Supplementary Appendix for Aylward *et al*., "Diel Cycling and Long-Term Persistence of Viruses in the Ocean's Euphotic Zone"

Contents

Supplementary Results and Discussion

Viral scaffold annotations and metagenome analyses

By sampling from the same microbial community continuously over a period of ~8 days we were able to track viral dynamics of diel periods. Our recovery of a relatively small set of 483 scaffolds compared to the sequencing effort (>500 Gbp of raw metagenomic data for 88 samples) is consistent with our finding of remarkable consistency in the viral assemblage through time. Indeed, a single metagenome from one time-point would have largely sufficed for purposes of only cataloguing the genetic diversity present. The vast majority of the different viral scaffolds were consistently present throughout the cruise, the only exceptions among the largest scaffolds (> 15 kbp in length) arising from VS18, VS87 and VS91, which appeared after ~2 days of sampling (Fig. 1). One of the largest and by far most abundant viral scaffolds (VS2) represents a novel virus with little identifiable homology to sequenced viral genomes or viral contigs from recent studies (Fig. 1, Fig. S4). This scaffold encodes a number of structural and tail proteins, suggesting it is a tailed bacteriophage from the Caudovirales order. Interestingly, despite its abundance in this dataset it was almost completely absent from metagenomes sequenced at Station ALOHA surface waters from 2010-2011 (Fig. 1b), suggesting its overall presence in the NPSG is episodic.

We sought to gain insight into the biological features of the viral populations identified in our datasets through analysis of gene and scaffold annotations. Inference of viral traits from sequence information is currently a major challenge in microbial ecology owing to the lack of information of viral interactions with their hosts *in situ* and the poorly characterized nature of most viral genomes. Additionally, because of the absence of universal marker genes in viral genomes, it is often difficult to ascertain if a contig or scaffold represents a complete genome or genome fragment. In this study, four scaffolds (VS1, 6, 12, and 153) could be confidently binned together into a complete genome due to whole-genome synteny with a reference (*Prochlorococcus* phage P-RSM1) among other characteristics (see Supplementary Methods). Another scaffold (VS3) was circular and therefore can be considered a complete genome. Other scaffolds (e.g., VS10, 30, 31, 58, and 60) likely represent fragments of complete genomes due to homology of the scaffold to only a particular region of a reference genome (Fig. S3). Large scaffolds with no homology to known viruses in our dataset (e.g., VS2, 3, 4, 5, and 7) may represent complete or near-complete genomes, but as some these viral groups do not exist as circular genomes, it is not possible to confirm this with metagenomic data.

Another prominent feature of our dataset and analyses is the ratio of abundances in the viral vs cellular size fractions (the "VC ratio") during the sampling period. The "viral fraction" metagenomes are not composed purely of viral sequences due multiple factors including potential cell lysis during filtration, naked cellular DNA present in environmental samples, and ultra-small cells passing through 0.2 μm pre-filters. Likewise, viral DNA present in the cellular fraction metagenomes may be the result of active intracellular viral infections, virion adherence to larger particles or cells, or the retention of large virions on 0.2 um filters. The contrasting VC ratios obtained for different viral scaffolds may nonetheless point to potential differences in the life history strategies of these groups. For example, viruses with short replication times and large burst sizes would be expected to have high VC ratios since their DNA would be predominantly found in extracellular viral particles. By contrast, viruses with long replication times or non-integrated intracellular dormancy (pseudolysogeny) may be expected to be more abundant in cellular material. Lower VC ratios could also be indications of lysogeny, but this appears unlikely here given the evidence we found for a primarily lytic viral assemblage (predominance of structural genes and endolysins, lack of marker genes for lysogeny, abundance of structural genes in our transcriptomic data, etc.) as well as previous findings of predominantly lytic viruses in the photic zone microbial communities (1, 2). Interestingly, the largest viral scaffolds identified here with low VC ratios belong to novel groups not identified previously (VS3, VS4, VS5, VS13, and VS15), suggesting that they may have eluded characterization with traditional culture-based viral analyses due to unusual life history strategies. One possible explanation entails some form of pseudolysogeny whereby viruses infect their host but remain dormant until conditions are appropriate for virion production. Analysis of VC ratios is not enough to confirm this, however, and future work is needed to elucidate the ecological strategies of these viruses.

Diel analysis of viral abundance profiles and transcription

All viral scaffolds were tested for diel periodicity in both the viral and cellular fraction metagenomic time series (in addition to the cellular RNA metatranscriptomic analyses) to identify possible diel fluctuations in viral relative abundances. No diel scaffolds were identified in the viral fraction metagenome time series, and 11 diel scaffolds were identified in the cellular fraction metagenomic time series (RAIN, corrected p-values < 0.1; Dataset S3). The peak abundance for these scaffolds was between 10-11 am, consistent with our finding of peak viral transcription shortly thereafter.

The finding of little to no diel periodicity of viral abundances in our metagenomic time series contrasts with our quantitative transcriptomic analyses and suggest that diel signatures of viruses are more readily detected in in mRNA rather than DNA. We detected 11 scaffolds with diel periodicity in our cellular fraction metagenomes out of 483 tested (2.3%), but even these findings are modest compared to results from the transcriptomic data, which yielded diel signatures in 26 scaffolds out of 170 tested (15.3%). We postulate that this due to the higher instability of mRNA compared to DNA, which allows for transcriptomic surveys to detect molecules that have been produced immediately before the time of sampling. The higher stability of DNA in viral particles likely produces noise in any diel signature of viral abundance that may exist, since the relative abundance detected in metagenomic surveys is produced by a combination of viruses released immediately before sampling and ambient viral particles that may be hours or even days old.

Our quantitative transcriptional analyses identified diel patterns in 26 scaffolds, 17 of which have homology to known cyanophage and exhibit peak transcriptional activity between noon and midnight, with most between 1200-1400 hrs (Fig. 2a, Dataset S3). *Prochlorococcus* is the most abundant cyanobacterium in the NPSG (4, 5) and the likely host for these viruses, and there are a number of reasons why peak viral activity within this host in the afternoon and evening would benefit viral reproduction. Firstly, many cyanophages including the groups identified in this study, encode AMGs involved in photosynthesis and energy acquisition, which has been hypothesized to play a role in shunting energy towards virion production during infection (6). These AMGs would only be able to benefit viruses if virion production took place in the daytime, which is one possible explanation for why viral replication in laboratory cyanophage-host infection studies were highest when provided light (7). Secondly, light irradiation has been shown to be a large driver of viral particle degradation (8), and viral lysis in the evening (after peak intracellular viral transcription) would prevent the degradation of viruses immediately after lysis and allow time to potentially infect a new host. Thirdly, the timing of cyanophage reproduction in the afternoon and evening likely benefits the viruses because this coincides with the time at which *Prochlorococcus* replicates its own genome, ensuring that both energy reserves and nucleotide monomers are available for viral replication. To confirm the synchronization of viral reproduction with *Prochlorococcus* genome replication we performed both Peak to Trough (bPTR) and index of replication (iRep) analyses for *Prochlorococcus* throughout the sampling period (Figs. 3, S12) (9). Results confirmed that genome replication took place near dusk (~1800 hrs), and analysis of peak transcription of *Prochlorococcus* genes involved in DNA replication and cell division confirmed that these genes

peaked shortly beforehand, as would be expected given the time necessary to produce functional proteins (Figs. 3, S12). Our findings of tightly-coordinated diel cyanophage reproduction that coincides with *Prochlorococcus* growth and genome replication all point towards key adaptations of cyanophage that have allowed them to synchronize their activities to the diel physiological cycles of their host.

Auxiliary metabolic genes

Numerous AMGs were identified in the viral scaffolds sequenced in this study. Many of the abundant cyanophage scaffolds encoded AMGs previously identified in marine viruses (6, 12, 13), such as photosystem genes *psbA* and *psbD*, plastoquinol oxidase, fatty acid desaturase, ribonucleotide reductase, glycine dehydrogenase, cobalt chetalase, and cytitidyltransferase. These genes were also highly expressed in the transcriptomes. It has been hypothesized that the activity of these genes manipulates host physiology in such a way to promote virion production (6, 14), but the full scope and precise activities of some of these AMGs remains enigmatic. The high expression of *psbA*, orders of magnitude higher than other AMGs, should be interpreted with caution since this is likely a result of the high similarity of viral gene copies with those of their host (15), leading to the mapping of host transcriptome reads to the viral copy.

One noteworthy finding was that of a *kaiC*-like regulator in VS2 (Fig. S4). KaiC is the core DNA-binding regulator of the cyanobacterial circadian clock, which is responsible for widespread repression of genes during diel cycling (16–18). Although the VS2-encoded *kaiC* homolog is highly divergent from homologous proteins in the picocyanobacteria *Prochlorococcus* and *Synechococcus*, it is tantalizing to think that this regulator may be the result of an ancient horizontal gene transfer from a cyanobacteria and now be responsible for the manipulation of host physiology during infection, possibly by shutting down host pathways not necessary for virion production. Under this scenario it would be predicted that the viral *kaiC* copy would be divergent from the host homolog, since operation of this gene within the established diel cycle of the host (including host-driven regulation of protein activity) would not necessarily be beneficial to the virus. Nevertheless, there is no other reason to suggest the host of VS2 is a cyanobacteria, and further work is needed to clarify the function of its encoded *kaiC*-like regulator.

Supplementary Methods

Field sampling

Samples were collected between 25 July and 5 August, 2015 in the North Pacific Subtropical Gyre (NPSG) during the Hawaii Ocean Experiment Legacy II cruise (KM1513). During this cruise samples were collected within the same water mass with an anticyclonic eddy by employing a Lagrangian drift strategy facilitated by the deployment of drogues (drifters) with a maximum depth of 15 m (Fig. 1a). Detailed information regarding the sampling regime on the cruise has been described in a previous study (19). General cruise information and associated biogeochemical and oceanographic measurements can be found online (hahana.soest.hawaii.edu/hoelegacy/hoelegacy.html).

Samples were collected every four hours during two periods of diel measurements from 26-30 July and 31 July-3 August, yielding samples from a total of 44 time-points. Water-column sampling was conducted at a depth of 15 m using a Niskin bottle rosette attached to a conductivity-temperature- depth (CTD) package (SBE 911Plus, SeaBird). At each time-point two replicate samples of 2 L of seawater were filtered through a 25 mm 0.2 µm Supor PES Membrane Disc filters (Pall, USA) housed in Swinnex units using a peristaltic pump. Filtrates from these samples were subsequently filtered through 25 mm 0.03 µm Supor PES Membrane Disc filters. This sampling strategy resulted in the acquisition of 3 samples total from each time-point, two corresponding to what are referred to as the "cellular fraction" (> 0.2 µm) and one corresponding to what are referred to as the "viral fraction" (0.2 µm > 0.03 µm). Filtration time on the 0.2 µm filters ranged from 15-20 minutes, while time on the 0.03 µm filtrations ranged from 30-50 min. Immediately after collection all filters were placed in RNALater (Ambion, Grand Island, NY) and stored at -80°C until processing. One cellular fraction sample and the viral fraction sample were used for subsequent DNA extraction and metagenome construction, while the second cellular fraction sample was used for RNA extraction and transcriptome construction.

Metagenome and quantitative metatranscriptome sequencing

DNA extractions were performed as previously described (19). Briefly, filters were thawed on ice, the RNALater was removed, and 400 µl of sucrose lysis buffer was added (final concentrations: 40 mM EDTA, 50 mM Tris (pH 8.3), and 0.75 M sucrose). Cell homogenization was performed using a Tissue Lyser (Qiagen, Germantown, MD) programmed at 30 Hertz for two rounds lasting 1 min each. 100 µl of sucrose lysis buffer containing 0.5 mg ml^{-1} lysozyme

was added before incubating in a rotating hybrid oven at 37°C for 30 min. Afterwards, 50 µl of sucrose lysis buffer containing Proteinase K (0.8 mg ml^{-1}) was added, followed by the addition of 50 µl of 10% SDS. Samples were incubated in a rotating hybrid oven at 55°C for 2 hrs. DNA purification was robotically performed using Chemagen MSM I instrument with the Saliva DNA CMG-1037 kit (Perkin Elmer, Waltham, MA) and DNA quantification was determined using Picogreen dsDNA kit (Invitrogen, Waltham MA). For cellular fraction metagenomes 250 ng of gDNA was sheared using Covaris M220 to a target insert size of 550 bp based on manufacturer's recommendation using Microtube-50 AFA fiber tubes. Cellular fraction metagenomes were prepared for sequencing using Illumina's TruSeq Nano LT library preparation kit. Viral fraction metagenomes were prepared by sharing 30 ng of gDNA using a Covaris M220 to a target insert size of 350 bp based on manufacturer's recommendation with Microtube-15 AFA Bead tubes. Sheared gDNA was loaded into Illumina's Neoprep library automation instrument using a Neoprep compatible TruSeq Nano LT kit. RNA extractions were performed by removing RNALater followed by the addition of 300 µl of Ambion denaturing solution directly to the filter then vortexed for 1 min. Prior to purification, 750 µl of nuclease free water was added. Samples were robotically purified and DNase treated using Chemagen MSM I instrument with the tissue RNA CMG-1212A kit (Perkin Elmer, Waltham, MA). RNA quality was assessed using the Fragment Analyzer high sensitivity reagents (Advanced Analytical Technologies, Inc.) and quantified using Ribogreen (Invitrogen, Waltham MA). Metatranscriptomic libraries were prepared for sequencing with the addition of 5–50 ng of Total RNA to the ScriptSeq cDNA V2 library preparation kit (Epicentre, Chicago, IL).

Molecular standard mixtures used for quantitative transcriptomics were prepared and implemented as previously described (20). Briefly, fourteen standards were generated from DNA templates via T7 RNA polymerase in vitro transcription (IVT) using the MEGAscript High Yield Transcription Kit (Ambion). DNA templates were generated directly from the genome of *Sulfolobus solfataricus* through PCR amplification and incorporation of the T7 promoter. Prior to RNA purification, 50 µl of each standard group was added to the sample lysate targeting a final standard concentration of approximately 1% to each sample based on expected total RNA yield. Metagenomic and metatranscriptomic samples were sequenced with an Illumina Nextseq500 system using V2 high output 300 cycle reagent kit with PHIX control added for metagenomic (1%) and for metatranscriptomic (5%) libraries. Both cellular fraction metagenomes, viral fraction metagenomes, and transcriptomes were multiplexed on two runs each. Dataset S1 contains details on the raw data generated in this manner.

Metagenome assembly

Each of the 44 metagenomes from the cellular and viral size fractions (88 samples total) were assembled individually using Mira v. 4.9.5_2 (21) with parameters "-AS:nop=6:sd=yes, -CL:pec=yes:spx174=yes:fpx174=yes:qc=yes", similar to previous methods (22). Two of the viral fraction metagenomes (S17 and S69) were too large for individual assembly, and for these the raw reads were split into groups of 10 million reads, each of which was subsequently assembled individually using Mira and then pooled. From the resulting assemblies contigs were analyzed using VirSorter (23) as provided on the CyVerse infrastructure (24, 25), with contigs < 3 Kb in length excluded as per tool guidelines. All reads used to assemble contigs from VirSorter categories 1 and 2 were then pooled and re-assembled using SPAdes v. 3.7 using the metagenome assembly option (26). All contigs > 3 Kbp in length from this assembly were then analyzed with VirSorter a second time, with those contigs annotated in categories 1 and 2 retained as putative viral contigs. Scaffolding was then performed using SSPACE 3.0 (27), and the resulting scaffolds used in downstream analyses.

Viral scaffold annotation

Genes for all metagenomes were predicted using Prodigal v. 2.6 (28) (parameters -p meta and -c). Annotations of predicted proteins were performed through comparison to the Pfam (29) (v 30) and VOG (http://vogdb.org; downloaded April 1, 2017;) databases using the hmmsearch utility in HMMER3 (30) (hmmsearch algorithm, e-value cutoff of 10e-3 used), with best hit retained. Comparison to the EggNOG database (31) was also performed using the eggnog-mapper utility (32). Lastly, predicted proteins were analyzed by comparison to the NCBI non-redundant protein databases (ftp.ncbi.nlm.nih.gov/blast/db/FASTA/nr.gz) using BLASTP (33) (e-value 10e-5 cutoff). To annotation putative auxiliary metabolic genes (AMGs) all protein annotations were analyzed manually.

Scaffold relative abundance estimates and VC ratio calculation

Abundance of the viral scaffolds in the cellular and viral fraction metagenomes was estimated through read mapping and normalization by total base pairs present in each metagenome. Cleaned reads provided by Mira after assembly were sorted using the "repair.sh" utility in BBMap (https://github.com/BioInfoTools/BBMap) and then mapped using the mem utility in BWA (34) with default parameters. Reads mapping with alignments < 45 bp in length and 95% identity were removed using msamtools (35), broadly consistent with previous methods (22). SAM and BAM files were processed using SAMtools (36), and coverage estimates

were generated using the genomecov utility in BEDTools (37). To account for library size, coverage estimates were divided by the total base-pairs sequenced in each metagenome. Because this procedure leads to values much smaller than 1, all relative abundance estimates were then multiplied by 10^{10} for convenience in downstream processing. These relative abundance values are provided in Dataset S2. The same methods were employed for the calculation of the abundance of individual viral genes, but with a fasta file of viral genes used for mapping using BWA. These relative abundance values are provided in Dataset S5. To estimate the relative abundance of scaffolds in different size fractions, a ratio of the relative abundance of scaffolds in the viral fraction vs. the cellular fraction was calculated (termed the VC ratio here). These values are provided in Dataset S2. Sample S69 was omitted in visualizations (for example, in Fig. 1a) given the amount of sequencing for this viral fraction metagenome was considerably smaller than in other samples (see Dataset S1 for details), and the zero values obtained for many scaffolds were not considered useful for determining VC ratios.

Assignment of putative taxonomy and hosts to viral scaffolds

To assign viral scaffolds to a putative taxonomic lineage annotations from the Pfam, EggNOG, VOG, and NCBI NR databases were analyzed manually. As a general rule, scaffolds with >20% of their encoded proteins matching to a particular reference genome were assigned a putative taxonomy of that reference. Exceptions to this general rule were made when marker genes indicative of particular viral groups were identified, in which case scaffolds with fewer than 20% of their encoded proteins matching to a reference were classified using these markers. These markers included baseplate wedge proteins, usually indicative of *Myoviridae*, or T7-like DNA polymerase, usually indicative of *Podoviridae*, consistent with previous studies (38–44). Additionally, the VOG database provides taxonomic resolution for each viral protein family, and scaffolds with > 2 best hits to a single lineage were classified accordingly. For putative host assignments, scaffolds with homology to known viruses (>20% of their encoded proteins having hits) were assigned the putative hosts of their references, and AMGs annotated as photosynthetic proteins were used for assignment of cyanobacterial hosts.

Comparison of scaffolds to reference genomes and contigs

To identify if the viral scaffolds were similar to genome references or contigs sequenced in previous metagenomic surveys a unidirectional average amino acid identity (AAI) comparison was performed against a database of known sequences. The database contained all sequenced

DNA viruses available in the NCBI RefSeq database (45) as of March 1st, 2016, together with the viral contig datasets sequenced in the Earth Virome (46), Global Ocean Virome (47), VirSorter (48), uvMED (49), and uvDeep (50) datasets. Proteins from all of these datasets was predicted using Prodigal v. 2.60 using the "-p meta" option with the exception of the complete genomes in NCBI, for which the available protein predictions were used. Proteins predicted from the scaffolds were queries against this database using LAST v. 756 (51), with hits with bit scores > 50 retained. A reference genome/contig was considered a match if the alignment fraction (AF), or percent of proteins in the queried scaffold having best hits to the same reference, was > 50. This is generally consistent with the methodology outlined in the Earth Virome study (46), where it was found that genomes with >90% AAI and >50% AF best recapitulated known viral species designations assigned by the International Committee on the Taxonomy of Viruses (ICTV). Full results are available in Dataset S4.

Select viral scaffolds for which the best hits of >20% of their predicted proteins were assigned to the same reference were analyzed further using a whole-genome alignment approach (Fig. S3). Scaffolds were aligned to their best reference genome using the "promer", "delta-filter", and "mummerplot" utilities in the MUMmer tool (52).

Scaffold binning

The scaffolds VS1, VS6, VS12, and VS153 were binned together (here this bin is referred to as Viral Group 1, or VG1) based on their combined whole-genome alignment with *Prochlorococcus* phage P-RSM1 (Figure S2). To confirm that this binning was warranted independent binning based on tetranucleotide frequencies (TNF) and co-abundance was also performed, as these characteristics have been shown to be useful in the phylogenetic binning of prokaryotic genomes (53). To this end weighted Pearson correlations of scaffolds was calculated based on a combined co-abundance profile of both cellular and viral fraction abundances (88 samples) as well as TNF for each of the scaffolds. Co-abundance and TNF metrics were weighted equally using the R package "weights" (https://cran.r-project.org/web/packages/weights/index.html). Clustering was then performed by converting the Pearson correlation values to distances by subtracting from 1 and generating a complete linkage clustering dendrogram in the R package "hclust". VS1, VS6, VS12, and VS153 clustered together in this dendrogram, consistent with their similar abundance profiles and TNF, confirming results of the whole-genome alignment to Cyanophage P-RSM1 (code and dendrogram available online at https://github.com/faylward/CSHLII.)

The four scaffolds VS1, VS6, VS12, and VS153 were the only scaffolds for which non-overlapping whole-genome alignment to a reference phage genome was observed, and so binning was only performed on these scaffolds. General binning of viral scaffolds based on TNF and co-abundance profiles was not performed since it is largely unknown how well these binning metrics will apply to fragmented viral genomes in metagenomic datasets, or what parameters are most appropriate. A great deal of effort has recently been focused on the development of new tools and appropriate parameters for the binning of bacterial and archaeal sequences (54–58), and in the future extension of these efforts to include viral genomes will be an important advance.

Transcriptome analysis

The initial processing of the 44 transcriptomes, quantification of the molecular standard spike-ins, and normalization transcript abundances was done in a manner identical to that previously reported (19). Briefly, reads were trimmed using Trimmomatic v. 0.27 (parameters: ILLUMINACLIP::2:40:15) (59), end-joined using PandaSeq v. 2.4 (parameters: -F -6 -t 0.32, quality cutoff of 0.32)(60), and quality-filtered using sickle v. 1.33 (length threshold set to 50). Reads corresponding to rRNA were then removed using sortmerna v. 2.1 (61) to obtain a final set of non-rRNA reads for each sample. For viral analyses non-rRNA reads were mapped against genes predicted from the viral scaffolds analyzed in this study using LAST (default parameters) with hits having ≥ 90% identity retained.

For quantitative normalization, non-rRNA reads were mapped to the standards using LAST. Five standards with consistent results within each time-point were used for calculating normalization coefficients (20) (standards S3, S5, S6, S10, and S11). For each time-point the average normalization coefficient for these five standards was multiplied by the reads mapped to each viral gene in that sample to derive estimates of transcripts per liter. This normalized count table was used for subsequent bioinformatic analyses. For aggregate transcriptional profiles of viral scaffolds all reads mapping to genes on a given scaffold were added. For VS399 reads mapping to the putative *psbA* homolog were not used on subsequent analyses because of the similarity of phage and host gene copies, making it unclear if the mapped reads were of viral origin.

To assess if total viral transcripts were more abundant in the evening samples a Mann-Whitney U test was performed in R using the wilcox.test function. For this test the estimated total viral transcripts per liter of the afternoon and evening timepoints (1400, 1800, and 2200 hrs) were compared to the morning timepoints (0200, 0600, and 1000 hrs). Thresholds of

detection for the quantitative transcriptomes (as shown in Fig. 2b) represent the abundance that would be calculated for a single transcript mapping. This is generally consistent with previous methods (62).

Fragment recruitment analyses

To evaluate the inter-annual presence and variability of viral populations, reads from a metagenomic time-series conducted at Sta. ALOHA were mapped against the viral scaffolds assembled in this study (Fig. 2b, Fig. S9). The metagenomic datasets correspond to the 1.6 > 0.2 µm size fraction of samples collected at a depth of 25m from 2010-2011 (22). Reads quality-trimmed with Mira (same procedures as outlined above) were mapped using both BLASTn (nucleotide identity) and BLASTx (amino acid identity) with an e-value threshold of 1e-5 used. Translated queries in BLASTx typically produce a large number of divergent hits which correspond to conserved proteins or domains; our analysis sought to identify only if sequences with high amino acid identity were present in the metagenomes, and to remove these divergent sequences only the upper quartile of amino acid identity hits were considered further. Nucleotide identity searches do not recover the same high abundance of divergent hits, and so for BLASTn queries all hits were retained for subsequent analyses. Abundance estimates reported correspond to the percent of reads mapping to a given reference, normalized by scaffold length.

Diel periodicity analyses

In the metagenomic datasets (viral and cellular fraction time-series) tests for diel periodicity were performed on the relative abundance estimates for viral scaffolds in both the viral and cellular fraction time series. For the quantitative transcriptomic datasets tests were performed on both the aggregate scaffold and individual gene transcriptomic profiles (units of transcripts/L in both cases). Because many viral scaffolds and genes had few or no reads mapping in the transcriptomic datasets, we pre-filtered out low abundance scaffolds (those with < 2 reads mapping per time-point, on average). Pre-filtering low-abundance entries has been shown to increase the statistical power of tests (63), and this general approach is implemented broadly in transcriptomic workflows (64). For periodicity tests the algorithm Rhythmicity Analysis Incorporating Non-Parametric Methods was used (implemented in the RAIN package (65) in R). Resulting p-values corrected using the Benjamini and Hochberg method (66) and corrected p-values < 0.1 considered significant, broadly consistent with previous methods (10). Because RAIN does not provide interpolated peak time estimates in

discrete time series, harmonic regression was implemented in base R for this purpose, similar to previous methods (10, 11).

Identification of Prochlorococcus *DNA replication time*

The growth rate estimations were performed on *Prochlorococcus* sp. MIT 0604 (CP007753) after a trimming step in order to remove genomic islands not present in metagenomes from HOE legacy cruise II. Reads from each 44 metagenomes were mapped on the original genome using default parameters of Bowtie2 (67). After the merge of resulting BAM files using SAMtools (36), the mean coverage over 1 kbp sliding windows and over the whole genome was calculated using BEDTools (68). Windows with mean coverage lower than 5% of the whole genome median coverage was excluded from the reference genome for use in these analyses. Twelve percent of *Prochlorococcus* sp. MIT 0604 reference genome was excluded from the analyses using this procedure. The trimmed genome was then used to estimate the peak-to-trough ratio (bPTR) and the index of replication (iRep) as previously described (69).

To identify transcriptional patterns in *Prochlorococcus* marker genes associated with DNA replication and cell division we queried protein predictions from our metagenomic datasets with the *Prochlorococcus* sp. MIT 0604 proteins DnaA (chromosomal replication initiation protein), DnaB (DNA helicase) DnaE (DNA polymerase III), PolA (DNA polymerase I), and FtsZ (Cell division protein) using Blastp (35) (default parameters) using a 95% amino acid identity threshold. Reads from our transcriptomes were then mapped (95% identity threshold) and normalized as described above. For PCA analyses a distance matrix of replication times (iRep and bPTR), *Prochlorococcus* marker gene transcript abundances, and diel cyanophage transcriptional abundances was generated by subtracting Pearson correlations from 1. The PCA analysis was performed using *XLStat version 2009.1.02.*

Supplementary Tables

Dataset 1. Sequencing statistics for the metagenomes and metatranscriptomes used in this study.

Dataset 2. Coverage values for the viral scaffolds in all of the metagenomes analyzed in this study.

Dataset 3. Annotations and statistics for the viral scaffolds analyzed in this study, and results for all scaffold-based diel analyses.

Dataset 4. Comparison of viral scaffolds to reference genomes and contigs, together with the results of read mapping from HOT metagenomes.

Dataset 5. Annotations for the viral genes on the viral scaffolds, together with results of all gene-based diel analyses.

Supplementary References

1. Wilcox RM, Fuhrman JA (1994) Bacterial viruses in coastal seawater: lytic rather than lysogenic production. *Mar Ecol Prog Ser* 114:35–45.

2. De Corte D, Sintes E, Yokokawa T, Reinthaler T, Herndl GJ (2012) Links between viruses and prokaryotes throughout the water column along a North Atlantic latitudinal transect. *ISME J* 6(8):1566–1577.

3. Doron S, et al. (2016) Transcriptome dynamics of a broad host-range cyanophage and its hosts. *ISME J* 10(6):1437–1455.

4. Campbell L, Vaulot D (1993) Photosynthetic picoplankton community structure in the subtropical North Pacific Ocean near Hawaii (station ALOHA). *Deep Sea Res Part I* 40(10):2043–2060.

5. Bryant JA, et al. (2016) Wind and sunlight shape microbial diversity in surface waters of the North Pacific Subtropical Gyre. *ISME J* 10(6):1308–1322.

6. Thompson LR, et al. (2011) Phage auxiliary metabolic genes and the redirection of cyanobacterial host carbon metabolism. *Proc Natl Acad Sci U S A* 108(39):E757–64.

7. Lindell D, Jaffe JD, Johnson ZI, Church GM, Chisholm SW (2005) Photosynthesis genes in marine viruses yield proteins during host infection. *Nature* 438(7064):86–89.

8. Suttle CA, Chen F (1992) Mechanisms and rates of decay of marine viruses in seawater. *Appl Environ Microbiol* 58(11):3721–3729.

9. Brown CT, Olm MR, Thomas BC, Banfield JF (2016) Measurement of bacterial replication rates in microbial communities. *Nat Biotechnol* 34(12):1256–1263.

10. Ottesen EA, et al. (2014) Ocean microbes. Multispecies diel transcriptional oscillations in open ocean heterotrophic bacterial assemblages. *Science* 345(6193):207–212.

11. Aylward FO, et al. (2015) Microbial community transcriptional networks are conserved in three domains at ocean basin scales. *Proc Natl Acad Sci U S A* 112(17):5443–5448.

12. Crummett LT, Puxty RJ, Weihe C, Marston MF, Martiny JBH (2016) The genomic content and context of auxiliary metabolic genes in marine cyanomyoviruses. *Virology* 499:219–229.

13. Mann NH, Cook A, Millard A, Bailey S, Clokie M (2003) Marine ecosystems: Bacterial photosynthesis genes in a virus. *Nature* 424(6950):741–741.

14. Breitbart M (2012) Marine Viruses: Truth or Dare. *Ann Rev Mar Sci* 4(1):425–448.

15. Lindell D, et al. (2004) Transfer of photosynthesis genes to and from Prochlorococcus viruses. *Proc Natl Acad Sci U S A* 101(30):11013–11018.

16. Ishiura M, et al. (1998) Expression of a gene cluster kaiABC as a circadian feedback process in cyanobacteria. *Science* 281(5382):1519–1523.

17. Dong G, et al. (2010) Elevated ATPase Activity of KaiC Applies a Circadian Checkpoint on Cell Division in Synechococcus elongatus. *Cell* 140(4):529–539.

18. Cohen SE, Golden SS (2015) Circadian Rhythms in Cyanobacteria. *Microbiol Mol Biol Rev* 79(4):373–385.

19. Wilson ST, et al. (2017) Coordinated regulation of growth, activity and transcription in natural populations of the unicellular nitrogen-fixing cyanobacterium Crocosphaera. *Nature Microbiology* 2:17118.

20. Gifford SM, Becker JW, Sosa OA, Repeta DJ, DeLong EF (2016) Quantitative Transcriptomics Reveals the Growth- and Nutrient-Dependent Response of a Streamlined Marine Methylotroph to Methanol and Naturally Occurring Dissolved Organic Matter. *MBio* 7(6). doi:10.1128/mBio.01279-16.

21. Chevreux B, et al. (2004) Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. *Genome Res* 14(6):1147–1159.

22. Mende DR, et al. (2017) Environmental drivers of a microbial genomic transition zone in the ocean's interior. *Nat Microbiol*. doi:10.1038/s41564-017-0008-3.

23. Roux S, Enault F, Hurwitz BL, Sullivan MB (2015) VirSorter: mining viral signal from microbial genomic data. *PeerJ* 3:e985.

24. Merchant N, et al. (2016) The iPlant Collaborative: Cyberinfrastructure for Enabling Data to Discovery for the Life Sciences. *PLoS Biol* 14(1):e1002342.

25. Goff SA, et al. (2011) The iPlant Collaborative: Cyberinfrastructure for Plant Biology. *Front Plant Sci* 2:34.

26. Bankevich A, et al. (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 19(5):455–477.

27. Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W (2011) Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* 27(4):578–579.

28. Hyatt D, et al. (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 11:119.

29. Finn RD, et al. (2015) The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res* 44(D1):D279–D285.

30. Eddy SR (2011) Accelerated Profile HMM Searches. *PLoS Comput Biol* 7(10):e1002195.

31. Huerta-Cepas J, et al. (2016) eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res* 44(D1):D286–93.

32. Huerta-Cepas J, et al. (2017) Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. *Mol Biol Evol*. doi:10.1093/molbev/msx148.

33. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215(3):403–410.

34. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25(14):1754–1760.

35. Arumugam M, Harrington ED, Foerstner KU, Raes J, Bork P (2010) SmashCommunity: a metagenomic annotation and analysis tool. *Bioinformatics* 26(23):2977–2978.

36. Li H, et al. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25(16):2078–2079.

37. Quinlan AR (2014) BEDTools: The Swiss-Army Tool for Genome Feature Analysis. *Curr Protoc Bioinformatics* 47:11.12.1–34.

38. Mizuno CM, Rodriguez-Valera F, Garcia-Heredia I, Martin-Cuadrado A-B, Ghai R (2013) Reconstruction of novel cyanobacterial siphovirus genomes from Mediterranean metagenomic fosmids. *Appl Environ Microbiol* 79(2):688–695.

39. Huang S, Wang K, Jiao N, Chen F (2012) Genome sequences of siphoviruses infecting marine Synechococcus unveil a diverse cyanophage group and extensive phage-host genetic exchanges. *Environ Microbiol* 14(2):540–558.

40. Sullivan MB, et al. (2009) The genome and structural proteome of an ocean siphovirus: a new window into the cyanobacterial "mobilome." *Environ Microbiol* 11(11):2935–2951.

41. Sullivan MB, Coleman ML, Weigele P, Rohwer F, Chisholm SW (2005) Three Prochlorococcus Cyanophage Genomes: Signature Features and Ecological Interpretations. *PLoS Biol* 3(5):e144.

42. Dekel-Bird NP, et al. (2013) Diversity and evolutionary relationships of T7-like podoviruses infecting marine cyanobacteria. *Environ Microbiol* 15(5):1476–1491.

43. Schmidt HF, Sakowski EG, Williamson SJ, Polson SW, Wommack K (2013) Shotgun metagenomics indicates novel family A DNA polymerases predominate within marine virioplankton. *ISME J* 8(1):103–114.

44. Adriaenssens EM, Cowan DA (2014) Using signature genes as tools to assess environmental viral ecology and diversity. *Appl Environ Microbiol* 80(15):4470–4480.

45. O'Leary NA, et al. (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* 44(D1):D733–45.

46. Paez-Espino D, et al. (2016) Uncovering Earth's virome. *Nature* 536(7617):425–430.

47. Roux S, et al. (2016) Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature* 537(7622):689–693.

48. Roux S, Hallam SJ, Woyke T, Sullivan MB (2015) Viral dark matter and virus-host interactions resolved from publicly available microbial genomes. *Elife* 4. doi:10.7554/eLife.08490.

49. Mizuno CM, Rodriguez-Valera F, Kimes NE, Ghai R (2013) Expanding the marine virosphere using metagenomics. *PLoS Genet* 9(12):e1003987.

50. 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(4). doi:10.1128/mBio.00805-16.

51. Kiełbasa SM, Wan R, Sato K, Horton P, Frith MC (2011) Adaptive seeds tame genomic sequence comparison. *Genome Res* 21(3):487–493.

52. Kurtz S, et al. (2004) Versatile and open software for comparing large genomes. *Genome Biol* 5(2):R12.

53. Sedlar K, Kupkova K, Provaznik I (2017) Bioinformatics strategies for taxonomy independent binning and visualization of sequences in shotgun metagenomics. *Comput Struct Biotechnol J* 15:48–55.

54. Eren AM, et al. (2015) Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ* 3:e1319.

55. Alneberg J, et al. (2014) Binning metagenomic contigs by coverage and composition. *Nat Methods* 11(11):1144–1146.

56. Kang DD, Froula J, Egan R, Wang Z (2015) MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* 3:e1165.

57. Lin H-H, Liao Y-C (2016) Accurate binning of metagenomic contigs via automated clustering sequences using information of genomic signatures and marker genes. *Sci Rep* 6:24175.

58. Imelfort M, et al. (2014) GroopM: an automated tool for the recovery of population genomes from related metagenomes. *PeerJ* 2:e603.

59. Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30(15):2114–2120.

60. Masella AP, Bartram AK, Truszkowski JM, Brown DG, Neufeld JD (2012) PANDAseq: paired-end assembler for illumina sequences. *BMC Bioinformatics* 13:31.

61. Kopylova E, Noé L, Touzet H (2012) SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics* 28(24):3211–3217.

62. Gifford SM, Sharma S, Rinta-Kanto JM, Moran MA (2011) Quantitative analysis of a deeply sequenced marine microbial metatranscriptome. *ISME J* 5(3):461–472.

63. Bourgon R, Gentleman R, Huber W (2010) Independent filtering increases detection power for high-throughput experiments. *Proc Natl Acad Sci U S A* 107(21):9546–9551.

64. Rau A, Gallopin M, Celeux G, Jaffrézic F (2013) Data-based filtering for replicated high-throughput transcriptome sequencing experiments. *Bioinformatics* 29(17):2146–2152.

65. Thaben PF, Westermark PO (2014) Detecting rhythms in time series with RAIN. *J Biol Rhythms* 29(6):391–400.

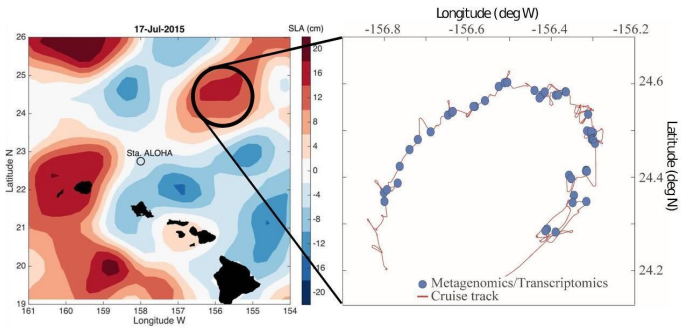
66. Yoav Benjamini YH (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Series B Stat Methodol* 57(1):289–300.

67. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nat Methods* 9(4):357–359.

68. Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26(6):841–842.

69. Brown CT, Olm MR, Thomas BC, Banfield JF (2016) Measurement of bacterial replication rates in microbial communities. *Nat Biotechnol* 34(12):1256–1263.

a.



b.

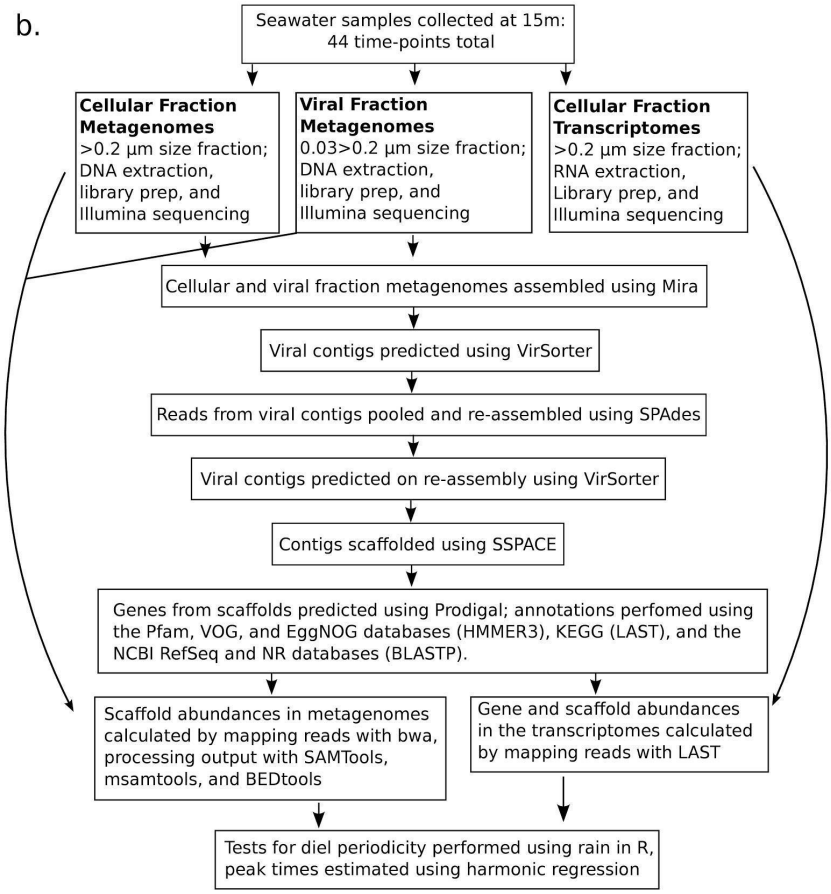


Figure S1. **Cruise overview and bioinformatic workflow.** a) Plot of Sea Level Anomaly (SLA) taken in the days before the start of the cruise (left), with the solid circle indicating the anticyclonic eddy targeted for sampling. The inset on the right shows the cruise track for the duration of the sampling period, with sample locations indicated with solid blue dots. Adapted from previously reported data[1]. b) Bioinformatic workflow used in this study for the assembly and annotation of dominant viral population genomes. See Supplementary Methods for details.

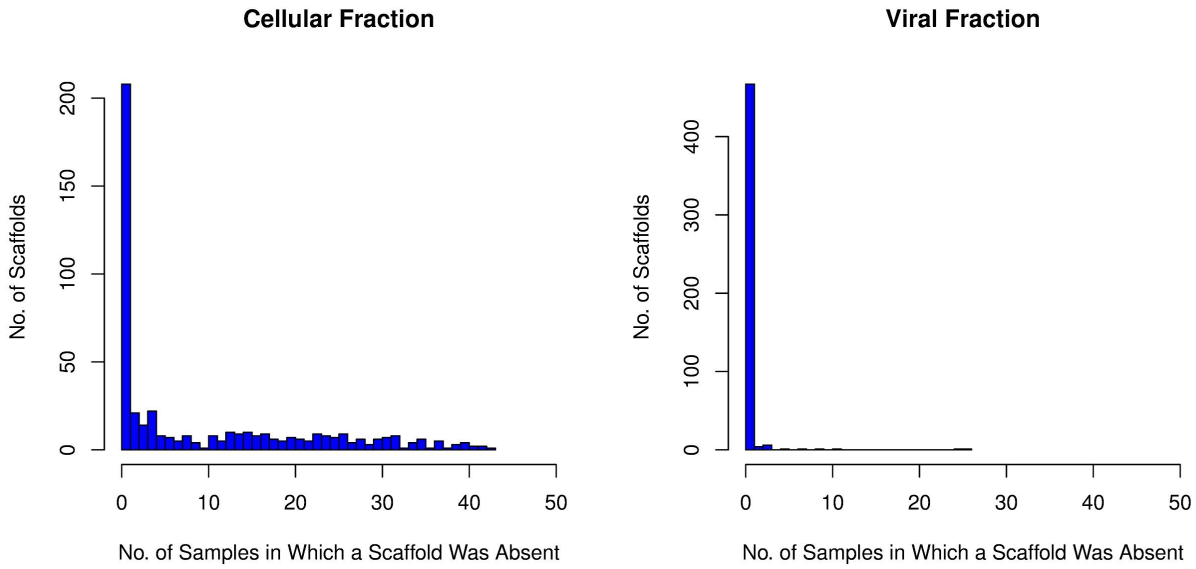


Figure S2. Occurrence table of the 483 viral genomic scaffolds identified in this study. The y-axis provides the number of viral genomic scaffolds that were not identified for a particular number of time-points (x-axis) throughout the 44-time-point time-course.

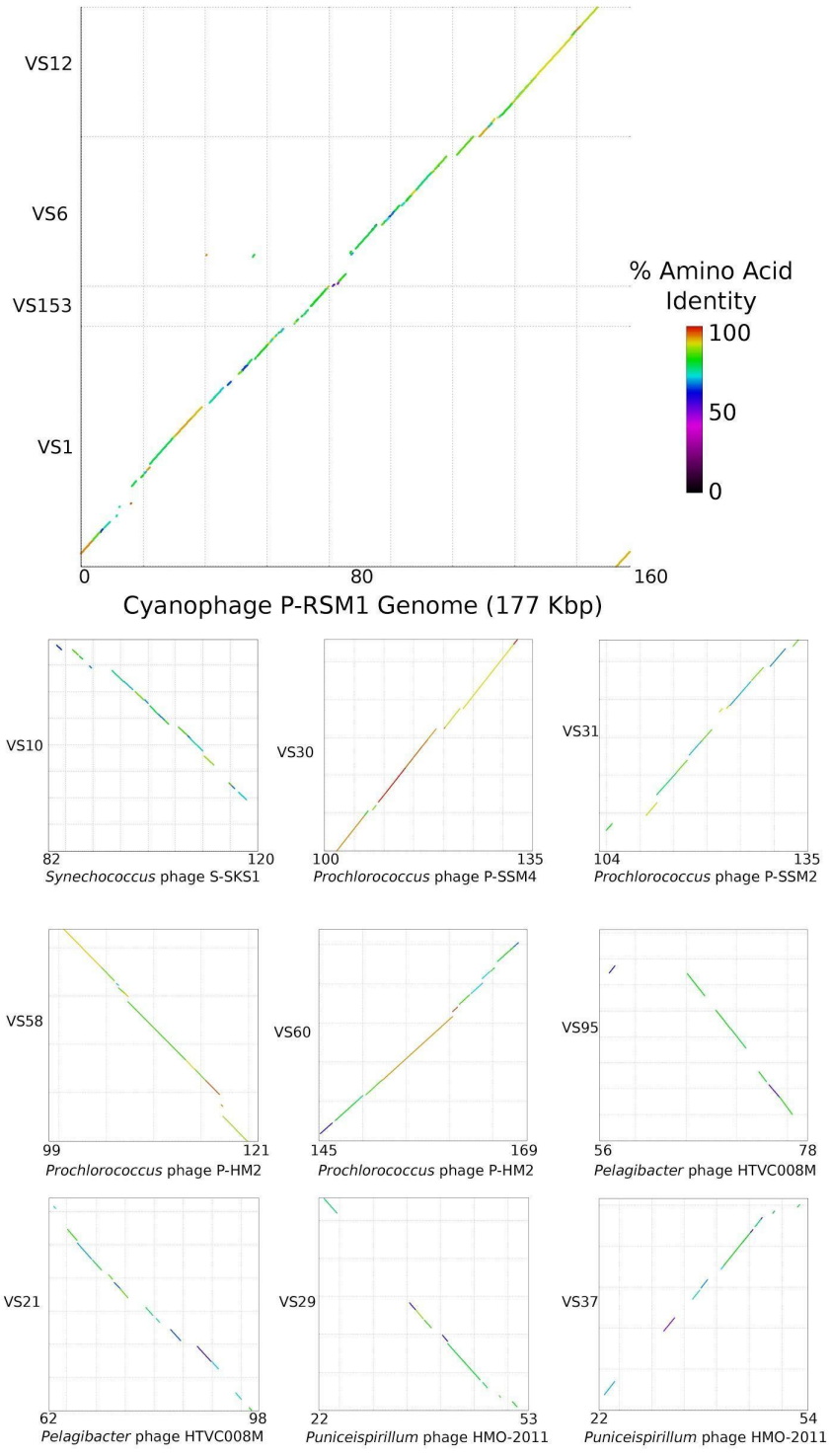


Figure S3. Synteny of viral scaffolds to reference viral genomes. Scaffold-to-reference genome alignment plots for select viral scaffolds with homology to sequenced viruses. The top plot shows Viral Group 1 (VG1), which is comprised of four scaffolds that combined span the entire genome of the Cyanophage P-RSM1. In each plot the viral scaffold coordinates are provided on the y-axis and reference genome coordinates are provided on the x-axis. Color indicates the percent amino acid identity of the aligned region.

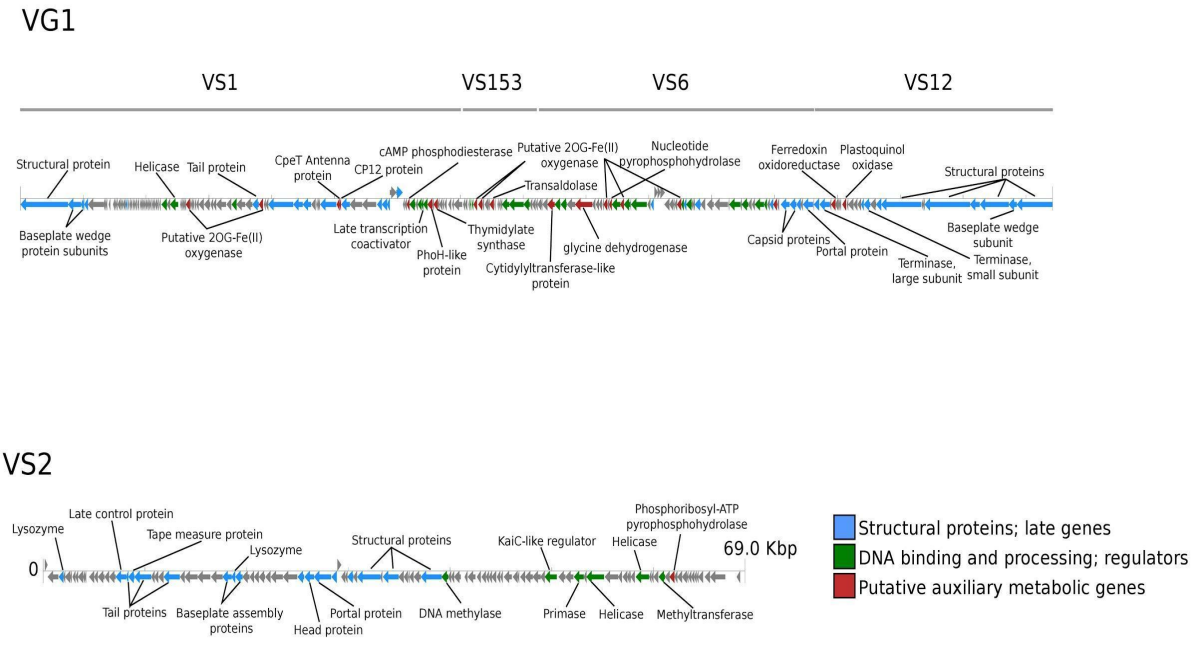


Figure S4. Genome annotation diagrams of the two most abundant viral groups identified in this study. VG1 (top) consists of the four scaffolds VS1, VS6, VS12, and VS153 (top), while VS2 (bottom) is represented by a single scaffold. Colors denote functional annotation categories.

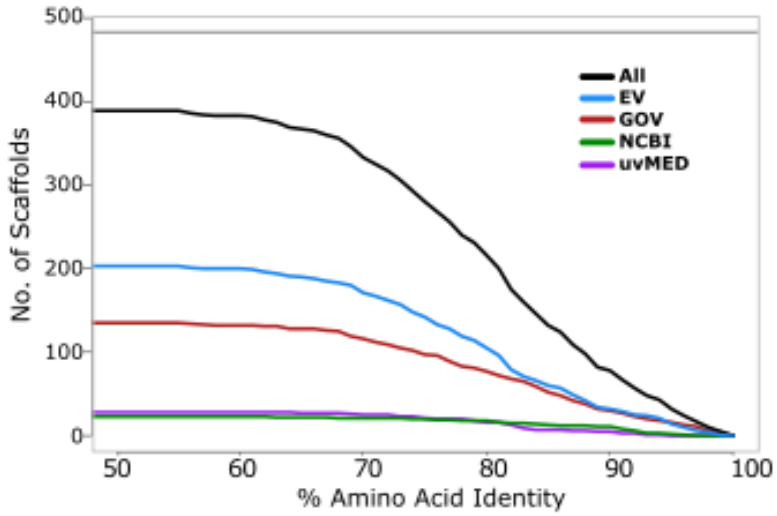


Figure S5. Classification of viral scaffolds database of sequenced viral genomes and viral metagenomic scaffolds. The y-axis gives the number of scaffolds that could be classified at a given AAI threshold, given on the x-axis. Different colors indicate results for different databases. Abbreviations: EV: Earth Virome; GOV: Global Ocean Virome; NCBI; NCBI Refseq genome collection; UVMed: Mediterranean DCM viruses. The upper grey line denotes the upper threshold of 483 scaffolds.

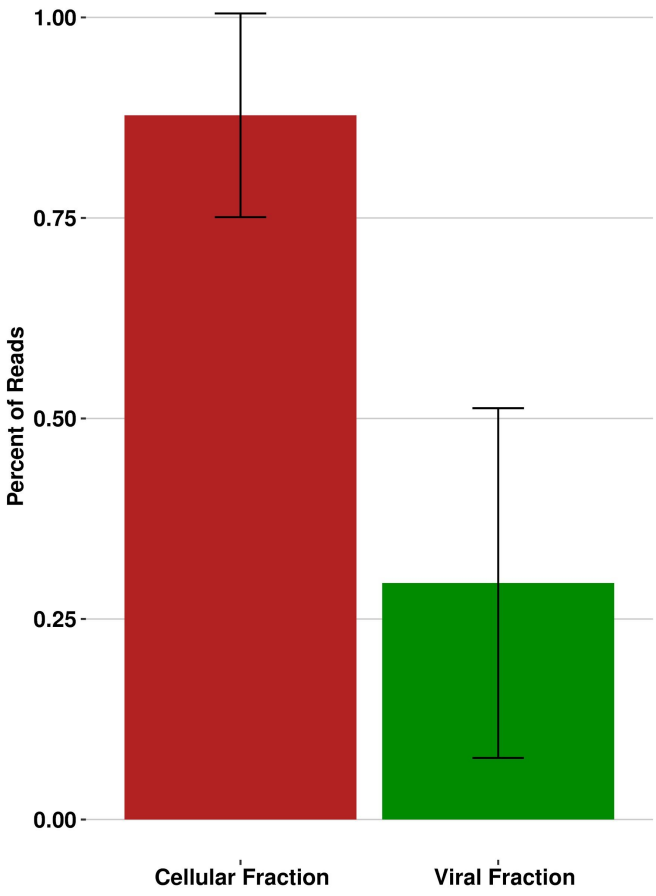


Figure S6. Comparison of rRNA content between different size fractions. Bars represent the average of the percent of rRNA reads identified in the 44 timepoints sequenced. Error bars denote standard deviation.

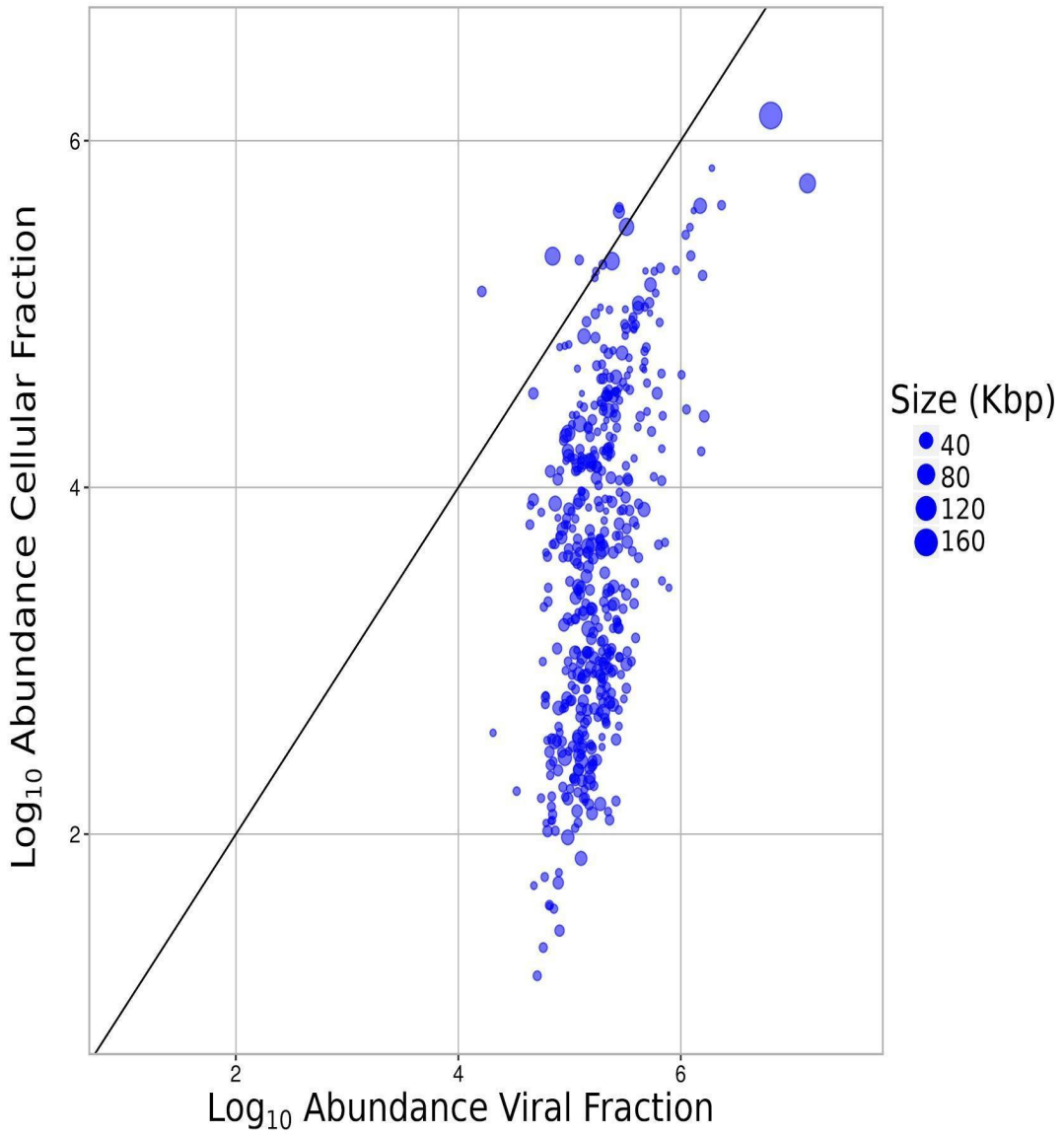


Figure S7. **Abundance of the viral scaffolds identified in this study in the viral fraction (x-axis) vs the cellular fraction (y-axis).** Dot size is proportional to the size of the scaffold. Relative abundance was calculated by normalizing the coverage of each scaffold by the total bp sequenced, and averaging across all time-points.

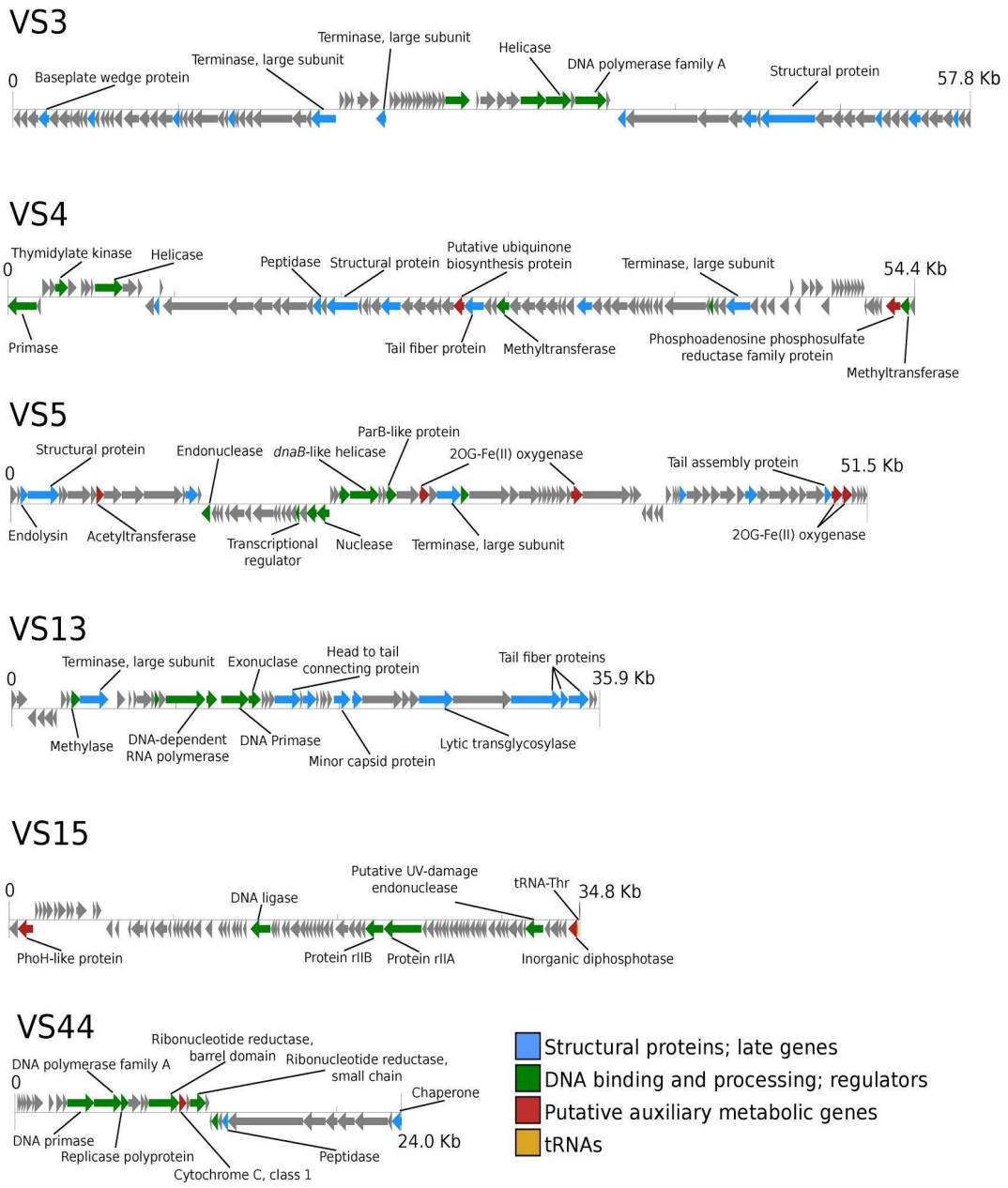


Figure S8. **Genome annotation diagrams of select viral scaffolds found to have high relative abundance in cellular fraction metagenomes compared to corresponding viral fraction metagenomes**. Colors indicate functional annotation categories.

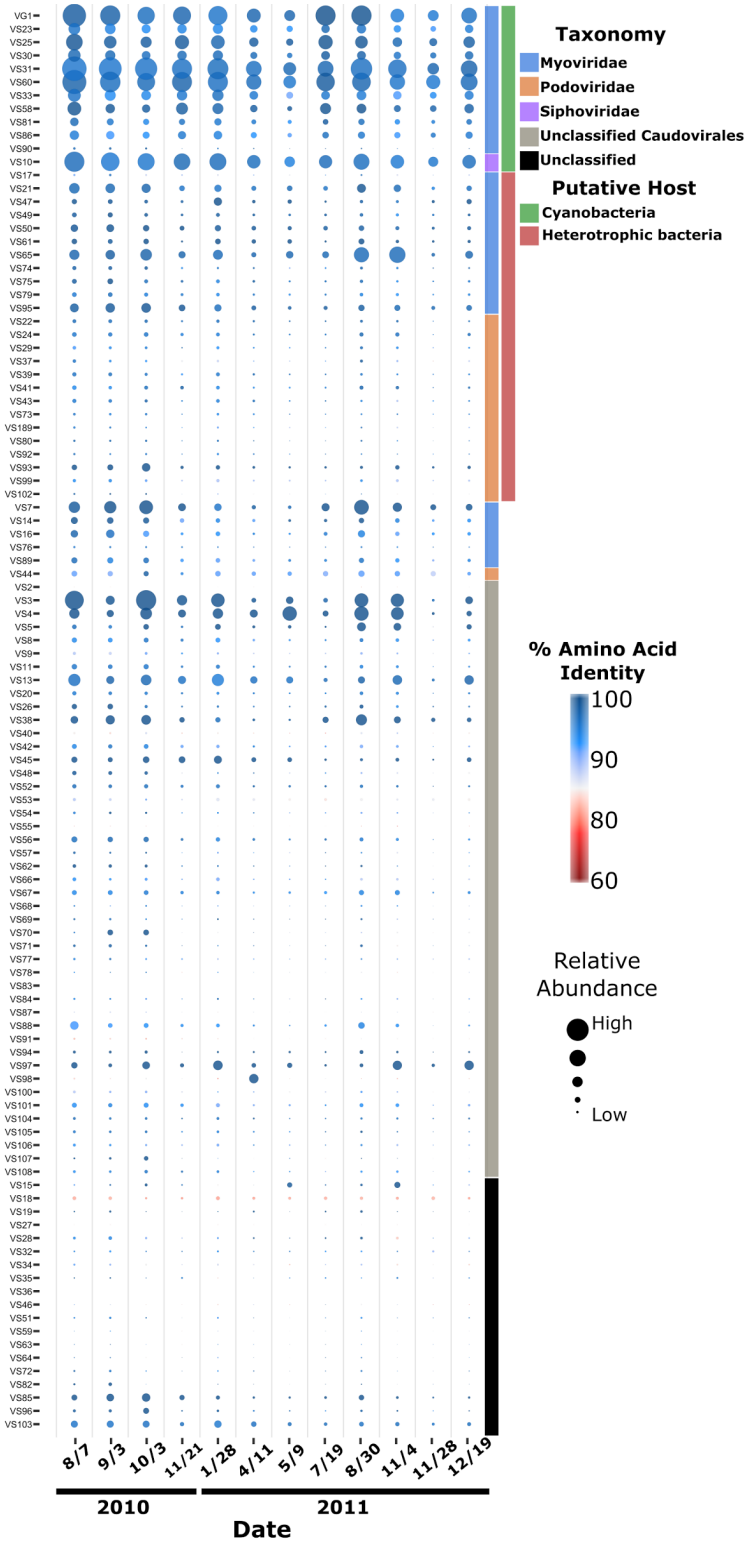


Figure S9. Inter-annual abundance of viral groups at Sta. ALOHA. Results of translated read-recruitment analyses (using BLASTx) of 12 metagenomes sequenced at 25m at Station ALOHA in 2010 and 2011 (dates on the x-axis) against viral scaffolds assembled in this study (y-axis). Scaffold order and color bars designating putative host and taxonomic assignments are identical to those in Fig. 1. The size of the dots is proportional to the relative abundance of hits in a given metagenome, while color denotes the average percent identity of those hits. Relative abundance estimates were normalized by library size and scaffold length.

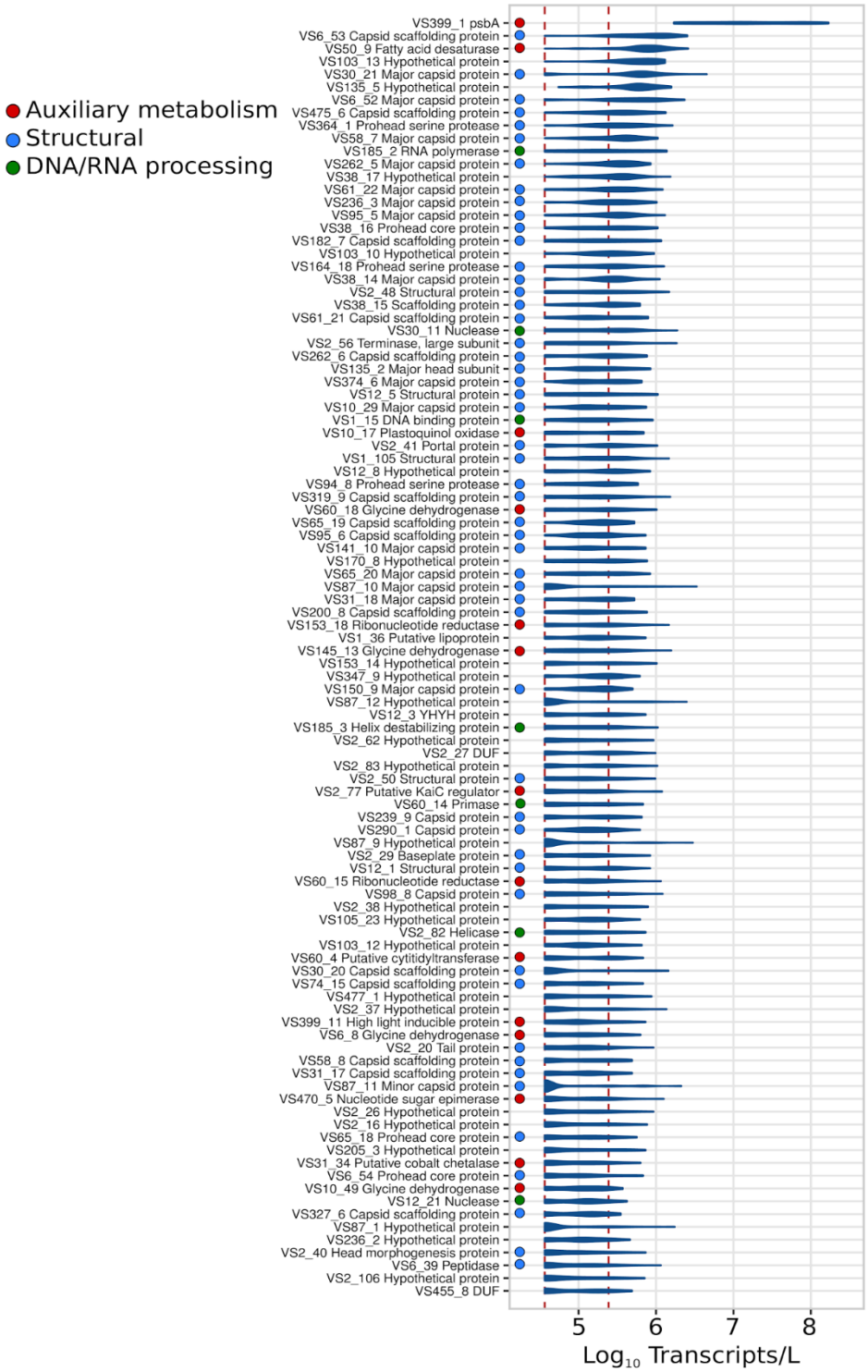


Figure S10. Most abundant viral transcripts identified in this study (y-axis, descending in order of median abundance). The x-axis gives transcripts per liter. Violin plots show the distribution of abundances for a given transcript. The red lines denote the thresholds of detection as in Figure 3. Colored dots denote functional annotation category. The high abundance of VS399_1 is likely due to the high similarity of phage and host *psbA* copies, and reads mapping to this gene were excluded from subsequent analyses.

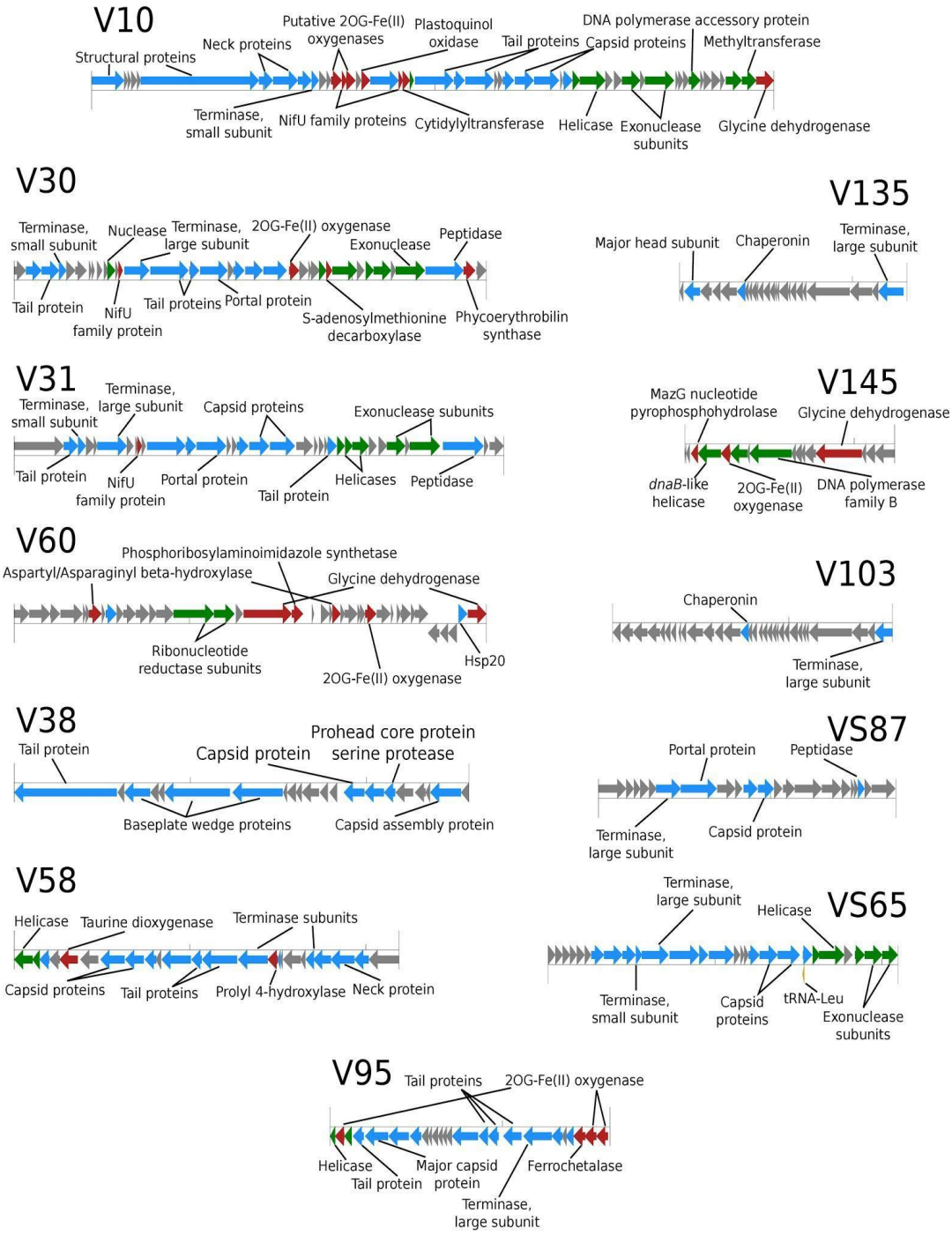


Figure S11. Genome diagrams of viral scaffolds abundant in the transcriptomes. Colors denote functional annotation categories.

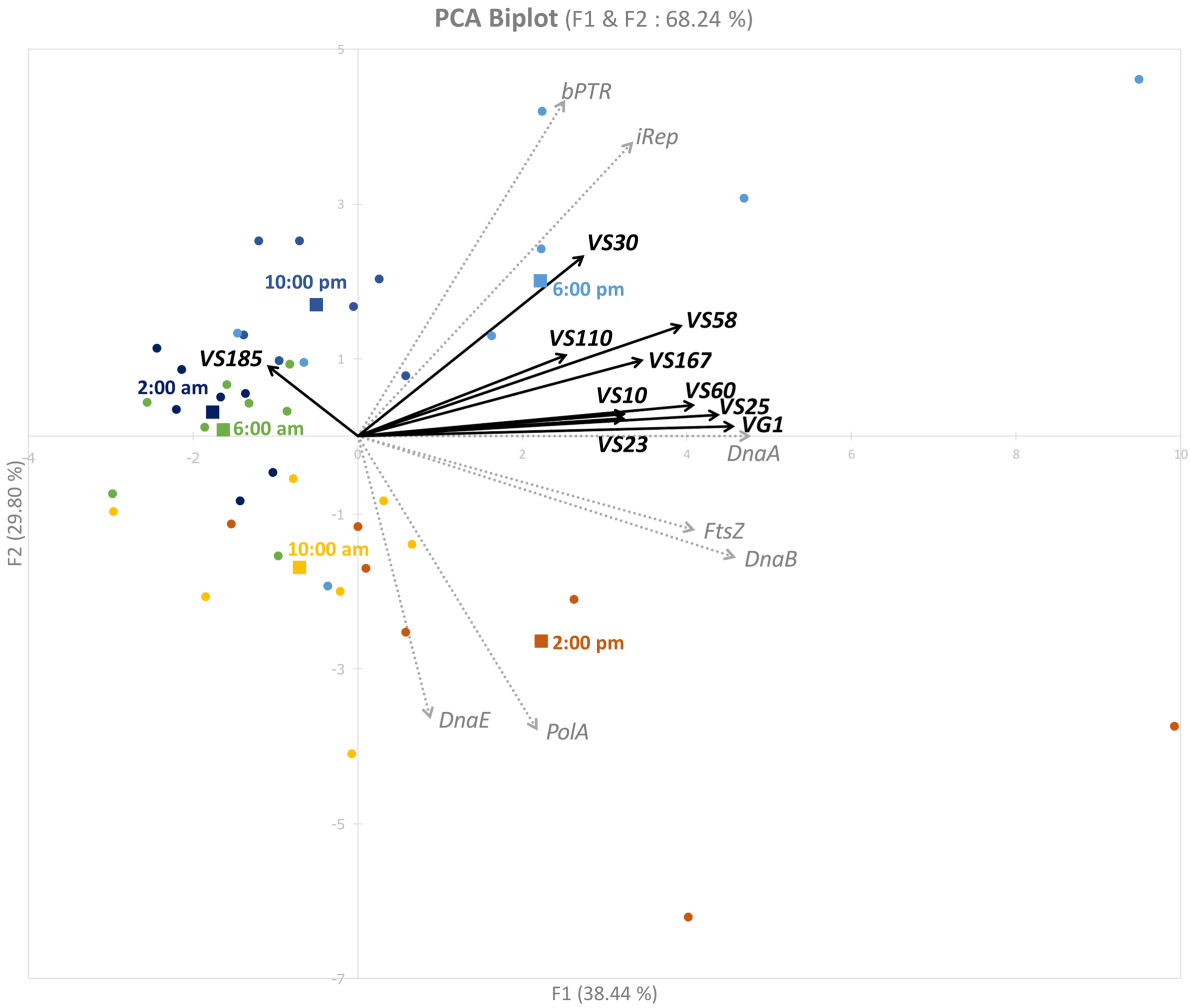


Figure S12. Synchrony of cyanophage activity and *Prochlorococcus* replication and division. PCA generated from diel cyanophage transcriptional profiles, transcriptional profiles of *Prochlorococcus* DNA replication and cell division marker genes, and two metrics for estimating the timing of *Prochlorococcus* DNA replication (iRep and bPTR). Arrows indicate the direction and magnitude of variables. The 44 time points used as observations are displayed by circles colored by the sampling time, with squares representing centroids. b) Temporal profiles for the data used in panel a. Units for the transcriptional profiles are X 10^5 transcripts/L.

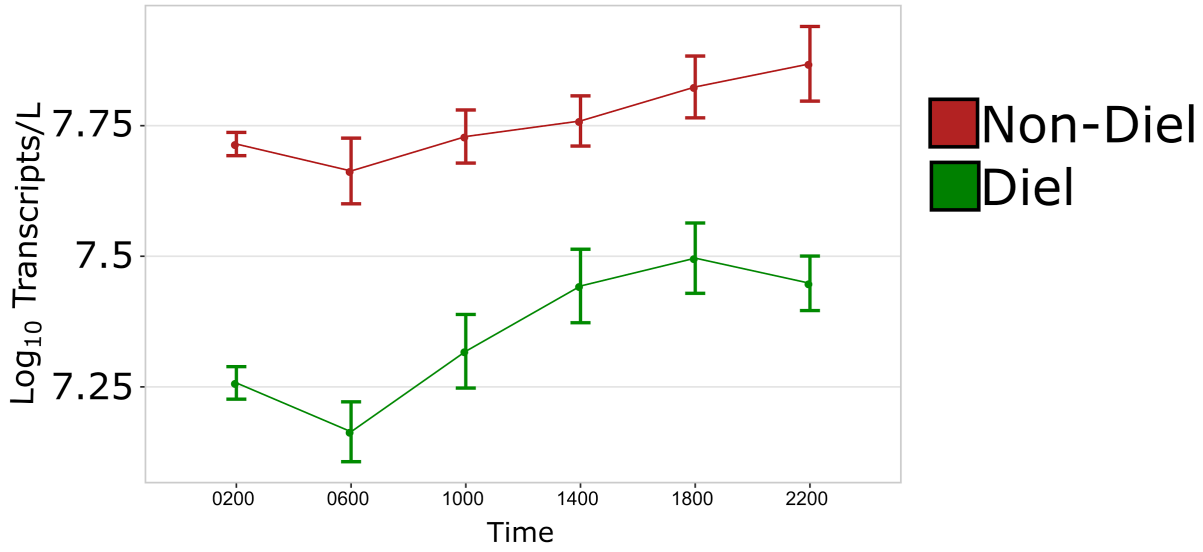


Figure S13. Abundance of transcripts mapping to viral scaffolds with significantly diel expression patterns (green) or no detectable diel periodicity in expression (red). Error bars represent standard error.