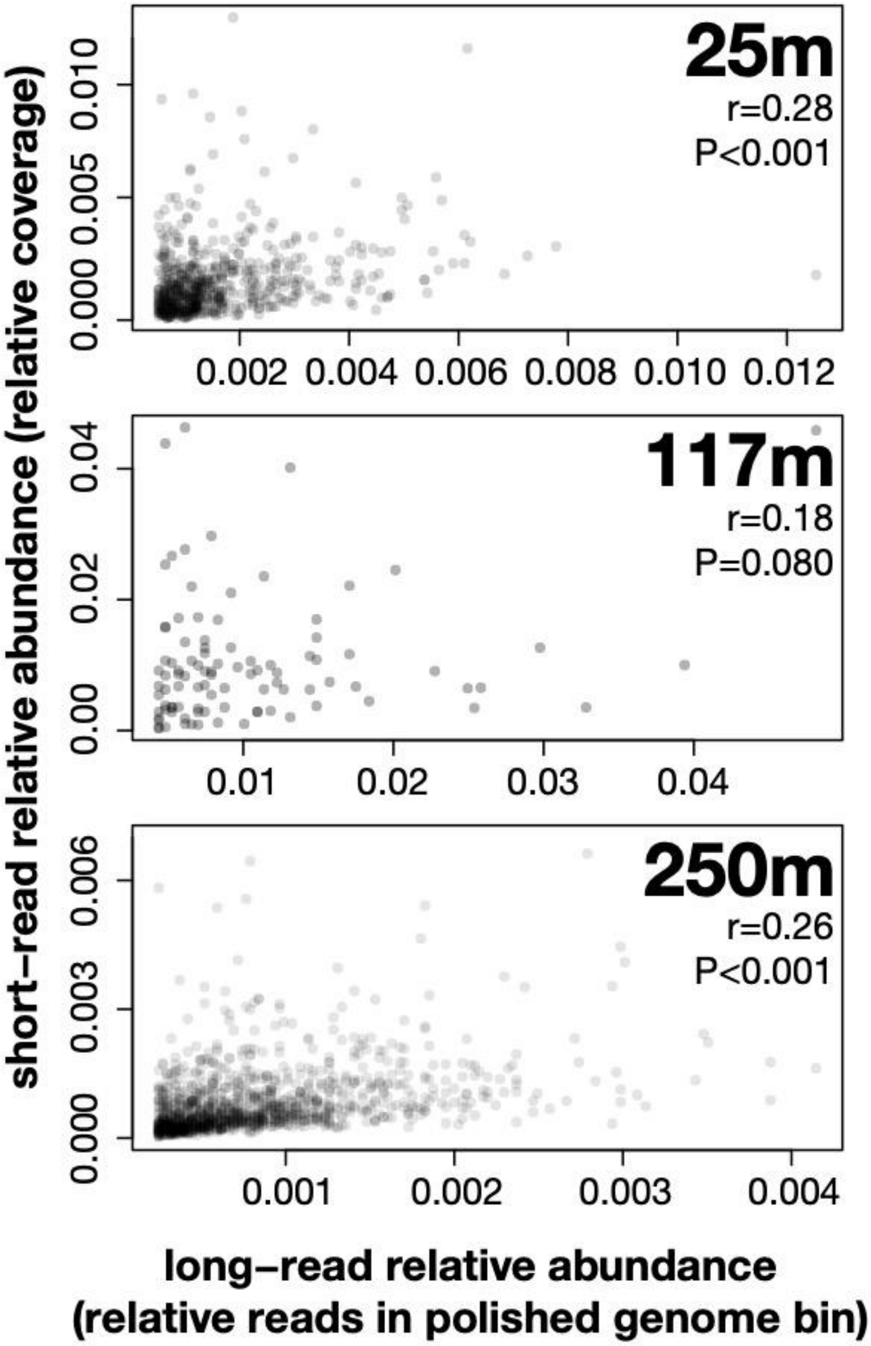
Supplementary Figure S4. Dot plot showing complex repeats in a polished virus genome. The ~40.4 kb polished draft genome AFVG_250M480 obtained from the 250 m sample reveals a 4.2 kb region of complex repeats that is predicted to result in significant difficulties for short-read assembly tools.



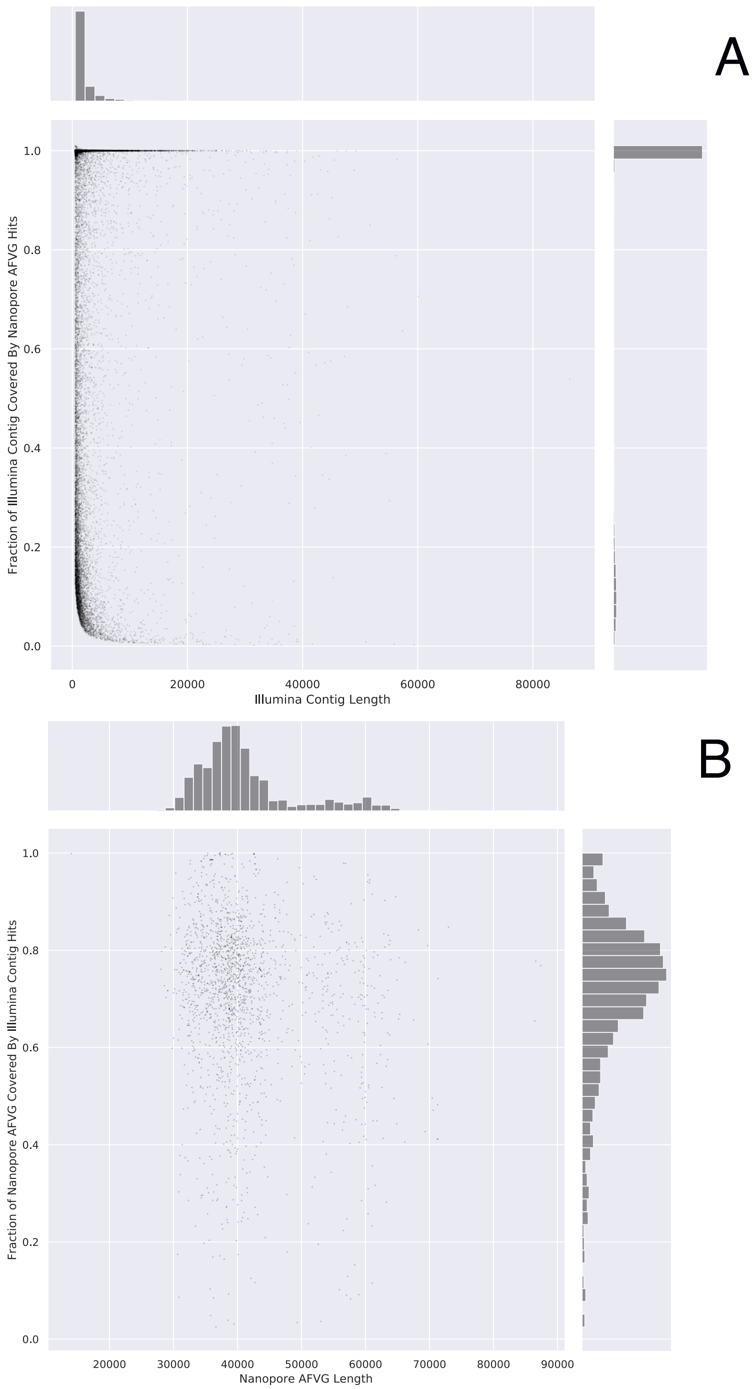
Supplementary Figure S5. Phage genome quality improves with increasing number of nanopore reads for polishing. All polished draft genomes produced from the three sample depths (25, 117, and 250 m) were assessed for genome quality by calculating the mean annotated protein coding sequence (CDS) length. The annotated CDS length of the uvMED genomes served as a benchmark for reference-quality marine phage genomes. Polished phage genomes are grouped by the number of Oxford Nanopore Technologies (ONT) reads used during the polishing steps (i.e. the number of reads in each alignment cluster minus one to serve as the draft genome sequence). Additional short read polishing of the nanopore-polished AFVGs (ONT only) was done using Pilon (Walker et al. 2014). These short read polished AFVGs (ONT + Pilon) showed increased mean CDS lengths in each of the coverage bins. When ten reads from a lambda phage spike-in were admitted into the phage discovery pipeline alongside environmental phage reads, we reconstituted the lambda genome at 99.81% accuracy. When 22 reads were admitted, accuracy increased to 99.92%.



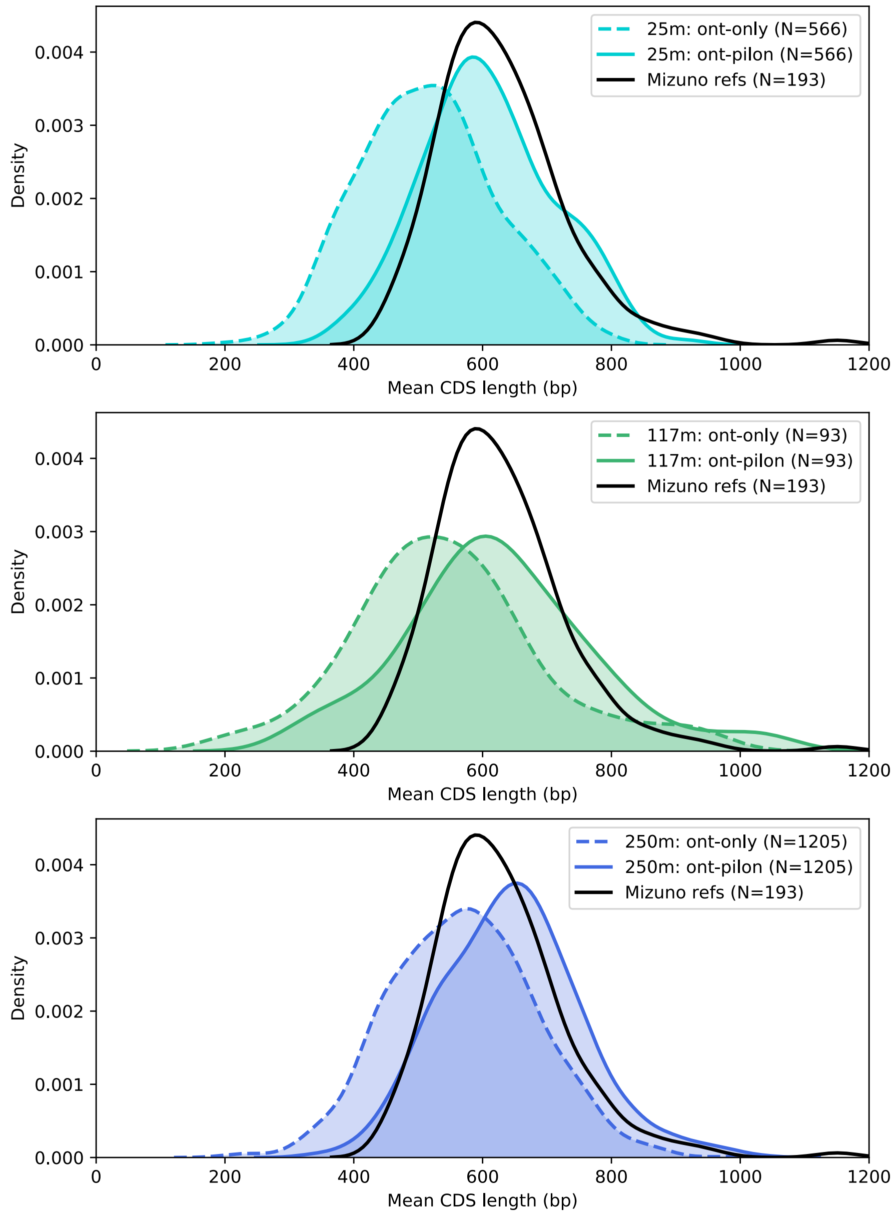
Supplementary Figure S6. Polished draft genome yields and length distributions. The genome lengths of the polished AFVGs obtained from the 25 m (left), 117 m (middle), and 250 m (right) virus-enriched DNA collected from seawater.



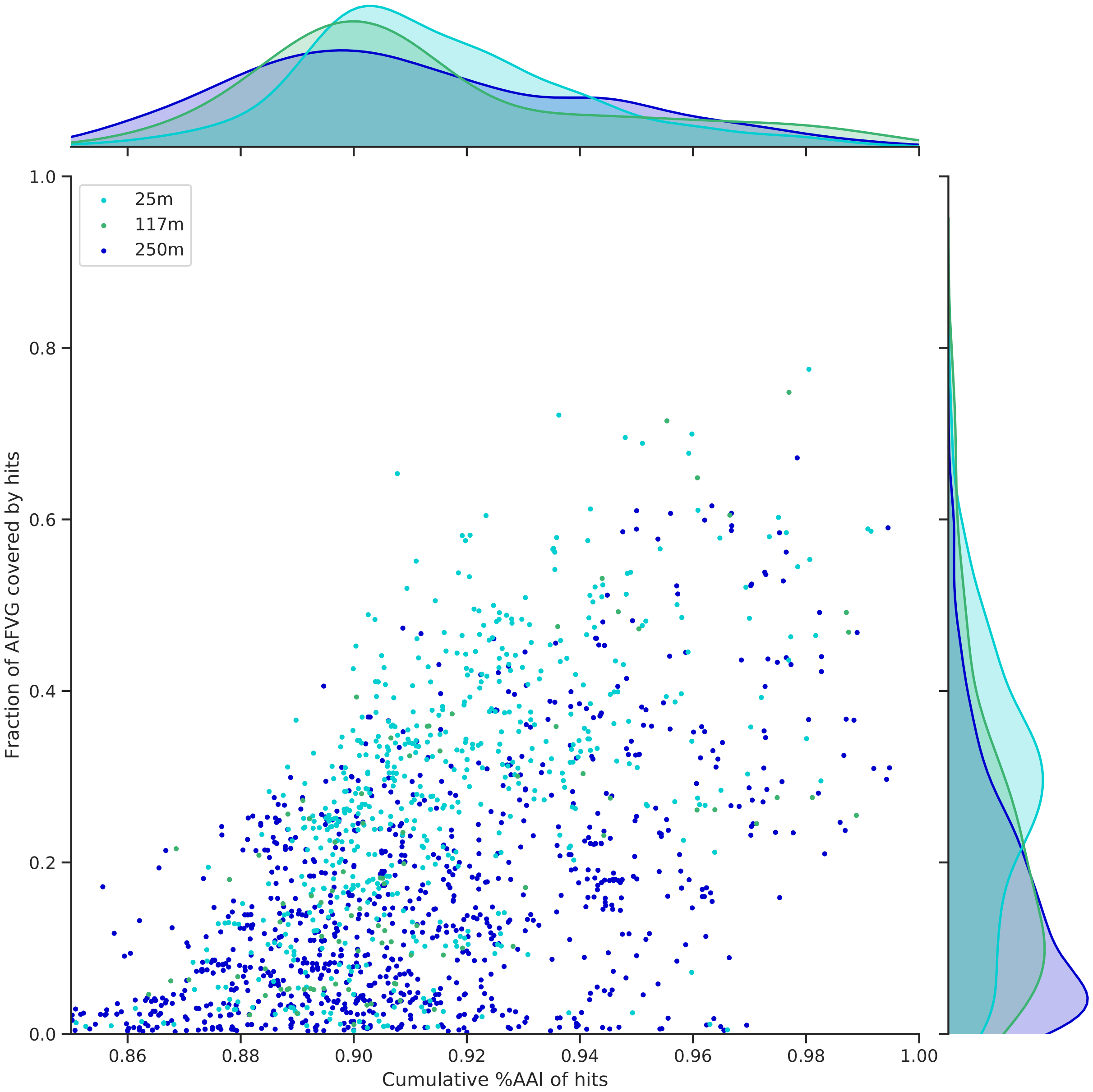
Supplementary Figure S7. AFVG bin read abundances versus short read sequence coverage. Each point represents one AFVG. X-axis represents normalized nanopore read abundance, calculated from the number of reads in an AFVG alignment cluster divided by total number of reads in all AFVG alignment clusters. The Y-axis represents normalized Illumina short read abundance, calculated by dividing coverage to an AFVG by the total coverage to all AFVGs. Correlation coefficients and P-values are shown from Pearson correlation.



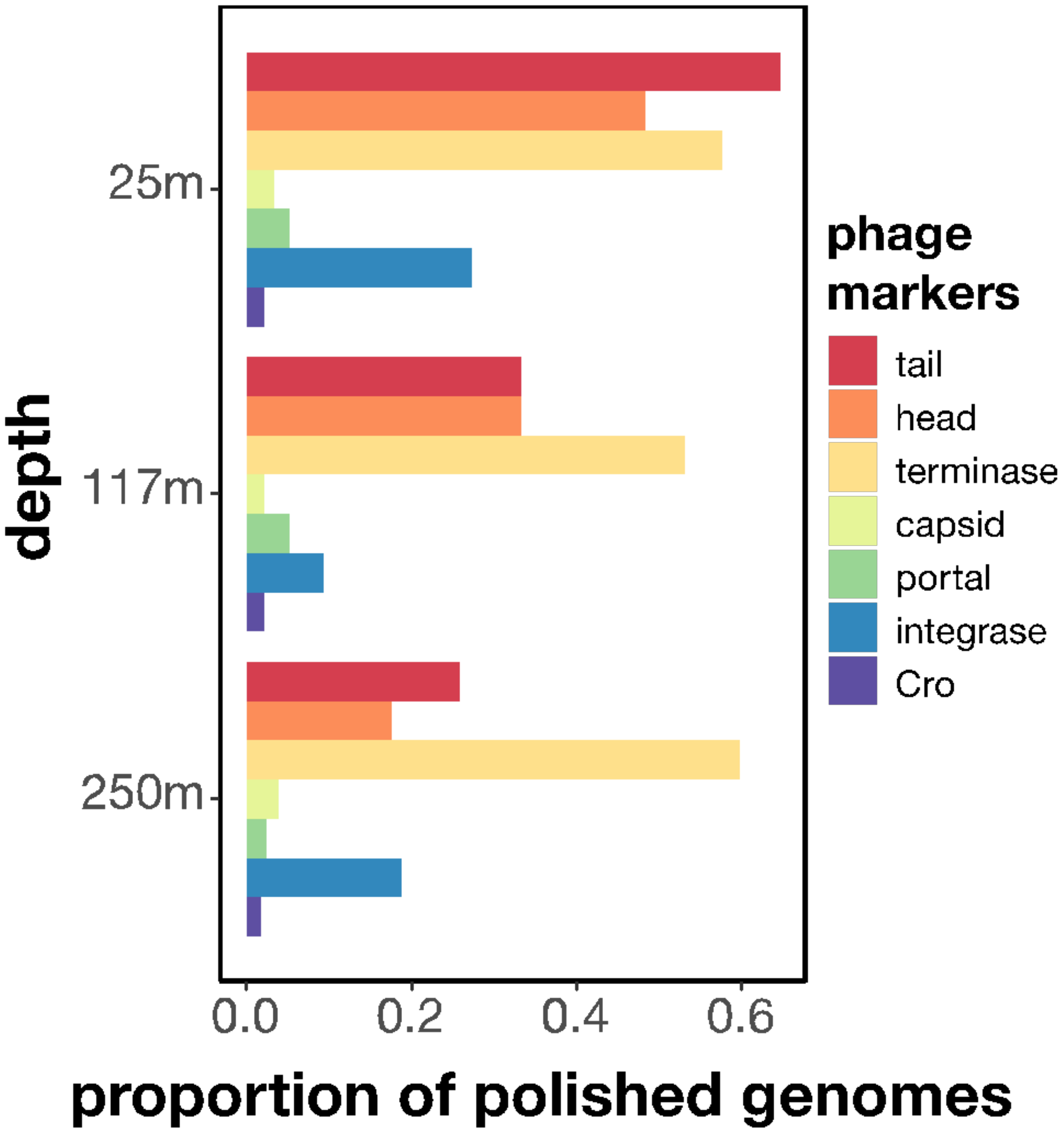
Supplementary Figure S8. Contigs from assembled short-reads vs AFVGs. (A) Each point in the central scatter plot represents one contig, and its position indicates the length of the contig and the fraction covered by its best match to any AFVGs. The histogram in the right margin shows the distribution of y values, revealing that most hits cover the full contig. Similarly, the histogram in the top margin shows the contig length distribution. **(B)** Each point in the central scatter plot represents one AFVG showing its length and the fractional length of its best hit to any one contig.



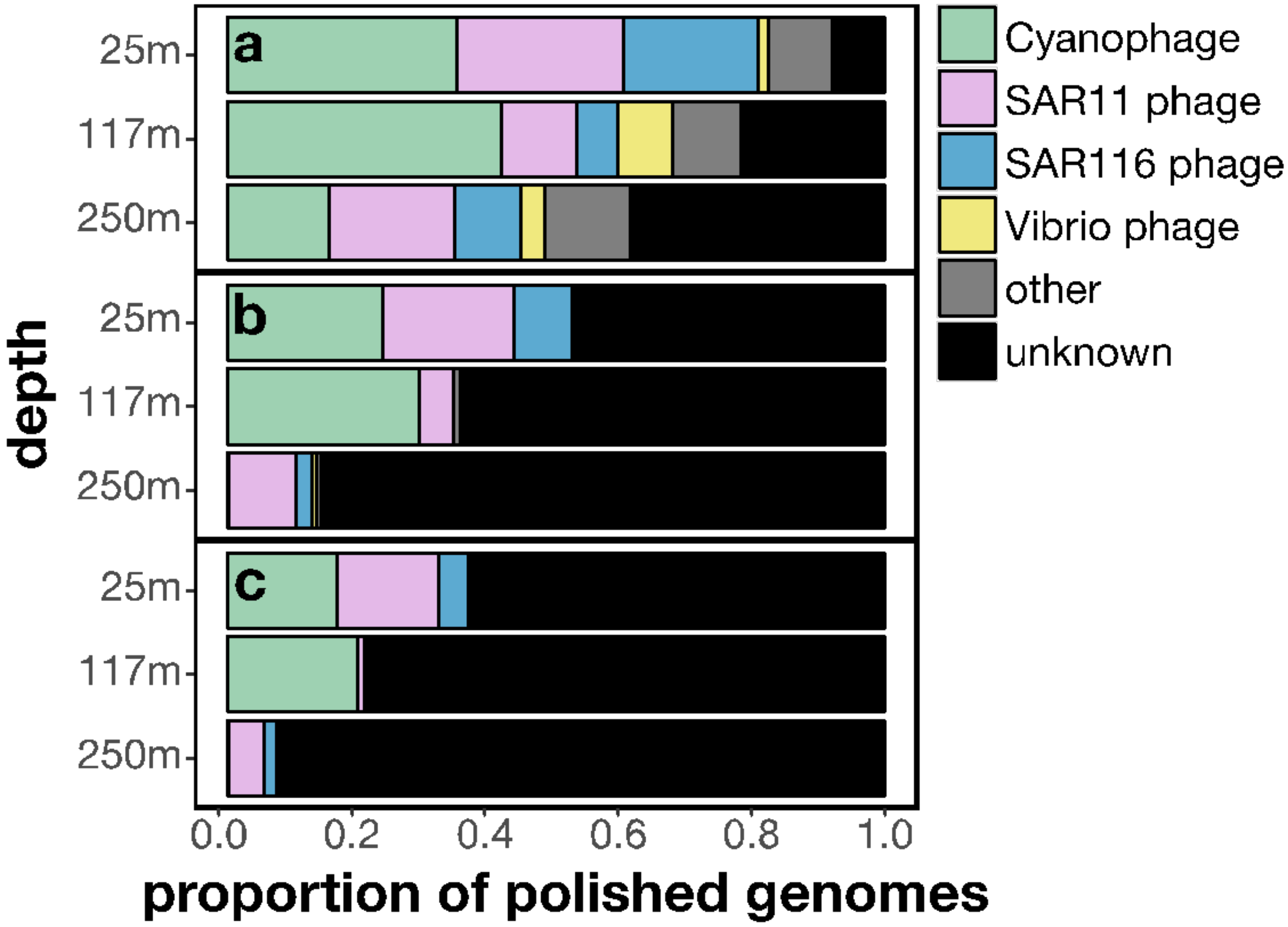
Supplementary Figure S9. AFVG quality improvements from short read polishing. The distributions of the mean annotated CDS length for each AFVG from each sample (top: 25 m, middle: 117 m, bottom: 250 m). Overall mean CDS lengths were increased when the nanopore-only polished AFVGs (dashed line) were subject to an additional round of polishing using sample-matched short reads via Pilon (solid line). The distribution of mean CDS lengths for the uvMED marine phage reference genomes is shown for comparison (solid black line).



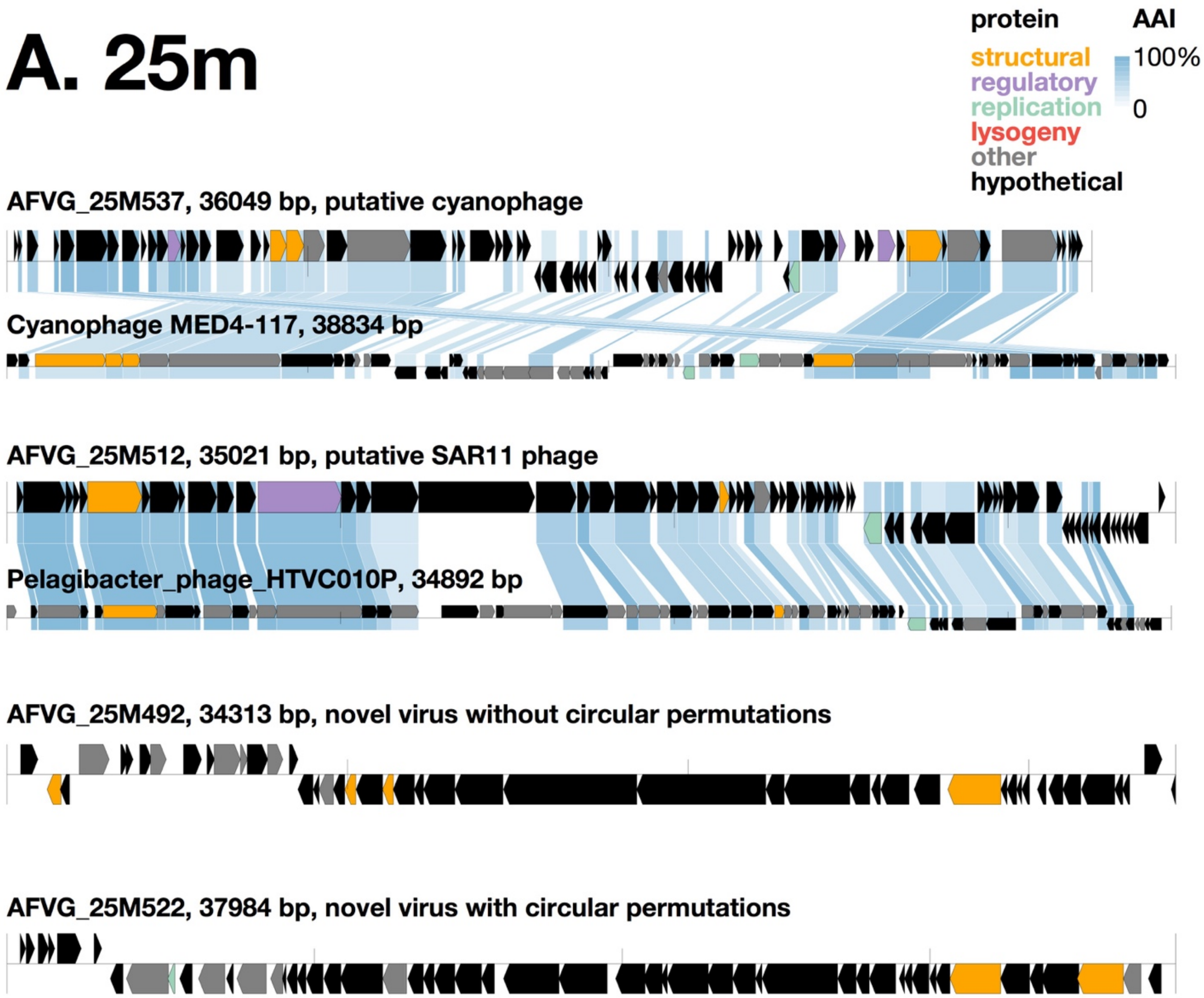
Supplementary Figure S10. Similarity of AFVG genes to genes in the NCBI RefSeq release 84 database. Each point represents an AFVG colored by the depth at which it was found. The y-axis encodes the fraction of genes with matches to reference genes and the x-axis encodes the cumulative percent amino acid identity (%AAI) of matches to each AFVG. The marginal histograms show the distribution of values for "cumulative %AAI" (top) and "fraction of genes" (right) grouped by depth (Blue, 25 m sample; Green, 117 m sample; Purple, 250 m sample).



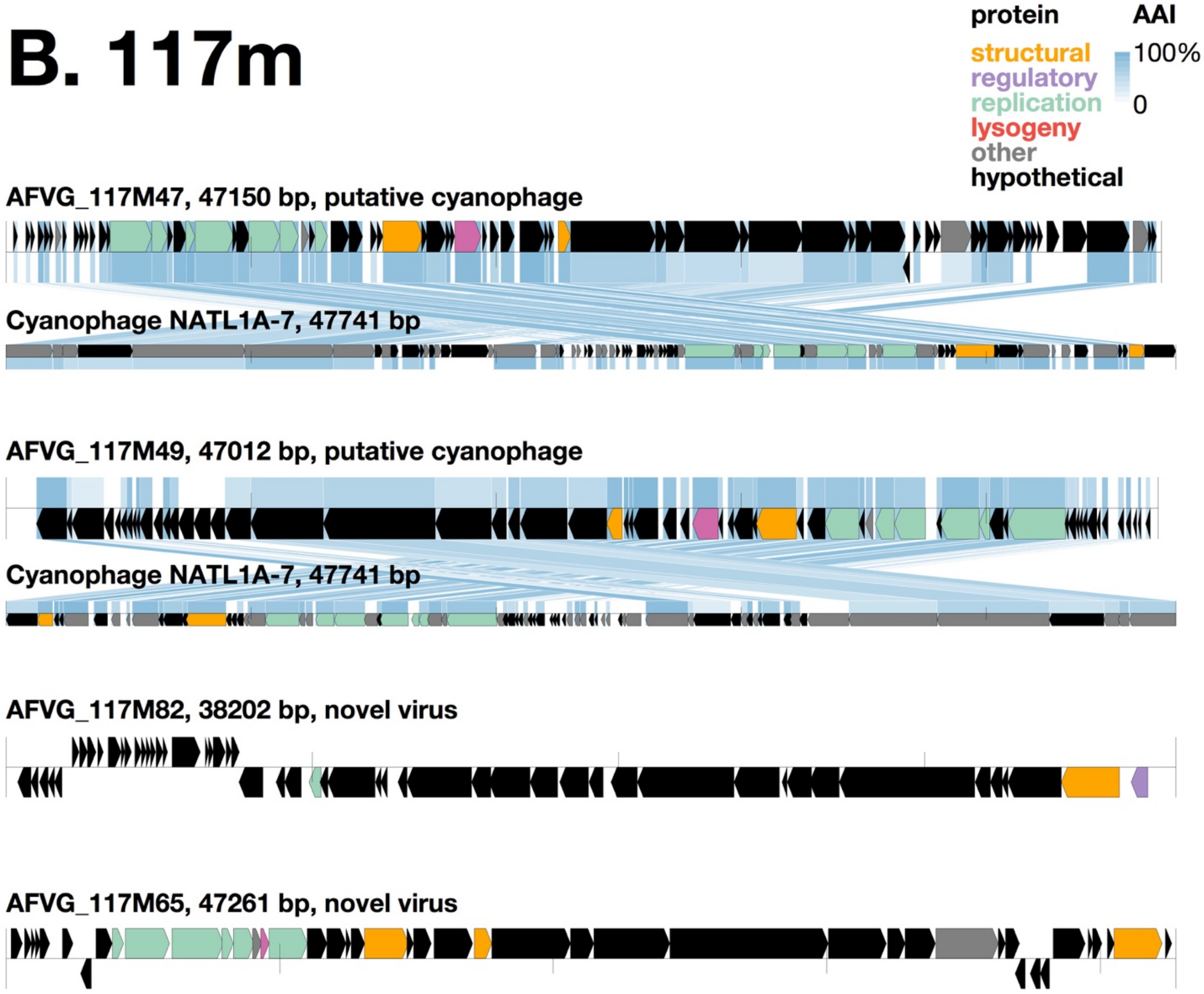
Supplementary Figure S11. Virus marker genes found in AFVGs from each sample. Proportion of polished nanopore AFVGs at each depth having one or more top protein hit (Pfam bit score >30) to the following virus and prophage marker proteins: terminase, tail, head, capsid, portal, integrase, and Cro.



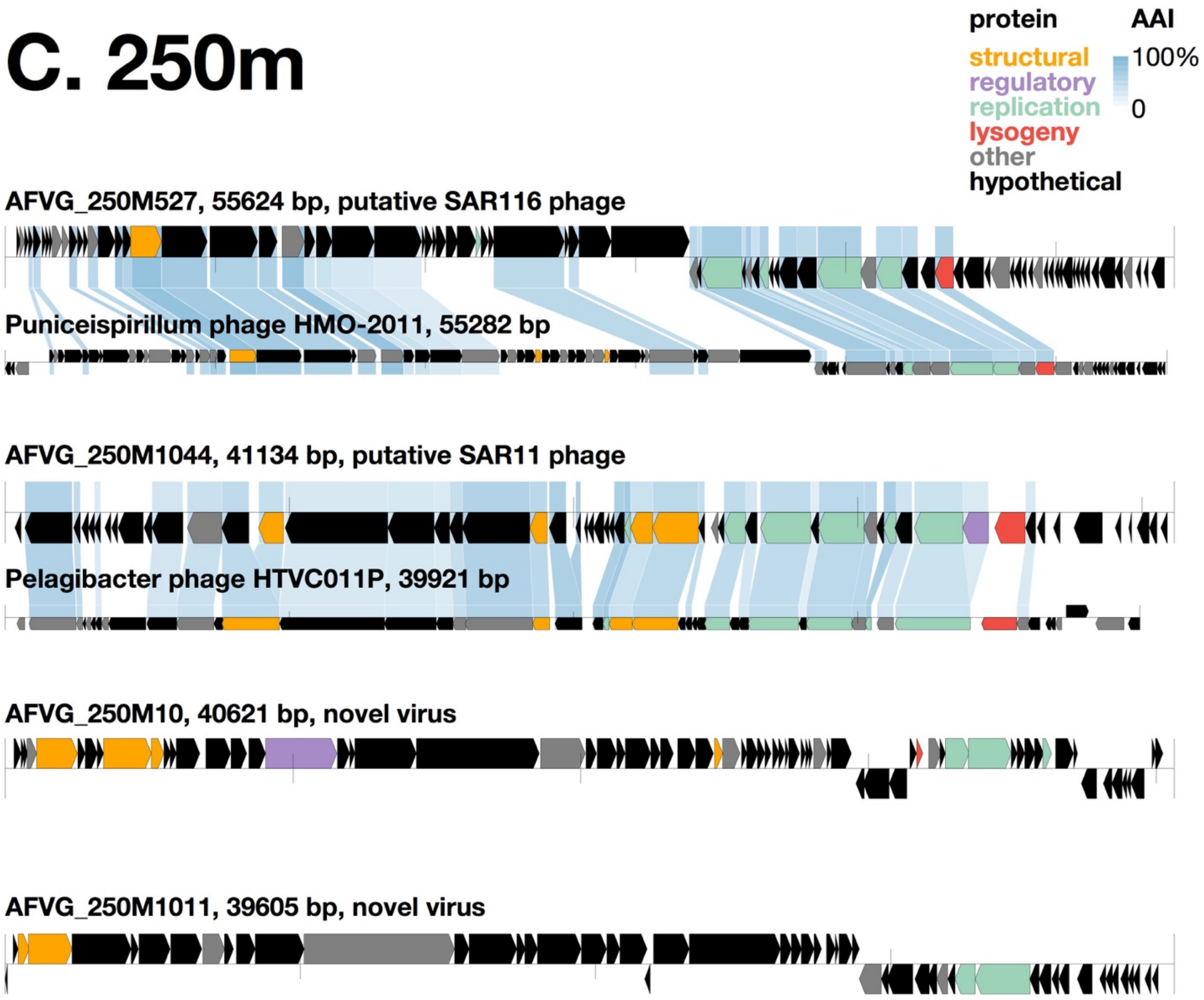
Supplementary Figure 12. Putative virus taxon types identified in AFVGs in each sample. Proportion of polished nanopore genomes functionally annotated to common virioplankton taxa on RefSeq release 92 (O'Leary et al. 2016) using cut-offs based on the following number of AFVG proteins in a single AFVG having a top match to a single genus: (**A)** one or more proteins, **(B)** three or more proteins, and **(C)** five or more proteins.


A. 25m
protein
structural
regulatory
replication
lysogeny
other
hypothetical
AAI
100%
0
AFVG_25M537, 36049 bp, putative cyanophage
Cyanophage MED4-117, 38834 bp
AFVG_25M512, 35021 bp, putative SAR11 phage
Pelagibacter_phage_HTVC010P, 34892 bp
AFVG_25M492, 34313 bp, novel virus without circular permutations
AFVG_25M522, 37984 bp, novel virus with circular permutations

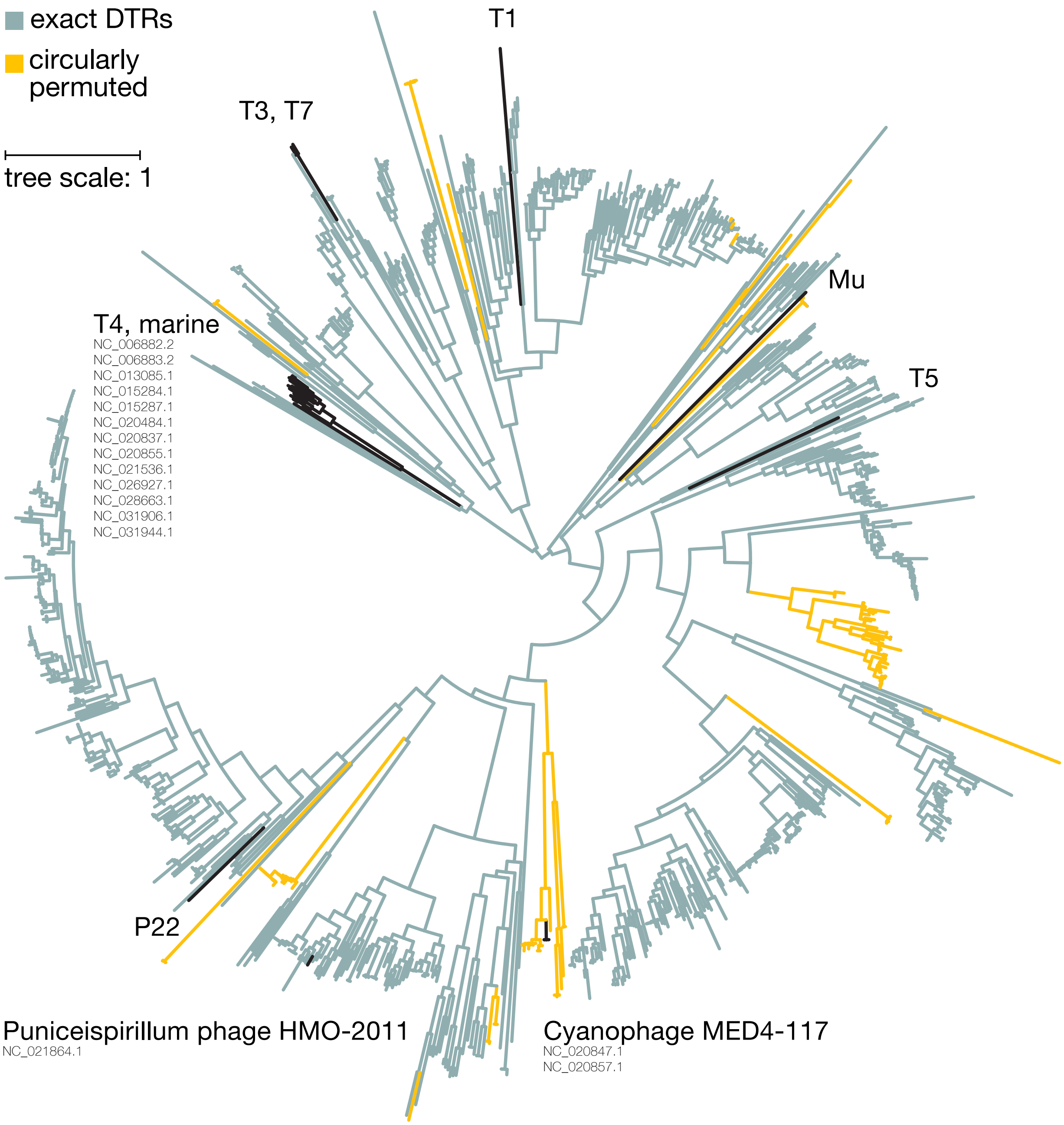
Supplementary Figure 13A. Annotation plots of representative AFVGs from 25 m. The genomic structure and functional annotations of four abundant AFVGs recovered from the 25 m sample. Protein coding sequences are color-coded based on functional annotations to Pfam (El-Gebali et al. 2019) (bit score >30). For some AFVGs with multiple hits to the same reference genome on RefSeq release 92 (O'Leary et al. 2016), that reference genome is included below, with blue shading representing amino acid identity (AAI) between homologous proteins.


B. 117m
protein
structural
regulatory
replication
lysogeny
other
hypothetical
AAI
100%
0
AFVG_117M47, 47150 bp, putative cyanophage
Cyanophage NATL1A-7, 47741 bp
AFVG_117M49, 47012 bp, putative cyanophage
Cyanophage NATL1A-7, 47741 bp
AFVG_117M82, 38202 bp, novel virus
AFVG_117M65, 47261 bp, novel virus

Supplementary Figure 13B. Annotation plots of representative AFVGs from 117 m. The genomic structure and functional annotations of four abundant AFVGs recovered from the 117 m sample. Protein coding sequences are color-coded based on functional annotations to Pfam (El-Gebali et al. 2019) (bit score >30). For some AFVGs with multiple hits to the same reference genome on RefSeq release 92 (O'Leary et al. 2016), that reference genome is included below, with blue shading representing amino acid identity (AAI) between homologous proteins.


C. 250m
protein
structural
regulatory
replication
lysogeny
other
hypothetical
AAI
100%
0
AFVG_250M527, 55624 bp, putative SAR116 phage
Puniceispirillum phage HMO-2011, 55282 bp
AFVG_250M1044, 41134 bp, putative SAR11 phage
Pelagibacter phage HTVC011P, 39921 bp
AFVG_250M10, 40621 bp, novel virus
AFVG_250M1011, 39605 bp, novel virus

Supplementary Figure 13C. Annotation plots of representative AFVGs from 250 m. The genomic structure and functional annotations of four abundant AFVGs recovered from the 250 m sample. Protein coding sequences are color-coded based on functional annotations to Pfam (El-Gebali et al. 2019) (bit score >30). For some AFVGs with multiple hits to the same reference genome on RefSeq release 92 (O'Leary et al. 2016), that reference genome is included below, with blue shading representing amino acid identity (AAI) between homologous proteins.



Supplementary Figure 14. Phylogenetic relationships of AFVG and reference sequence large subunit terminase protein sequences. A maximum likelihood phylogenetic tree was constructed from phage terminase large-subunit domains of AFVGs, plus reference sequences from well-characterized enterobacteriophages, and several marine planktonic cyanophage and bacteriophage cultured isolates. The latter group of reference sequences included bacteriophage that infect the marine photoautotrophic cyanobacteria (*Prochlorococcus* and *Synechococcus*) and common heterotrophic bacterioplankton (*Pelagibacter* and *Puniceispirillum*), with RefSeq accessions indicated in the labels. The black lines indicate reference lineages, and blue and orange lines show AFVGs with either exact DTRs, or circularly permuted ends, respectively.

Supplementary References

Arumugam M, E.D. H, K.U. F, J. R, P. B. 2010. SmashCommunity: a metagenomic annotation and analysis tool. *Bioinformatics* **26**: 2977–2978.

Beaulaurier J, Zhu S, Deikus G, Mogno I, Zhang XS, Davis-Richardson A, Canepa R, Triplett EW, Faith JJ, Sebra R et al. 2018. Metagenomic binning and association of plasmids with bacterial host genomes using DNA methylation. *Nat Biotechnol* **36**: 61-69.

Casjens SR, Gilcrease EB. 2009. Determining DNA packaging strategy by analysis of the termini of the chromosomes in tailed-bacteriophage virions. *Methods Mol Biol* **502**: 91-111.

El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, Qureshi M, Richardson LJ, Salazar GA, Smart A et al. 2019. The Pfam protein families database in 2019. *Nucleic Acids Res* **47**: D427-D432.

Eren A, Vineis J, Morrison H, ML S. 2013. Correction: A Filtering Method to Generate High Quality Short Reads Using Illumina Paired-End Technology. *PLoS ONE 8* **8**: (6).

Huerta-Cepas J, Serra F, Bork P. 2016. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Molecular Biology and Evolution,* 33:1635–1638.

Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL. 2004. Versatile and open software for comparing large genomes. *Genome Biol* **5**: R12.

Laczny CC, Pinel N, Vlassis N, Wilmes P. 2014. Alignment-free visualization of metagenomic data by nonlinear dimension reduction. *Sci Rep* **4**: 4516.

Li H. 2009. The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics. *Bioinformatics* **25**: 2078-2079.

Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv* **q-bio**.

Li D, Liu CM, Luo R, Sadakane K, Lam TW. 2015. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**: 1674-1676.

McInnes L, Healy J, Astels S. 2017. hdbscan: Hierarchical density based clustering. *Journal of Open Source Software* **2**: 205.

McInnes L, Healy J, Melville J. 2018. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv e-prints*.

Mizuno CM, Ghai R, Saghaï A, López-García P, Rodriguez-Valera F. 2016. Genomes of Abundant and Widespread Viruses from the Deep Ocean. *mBio* 7: e00805-00816. Morgan N. Price ,Paramvir S. Dehal, Adam P. Arkin. 2010. FastTree 2 – approximately maximum-likelihood trees for large alignments. PLoS One **5**:e9490.

O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D et al. 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* **44**: D733-745.

R Core Team. 2018. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL. https://www.R-project.org/.

Saeed I, Tang SL, Halgamuge SK. 2012. Unsupervised discovery of microbial population structure within metagenomes using nucleotide base composition. *Nucleic Acids Res* **40**: e34.

Sievers F, Wilm A, Dineen DG, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, Thompson JD, Higgins DG. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology* **7**:539.

Teeling H, Waldmann J, Lombardot T, Bauer M, Glockner FO. 2004. TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. *BMC Bioinformatics* **5**: 163.

Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA,Zeng Q, Wortman J, Young SK, Earl AM. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9**:e112963.