bioRxiv preprint doi: https://doi.org/10.1101/2024.08.12.607631; this version posted August 15, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Phylogenetic proximity drives temporal succession of marine giant viruses in a five-year metagenomic time-series

Sarah M. Laperriere[1†*], Benjamin Minch[2†], JL Weissman[1,3,4], Shengwei Hou[1,5]*, Yi-Chun Yeh[1], J. Cesar Ignacio-Espinoza[1], Nathan A. Ahlgren[6], Mohammad Moniruzzaman[2]*, and Jed A. Fuhrman[1]*

1 Department of Biological Sciences, University of Southern California, Los Angeles, California, USA
2 Department of Marine Biology and Ecology, Rosenstiel School of Marine, Atmospheric, and Earth Sciences, University of Miami, Miami, FL, USA

3 Department of Ecology and Evolution, Stony Brook University, Stony Brook, NY, USA

4 Institute for Advanced Computational Science, Stony Brook University, Stony Brook, NY, USA

5 Department of Ocean Science and Engineering, Southern University of Science and Technology, Shenzhen 518055, China

6 Department of Biology, Clark University, Worcester, MA USA

†These authors contributed equally.

* To whom correspondence may be addressed: fuhrman@usc.edu, m.monir@miami.edu, housw2010@gmail.com

Abstract

Nucleocytoplasmic Large DNA Viruses (NCLDVs, also called giant viruses) are widespread in marine systems and infect a broad range of microbial eukaryotes (protists). Recent biogeographic work has provided global snapshots of NCLDV diversity and community composition across the world's oceans, yet little information exists about the guiding 'rules' underpinning their community dynamics over time. We leveraged a five-year monthly metagenomic time-series to quantify the community composition of NCLDVs off the coast of Southern California and characterize these populations' temporal dynamics. NCLDVs were dominated by Algavirales (Phycodnaviruses, 59%) and Imitervirales (Mimiviruses, 36%). We identified clusters of NCLDVs with distinct classes of seasonal and non-seasonal temporal dynamics. Overall, NCLDV population abundances were often highly dynamic with a strong seasonal signal. The Imitervirales group had highest relative abundance in the more oligotrophic late summer and fall, while Algavirales did so in winter. Generally, closely related strains had similar temporal dynamics, suggesting that evolutionary history is a key driver of the temporal niche of marine NCLDVs. However, a few closely-related strains had drastically different seasonal dynamics, suggesting that while phylogenetic proximity often indicates ecological similarity, occasionally

bioRxiv preprint doi: https://doi.org/10.1101/2024.08.12.607631; this version posted August 15, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

phenology can shift rapidly, possibly due to host-switching. Finally, we identified distinct functional content and possible host interactions of two major NCLDV orders- including connections of Imitervirales with primary producers like the diatom *Chaetoceros* and widespread marine grazers like *Paraphysomonas* and Spirotrichea ciliates. Together, our results reveal key insights on season-specific effect of phylogenetically distinct giant virus communities on marine protist metabolism, biogeochemical fluxes and carbon cycling.

Introduction

Viruses are the most abundant and genetically diverse entities in marine systems [1, 2]. Orders of magnitude more abundant than prokaryotes and eukaryotes, viruses exert significant influence on ecosystem processes, such as structuring microbial communities and altering biogeochemical cycles [3, 4, 5]. Much of what is known about marine viruses stems from studying viruses that infect bacteria and a substantial knowledge gap exists regarding viruses that infect marine eukaryotes or how these eukaryotic-virus interactions modulate ecosystem processes. Globally, eukaryotes contribute significantly to ocean biomass and primary production [6, 7, 8], suggesting that eukaryotic viruses have the potential to exert significant control on marine systems.

Many of the viruses that infect marine eukaryotes belong to the phylum Nucleocytoviricota and are referred to as nucleocytoplasmic large DNA viruses (NCLDVs; [9, 10, 11]). This diverse monophyletic group is composed of several orders including Algavirales (Phycodnaviruses), Imitervirales (Mimiviruses), Pimascovirales, Pandoravirales, Chitovirales, and Asfuvirales, and infects many microbial eukaryotic lineages, including chlorophytes, haptophytes, dinoflagellates, choanoflagellates, and stramenopiles [12-18]. These viruses have large genomes up to 2.5 Mbp [19] harboring many genes for DNA replication, translation, and repair, as well as suites of genes implicated in the rewiring of host metabolic networks [20- 22, 12]. Virus-encoded genes involved in nutrient acquisition and utilization, photosynthesis, and carbon fixation alter the growth of eukaryotic hosts and potentially also the nutrient cycles in which these hosts are embedded [16, 12, 21-27].

NCLDVs are widely distributed throughout the ocean and display distinct latitudinal patterns in community structure that correlate with eukaryotic communities [12, 28-31]. Yet, with a few notable exceptions like the extensive work on viruses infecting the coccolithophore *E. huxleyi* [32-38], our understanding of these viral communities as a whole is based on spatial surveys and short-term studies, with little existing data on the long-term temporal dynamics of these viruses; the latter is crucial to assess their ecological succession and impact on the microbial eukaryote populations. [39]. Nevertheless, we know that many microbial eukaryote populations exhibit strong temporal patterns (e.g., [40-43]), suggesting that viral communities should also demonstrate similarly variable patterns in incidence and abundance. Patterns of seasonal dynamics have already been elucidated in prokaryotic viruses [44] but a better understanding of

bioRxiv preprint doi: https://doi.org/10.1101/2024.08.12.607631; this version posted August 15, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

the temporal dynamics of eukaryotic viruses and hosts is necessary to understand the impact of eukaryotic viruses on ecosystem processes.

We leveraged a multi-year, monthly metagenomic time series [44,45] to quantify the diversity of NCLDVs off the coast of Southern California and to characterize the temporal dynamics of these populations. Together, using this rich dataset we identified clusters of NCLDVs with distinct classes of seasonal and non-seasonal temporal dynamics and showed that closely related strains typically had similar dynamics. As an exception to this rule, in a select few cases, closely related strains had drastically different seasonal dynamics, suggesting that while phylogenetic proximity often indicates ecological similarity for marine NCLDVs, occasionally phenology can shift rapidly, possibly as a result of host-switching.

Methods

Sample Collection, DNA extraction, and sequencing

Surface seawater was collected monthly from 5 m as part of San Pedro Ocean Time-series (SPOT) (https://dornsife.usc.edu/spot/) using a 12 x12 Niskin bottle rosette directly into 20-L cubitainers. Seawater was filtered for viral metagenomes (< 0.22μm, n = 53) by serially passing 0.5 to 1 L of seawater through an 80μm mesh and a 0.2μm Sterivex filter (Millipore) onto a 0.02μm Anotop filter (Whatman). Similarly, DNA was collected for cellular metagenomes (0.22 - 1.2μm, n = 42) and rRNA gene amplicons (1.2- 80μm, n = 46) monthly between May 2009 and August 2014 and May 2009 and June 2013, respectively, by passing 15 to 20 L of seawater through 80μm mesh and a 1.2μm AE filter and onto a 0.22μm Sterivex filter.

Viral-fraction DNA was sent to JGI for library preparation and Illumina HiSeq-2000 2 x 150 bp sequencing under CSP proposal 2799 [44-45]. DNA for cellular metagenomes and cellular 18S rRNA gene amplicons were extracted using the Qiagen AllPrep DNA/RNA kit (Qiagen) and the DNA was cleaned and concentrated using AMPure beads (Beckman Coulter). 18S rRNA amplicons were amplified using universal primer sets 515Y and 926R as described by Yeh et al [46]. Cellular fraction metagenomic libraries were prepared and sequenced as described previously in Dart et. al. [45].

Read quality control and assembly

Viral and cellular metagenomic assemblies for each sample were downloaded from the JGI genome portal (proposal 2799). Reads from both metagenomes were subject to trimming using bbtools [47], read correction using bfc [48], and assembly using SPAdes [49] with a range of k-mers and the following options: ‘-m 2000 --only-assembler -k 33,55,77,99,127 --meta -t 32’.

bioRxiv preprint doi: https://doi.org/10.1101/2024.08.12.607631; this version posted August 15, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

18S rRNA amplicon sequences were trimmed using cutadapt [50] and amplicons were pooled using bbsplit with the PR2 database [51]. The 18S rRNA amplicons were analyzed using DADA2 [52] implemented in QIIME2 [53]. 18S rRNA reads were then denoised and filtered for chimeras using DADA2 commands and assigned taxonomy against the PR2 database [51].

Identification of NCLDV phylotypes

To identify NCLDV phylotypes, protein-coding regions for assemblies were predicted using Prodigal [54], and the protein sequences were searched for copies of the DNA Polymerase Beta (PolB) gene using hmmsearch (e-value 10^{-5}) in HMMER3 [55] using the polB model from Moniruzzaman et al. [26]. PolB sequences were dereplicated at 99% amino acid identity using cd-hit [56,57] requiring 90% coverage. Only sequences greater than 500 amino acids in length (seqkit; [58]) were considered for downstream analyses. PolB sequences were classified using GraftM [59] to remove any non-NCLDV PolB sequences.

Reads were mapped to the corresponding PolB nucleotide sequence using BWA-MEM [60] and sorted and indexed with samtools [61, 62]. Read coverage was calculated using ‘trimmed_mean’ in CoverM (v0.7) (https://github.com/wwood/CoverM) requiring 98% minimum read identity.

Metagenomic binning, quality control, and annotation

Metagenomic reads were aligned to contigs using BWA-MEM and sorted and indexed with samtools [61, 62]. Bins were assembled using MetaBat 2 [63] with ‘runMetaBat.sh -s 100000 -m 5000 –minS 75 –maxEdges 75’. NCLDV bins were identified and scanned for completeness using miComplete [64] using five conserved NCLDV marker genes, namely A32 ATPase, Major Capsid Protein (MCP), SFII helicase, B-family DNA polymerase (PolB), and VLTF3 transcription factor [26]. NCLDV bins with ≥ 3 of the five marker genes were retained for downstream analyses. NCLDV bins were screened with Viralrecall [65] using the NCLDV-specific database (NVOG), requiring a minimum of 3 hits to the NVOG database, and a minimum score of 0. Bins were dereplicated with dRep [66] at 95% nucleotide identity requiring coverage of 30% the shorter bin. Protein coding regions for the NCLDV bins were predicted using Prodigal [54] and annotated using eggNOG-mapper [67,68] using default settings.

VirHostMatcher [69] was used to predict putative hosts of the NCLDVs based on oligonucleotide frequency patterns. Reference transcriptomes from the Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP; [70, 71]) were used as the host database and d2* dissimilarity was calculated using the NCLDV bins. Here we report the lowest d2* dissimilarity for each bin below 0.4. To supplement this approach, a co-occurrence network was created to link NCLDV phylotypes to eukaryotic OTUs using FlashWeave [72]. The resulting correlations were filtered using a weight cutoff of 0.4 [73] and displayed using Cytoscape [74].

bioRxiv preprint doi: https://doi.org/10.1101/2024.08.12.607631; this version posted August 15, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Phylogenetic analyses of NCLDV marker genes

PolB phylotype sequences were aligned using mafft [75, 76] with '–retree 2' and alignments were trimmed using trimal [77] with the '-gappyout' flag. A maximum likelihood tree of PolB phylotypes was constructed using RAxML [78] with 'raxmlHPC-PTHREADS-SSE3 -k -T 80 -f a -m PROTGAMMAGTR -p 12345 -x 12345'.

A multigene tree was constructed using a concatenated alignment of five conserved NCLDV marker genes (A32, MCP, SFII, PolB, and VLTF3) using ncldv_markersearch.py [26]. Poorly aligned sequences and regions were removed using trimal [77] with '-gappyout'. A maximum likelihood tree was constructed using RAxML [78], specifically 'raxmlHPC-PTHREADS-SSE3 -k -T 80 -f a -m PROTGAMMAGTR -p 12345 -x 12345'. This tree was rooted using Chitovirales as an outgroup.

Time-series analysis and clustering

Bray-Curtis similarity was calculated using the *distance* function in the phyloseq R package (v3.19), [79]). Seasonal profiles were calculated as trimmed mean coverage normalized by the number of reads in each metagenomic sample for NCLDV phylotypes across samples. For host 16S rRNA and 18S rRNA sequences, relative abundance was used. Finally, data was linearly interpolated to create an evenly spaced time series with no missing data using the *ApproxFun* in the R stats package. Spectral densities were calculated using *spectrum* in the stats R package.

Using the time series generated above, *ptestg* in the ptest package [80, 81] using the method "extendedRobust" was applied to identify periodicity in these time series with a p-value cutoff of $p < 0.001$. Time series were clustered using z-score normalized monthly mean seasonal profiles with the function *repr_seas_profile* in the TSrepr package [82, 83]. The number of clusters was determined based on the minimum Davies-Bouldin index value [84].

Visualization was performed using a combination of functions from the ggplot2 [85], ggtree [86], and ggpubr [87] R packages as well as base R [88].

Data availability

The viral metagenomes are available at the JGI Genome portal under proposal ID 2799. Cellular metagenomes are available at the National Center for Biotechnology Information (NCBI) under BioProject PRJNA814250 (SRR18278848-SRR18278886) and 18S rRNA amplicons are available on EMBL-EBI under projects PRJEB48162 and PRJEB35673. NCLDV MAGs, PolB phylotypes, protein annotations, host prediction data, and phylogenetic trees are available in a Figshare repository

bioRxiv preprint doi: https://doi.org/10.1101/2024.08.12.607631; this version posted August 15, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(https://figshare.com/projects/Seasonal_Dynamics_of_Giant_viruses_in_Long-Term_metagenomic_time_series_SPOT_/206872).

Results

High NCLDV Prevalence and Diversity in the California Current

We found a diverse set of NCLDV phylotypes on the basis of the reconstructed PolB marker gene. We use the term phylotype throughout the text to indicate reconstructed NCDLV PolB gene sequences clustered at 99% amino acid identity. A total of 676 NCLDV phylotypes were identified in the two size fractions: 655 phylotypes were found in both the viral and cellular size fractions, and 15 and 6 phylotypes were found exclusively in the viral and cellular size fractions, respectively. The NCLDV community was dominated by Algavirales (59%) and Imitervirales (36%), with Pimascovirales having a minor contribution (0.7%) (Figure 1). This is consistent with a recent global survey that found Imitervirales and Algavirales to be the two dominant NCLDV orders in the world's oceans [12]. Imitervirales were more abundant in the larger size fraction (0.2 - 1.2μm) compared to the viral size fraction (0.02 - 0.2μm), consistent with the often-large size of their viral particles [89].

The eukaryotic community, based on 18S rRNA genes in the eukaryotic size fraction (1.2 - 80 μm), was dominated by Dinoflagellata and Ciliophora, with 18S rRNA capturing both heterotrophic and phototrophic eukaryotes. Eukaryotes in the smaller cellular fraction (0.2 - 1.2μm) were detected using chloroplast 16S rRNA. Eukaryotes in the Mamiellaceae order were abundant in the smaller cellular size fraction, including *Bathycoccus*, *Micromonas*, and *Ostreococcus*. Bacillariophyta and Prymnesiophyceae were also abundant in the eukaryotic size fraction (Figure S1).

NCLDV Populations Frequently Exhibit Strong Seasonal Dynamics

We quantified the relative coverage of 676 NCLDV PolB phylotypes across our five-year metagenomic times series in two separate filtered size fractions (0.02-0.2μm "viral" fraction and a 0.2-1.2μm "cellular" fraction). Overall, the NCLDV community exhibited seasonally predictable changes in composition (Figure 1), with peak community similarity (Bray-Curtis) occurring at 1-year intervals, in both sequenced size fractions. Cycles were apparent at the order level with Imitervirales dominating in the late summer and early fall and Algavirales dominating at all other times. These cycles were apparent, though less clearly so, in the corresponding eukaryotic community (Figure 1a, d). Among the NCLDV, the community found in the larger size fraction (0.2-1.2μm, "cellular"; Figure 1c, f) had stronger apparent fluctuations in composition than that found in the smaller fraction (0.02-0.2μm, "viral"; Figure 1b, e). This difference is likely due to the fact that many of the free-floating NCLDVs have sufficiently large

bioRxiv preprint doi: https://doi.org/10.1101/2024.08.12.607631; this version posted August 15, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

viral particles that they are likely to not be found in the smaller "viral" fraction. NCLDV particles at SPOT might be present in higher proportion in the cellular size fractions, contributing to the relatively stronger annual periodicity observed within this fraction. [19, 89-91].

While overall community similarity showed seasonal fluctuations, it was initially unclear if these were driven by fluctuations in the abundance of all or a subset of community members. We individually examined the temporal patterns in abundance of each NCLDV to classify strains as seasonal or non-seasonal (Figure 2), and to distinguish strains with annual peaks from those that peaked biannually. Using a likelihood-ratio test to compare fitted harmonic models we found 13 NCLDVs with strong seasonal dynamics. The same strains were recovered as significantly seasonal using Fisher's g test. We then applied a robust variant of Fisher's g test that recovered a greatly expanded set of significantly periodic strains (131). Many of these strains had unevenly sized annual peaks or "gap" years in which they did not peak, possibly explaining why less robust methods failed to recover them. Even among the strains in which no seasonality was detected using the more robust test, it was clear that some groups of strains showed some apparent seasonal dynamics, in that their peaks always occurred during a specific season even if they did not peak every year (Figure S2). This suggests that longer time series studies may have the power to detect the many potentially seasonal viral populations with a patchy distribution across years. Overall, viruses could be classified as (1) seasonal with one annual peak (17%), (2) seasonal with two annual peaks (corresponding to spring and fall upwellings) (2%), or (3) erratic with one or a couple of peaks seemingly randomly distributed across years (81%).

Closely Related NCLDVs Have Similar Seasonal Dynamics

We clustered phylotypes that exhibited a single annual peak into distinct seasonal clusters based on their mean seasonal profile (Figure 3). NCLDV phylotypes were clustered in each size fraction separately resulting in 10 and 9 clusters in the viral and cellular size fraction, respectively (Figure 3a, b). The clusters showed a succession of NCLDV phylotypes throughout the year, with 11 NCLDV peaking in spring (Mar-May), 7 peaking in summer (Jun-Aug), 21 peaking in fall (Sept-Nov), and 31 peaking in winter (Dec-Feb).

Phylogenetically related phylotypes tended to belong to the same or similar seasonal cluster (Figures 4-5). For pairs of phylotypes, the phylogenetic distance was negatively related to the agreement of the pair's seasonal abundance profiles (Figures 4c,5c). Prasinoviruses (Algavirales) typically belonged to the spring and early summer clusters, while Imitervirales belonged to clusters peaking in later summer and fall. Prasinoviruses were also hyperabundant compared to other NCLDV phylotypes (Figure 4b,5b). Interestingly, Algavirales typically peaked earlier (February) in the viral size fraction and later (April) in the cellular size fraction (Figures 4a-5a), suggesting that larger viruses peak later into the spring in response to their algal hosts blooming during this season [92,93].

bioRxiv preprint doi: https://doi.org/10.1101/2024.08.12.607631; this version posted August 15, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Finally, we note that even though closely related strains tend to have similar seasonal dynamics, occasionally a pair of closely related strains would have drastically different seasonal profiles (Figures 4a-5a). These changes could indicate host shift events and suggest that high-resolution time-series data may be a useful tool to detect viral host shifts along a phylogeny. Alternatively, it is also possible that some protist lineages within a phylogenetically cohesive group showed divergent seasonal dynamics. If the viruses associated with these protists tracked their dynamics, that could at least partly explain our observations.

Eukaryotic Hosts Exhibit a Weaker Seasonality than Their Viruses

We found far fewer eukaryotic taxa had detectable seasonal dynamics than we found for NCLDVs (18) consistent with overall weaker seasonal fluctuations found at the community level (Figure 1). While initially surprising, we suspect that the relatively low population densities of the hosts relative to the prokaryotic communities sequenced simultaneously may mean noise masks potential seasonal patterns. Ultimately, eukaryotic taxa clustered into fewer (3) seasonal clusters, mainly corresponding to the major annual upwelling events: early spring, late spring, and fall (Figure S5).

Distinct Functional Potentials Among Two NCLDV Families

Adding metagenomic bins to the phylotype analyses provided additional information regarding the NCLDV community, including genomic potential and putative hosts. Seventy NCLDV metagenome assembled genomes (MAGs) were assembled with ≥ 3 marker genes and were used for subsequent analyses. The MAGs ranged in size from 72-826 kb and contained 100-857 coding sequences, MAG GC content ranged between 20.2 and 54.3% (Figure S6). A multigene phylogenetic tree indicated that 31 MAGs belong to Algavirales, 38 MAGs to Imitervirales and one was Pimascovirales (Figure 6a). According to eggNOG annotations, Imitervirales MAGs had a particularly large number of genes belonging to COG functional groups, including DNA replication, translation, transcription, and cell wall envelope biogenesis (Fig 6b).

NCLDV MAGs also encoded diverse auxiliary metabolic genes (AMGs) homologous to sets of AMGs previously identified among viral genomes. There were clear order-level differences in the sets of AMGs encoded between Imitervirales and Algavirales MAGs (Figure 6b). Imitervirales MAGs encoded a diverse set of genes involved in energy metabolism and transport, which were all but absent among Algavirales MAGs. This is consistent with previous work showing that Imitervirale members encode diverse transporters both for the transport of dNTPs across the mitochondrial membrane [94] and for the transport of nutrients across the cell membrane [95]. Surprisingly few Algavirales MAGs from our study site encoded nutrient transporters, despite the apparent importance of Algavirales-encoded transporters in some other systems [95, 21]. This may suggest less nutrient limitation of Algavirus replication at SPOT

bioRxiv preprint doi: https://doi.org/10.1101/2024.08.12.607631; this version posted August 15, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

compared to other previously studied locations, which are often more oligotrophic; and we note Algaviruses are less abundant at SPOT in the more oligotrophic late summer and fall seasons. While both Imitervirales and Algavirales MAGs encoded light-harvesting orthologous groups of genes, each NCLDV order encoded distinct sets with specific functions, suggesting that these two orders modulate host photosynthesis using distinct pathways. Algavirales MAGs encoded many orthologous groups directly implicated in photosynthesis, whereas such groups were nearly absent among Imitervirales MAGs. On the other hand, several Imitervirales MAGs, as well as several Algavirales MAGs, encoded various chlorophyll-binding proteins. Finally, superoxide dismutases were found distributed across both Imitervirales and Algavirales MAGs, suggesting that superoxides produced by host metabolism or stress responses may pose a general threat to infecting NCLDVs.

Compositional Patterns Reveal Putative NCLDV Hosts

We identified putative host pairings for our NCLDV MAGs based on k-mer frequency and phylotypes based on a co-occurrence network (Fig 7a, b). Both methodologies allowed for the prediction of a diverse set of hosts for our recovered NCLDVs although we acknowledge that correlation and sequence-based methods have limitations and may sometimes yield false positives, however, they remain valuable tools for identifying valid host-virus pairings. Correlation-based methods have been used previously to successfully recover established giant virus-host linkages in different marine environments [96]. We report many putative connections between Imitervirales MAGs, which were more abundant during the warmer and more oligotrophic late summer and fall seasons, and *Chaetoceros* and *Paraphysomonas* genera (Figure 7a). *Chaetoceros* is a diverse genus of marine diatoms and is globally distributed [97, 98], yet this is the first indication, to our knowledge, of an NCLDV infecting a member of this genus. *Paraphysomonas* is a cosmopolitan phagotrophic clade [99] belonging to the chrysophytes, which were recently suggested to be a target of many novel Imitervirales lineages based on host-virus correlations in data from the Tara Oceans Project [12]. Relatively few hosts for Algavirales MAGs could be predicted with high confidence, but those that could belong to the genus *Micromonas*, a common, sometimes hyperabundant [100] marine photosynthetic picoeukaryote, occurring year-round at SPOT.

Additional to the k-mer frequency-based method, the co-occurrence approach allowed for the prediction of environmentally specific hosts for all three orders of recovered phylotypes (Figure 7b). Many of these connections (e.g., Mamiellophyceae: *Micromonas*) were consistent with results from the k-mer frequency approach, adding confidence to these predictions. A large number of associations were found between Algavirales and Syndiniales. In addition, we also recovered associations between Algavirales members and Dinophyceae, consistent with previous observations [101]. Finally, connections were also observed between Algavirales and Spirotrichea, which are widespread ciliates and a major trophic link from nanoplankton to higher trophic levels [102].

bioRxiv preprint doi: https://doi.org/10.1101/2024.08.12.607631; this version posted August 15, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Discussion

We leveraged a high-resolution multi-year metagenomic time series to assess seasonal patterns of NCLDV abundance in the California Current. This work demonstrates the advantages of long-term sequencing projects that sample multiple size fractions and have frequent sampling [44-46]. We found strong seasonal patterns in NCLDV phylotype abundance that differed systematically across size fractions, consistent with patterns seen for phage infecting cyanobacteria (cyanophage) in these same sample sets [45]. By incorporating multiple years of data in our analysis we were able to detect significant seasonal patterns even in the presence of abnormal years where a peak in abundance did not occur or was shifted.

While previous work has demonstrated the global distribution and numerical abundance of marine NCLDVs [12], these explorations have only captured a snapshot of NCLDV dynamics. By leveraging multi-year data, we show that these viral populations are extremely dynamic in time, often varying to a greater degree than even their putative hosts (Figures 1-5). The majority of phylotypes exhibit no seasonality and are apparently erratic in their patterns, similar to what has been seen for species-like populations of cyanophage at SPOT [45]. Phylotypes that are seasonal peak at different times, presumably following when their particular hosts bloom in abundance, a pattern also paralleled in cyanobacteria-cyanophage dynamics here [45]. Biogeographic investigations must consider this time-varying quality of NCLDV populations moving forward, as it is likely that many local populations vary in time. At the same time, this variation is non-random, as the community returns to a similar compositional state at one-year increments (Figure 1). The strong seasonality in the overall community of NCLDV follows similar patterns documented for phage, prokaryotic, and eukaryotic components of the microbial communities at SPOT [45-46, 103-105], emphasizing the interconnectedness of these communities. As global climate shifts, continued metagenomic surveillance will allow us to assess how seasonal dynamics change in marine microbial communities as environmental conditions are altered [106].

We found that NCLDV phylotypes group into distinct seasonal clusters, and that closely related phylotypes tend to share a cluster (Figures 4-5). This suggests that seasonal clusters reflect overall patterns in host identity since closely related strains are expected to share host taxa (Figure 7). Intriguingly, these patterns mean that temporal patterns of viral abundance may be used to place host-shift events on a viral phylogeny, independent of any knowledge about the host identity of any particular viral strain. Analysis of cyanophage similarly found that while phylogenetic proximity often indicates ecological similarity, closely related taxa sometimes differ in their phenology, indicative of host-switching. Thus, only small changes in genome sequences likely can drive shifts in the host range [107].

Finally, we found that the two major NCLDV orders found at our study site, Imitervirales and Algavirales, had distinct functional potentials based on an analysis of reconstructed viral MAGs (Figure 6). These two orders have distinct seasonal abundance patterns (Figure 1), suggesting

bioRxiv preprint doi: https://doi.org/10.1101/2024.08.12.607631; this version posted August 15, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

that the functional composition of the NCLDV community fluctuates over time and phylogeny and functional potential are key drivers of NCLDV dynamics. These patterns may ultimately be caused by different ecological strategies between these two orders of NCLDV, but further research is needed to validate these patterns. It is not yet clear how different AMGs may play a role in the seasonal variation of viral communities [12, 22, 26], but we believe this will be a fruitful area of research moving forward.

In closing, this time-series analysis of NCLDV not only yields insight into their temporal dynamics but also genomic differences across major orders of these viruses. The distinct phenologies of NCLDVs likely reflect how they ‘track’ with the particular hosts they infect, and we leverage this to identify putative new host connections, including a novel possible connection to the diatom phylum Bacillariophyta. We emphasize that given dynamic changes in viruses over time biogeographic studies need to account for this in their interpretation of patterns. As sequencing costs continue to decline [108], and new tools for automatic sampling are developed [109], we expect the field to increasingly move towards building a high-resolution understanding of viral populations and evolutionary dynamics in the environment.

Acknowledgments

This research was funded by the Joint Genome Institute (JGI), Simons Collaboration on Computational Biogeochemical Modeling of Marine Ecosystems / CBIOMES) grant [549943 to J.A.F.], Simons Award #653212 (JLW), Simons Award, NSF grants [EF-2125142 to J.A.F. NSF-OCE-1737409 to J.A.F.]; National Institutes of Health (NIH) grant 1R01GM120624-01A1 to N.A.A. and J.A.F. NSFC grant [42276163 to S.H.]; Shenzhen Science, Technology and Innovation Commission Programme [JCYJ20220530115401003 to S.H.]. M.M. and B.M. are supported through intramural funds provided by the Rosenstiel School of Marine, Atmospheric, and Earth Sciences, University of Miami.

Conflict of Interest Statement

The authors declare no conflict of interest.

Author contributions

JF, SH and MM supervised the research. JF secured external funding. SH designed and developed the NGS data analysis workflow. SL, JLW, SH and BM performed data analysis and writing of the manuscript. NA coordinated collaboration and data management and edited the manuscript. All authors critically reviewed and edited the manuscript.

bioRxiv preprint doi: https://doi.org/10.1101/2024.08.12.607631; this version posted August 15, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure Legends

Figure 1. NCLDV and host communities at SPOT vary seasonally in their composition. (a,d) The eukaryotic community assayed via 18S rRNA amplicon sequencing in terms of (a) the overall composition and (d) pairwise community similarity calculated between all sample pairs. (b,e) The relative abundance of NCLDV orders in the viral size fraction (0.02-0.2μm) was assayed via reconstructed PolB sequences. (c,f) The same was done for the cellular size fraction (0.2-1.2μm). In panels (d-f) red lines indicate the mean Bray-Curtis similarity for a given temporal separation.

Figure 2: Periodograms to detect seasonality in NCLDV time series. Periodograms represent a decomposition of the time series into a set of periodic functions whose frequencies are represented in the periodogram. High spectral density indicates the dominant frequencies in a time series, expressed in units of 1/years. The class of statistical tests applied in the main text calculates various summary statistics based on these periodograms to assess the significance of seasonal patterns. Here we show representative periodograms along with corresponding normalized coverage for viruses with biannual peaks, annual peaks, and no seasonal peaks.

Figure 3. NCLDV phylotypes fall into distinct seasonal clusters. Distinct clusters for the (a) viral and (b) cellular fractions were calculated for significantly seasonal NCLDV phylotypes using k-medoids clustering and the minimum Davies-Boudin index. Dark lines indicate normalized mean coverage of an individual phylotype in the cluster and the medoid of each cluster is shown in red.

Figure 4. Closely related phylotypes tend to have similar seasonal dynamics (viral size fraction; 0.02-0.2μm). (a) Phylogeny of NCLDV PolB genes and mean annual abundance profiles for each gene. (b) Abundance of each phylotype across samples. Note that Prasinoviruses are particularly abundant in this system. (c) Pairwise correlations in phylotype abundances across time plotted against patristic distances computed from the phylogeny in (a). An expanded figure with phylogenetic tree leaf labels is available in Figure S3.

Figure 5. Closely related phylotypes tend to have similar seasonal dynamics (cellular size fraction; 0.2-1.2μm). (a) Phylogeny of NCLDV PolB genes and mean annual abundance profiles for each gene. Observe that peak abundances for Prasinoviruses are shifted later in the year relative to those shown in the viral fraction in Fig 4. (b) The abundance of each phylotype across samples. (c) Pairwise correlations in phylotype abundances across time plotted against patristic distances computed from the phylogeny in (a). An expanded figure with phylogenetic tree leaf labels is available in Figure S4.

Figure 6: Phylogenetic and functional potential of NCLDV MAGs. (a) Multigene Maximum likelihood phylogenetic tree from a concatenated alignment of five conserved NCLDV genes from reference genomes (black) and MAGs (red). The node color indicates the reported host of reference NCLDVs, and the branch color indicates NCLDV order membership. (b) The tree from (a) subset to include only MAGs generated here. The heatmap (center) indicates the number of genes within each COG functional group and the bubble plot (right) indicates the number of genes involved in select metabolic functions.

bioRxiv preprint doi: https://doi.org/10.1101/2024.08.12.607631; this version posted August 15, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 7: Reconstructed NCLDV MAGs are predicted to infect diverse host taxa. (a) Phylogenetic tree of NCLDV MAGs and their predicted host taxa based on k-mer profile similarity (d2∗) with MMETSP reference genomes. High-confidence (d2∗ < 0.2) host predictions are shown in red, while lower-confidence host predictions are shown in blue. Peak season is also shown for MAGs with significant seasonality. (b) Co-occurrence network of NCLDV PolB phylotypes and eukaryote OTUs using 18S data. A weight cutoff of 0.4 was used to filter connections, and the number of connections is shown by the edge thickness. Known NCLDV and eukaryotic host pairs from cultured isolates are shown in red and connections match and d2* results are shown with a star (*). An expanded network and phylogenetic tree are available in Figure S7.

References

1. Suttle, C.A., 2005. Viruses in the sea. *Nature*, *437*(7057), pp.356-361.
2. Gregory, A.C., Zayed, A.A., Conceição-Neto, N., Temperton, B., Bolduc, B., Alberti, A., Ardyna, M., Arkhipova, K., Carmichael, M., Cruaud, C. and Dimier, C., 2019. Marine DNA viral macro-and microdiversity from pole to pole. *Cell*, *177*(5), pp.1109-1123.
3. Fuhrman, J.A., 1999. Marine viruses and their biogeochemical and ecological effects. *Nature*, *399*(6736), pp.541-548.
4. Weitz, J.S. and Wilhelm, S.W., 2012. Ocean viruses and their effects on microbial communities and biogeochemical cycles. *F1000 biology reports*, *4*.
5. Wilhelm, S.W. and Suttle, C.A., 1999. Viruses and nutrient cycles in the sea: viruses play critical roles in the structure and function of aquatic food webs. *Bioscience*, *49*(10), pp.781-788.
6. Jardillier, L., Zubkov, M.V., Pearman, J. and Scanlan, D.J., 2010. Significant CO2 fixation by small prymnesiophytes in the subtropical and tropical northeast Atlantic Ocean. *The ISME journal*, *4*(9), pp.1180-1192.
7. Li, W.K., 1994. Primary production of prochlorophytes, cyanobacteria, and eucaryotic ultraphytoplankton: measurements from flow cytometric sorting. *Limnology and Oceanography*, *39*(1), pp.169-175.
8. Worden, A.Z., Nolan, J.K. and Palenik, B., 2004. Assessing the dynamics and ecology of marine picophytoplankton: the importance of the eukaryotic component. *Limnology and oceanography*, *49*(1), pp.168-179.
9. Colson, P., De Lamballerie, X., Yutin, N., Asgari, S., Bigot, Y., Bideshi, D.K., Cheng, X.W., Federici, B.A., Van Etten, J.L., Koonin, E.V. and La Scola, B., 2013. “Megavirales”, a proposed new order for eukaryotic nucleocytoplasmic large DNA viruses. *Archives of virology*, *158*, pp.2517-2521.
10. Fischer, M.G., 2016. Giant viruses come of age. *Current Opinion in Microbiology*, *31*, pp.50-57.

bioRxiv preprint doi: https://doi.org/10.1101/2024.08.12.607631; this version posted August 15, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

11. Koonin, E.V. and Yutin, N., 2019. Evolution of the large nucleocytoplasmic DNA viruses of eukaryotes and convergent origins of viral gigantism. *Advances in Virus Research*, *103*, pp.167-202.
12. Endo, H., Blanc-Mathieu, R., Li, Y., Salazar, G., Henry, N., Labadie, K., de Vargas, C., Sullivan, M.B., Bowler, C., Wincker, P. and Karp-Boss, L., 2020. Biogeography of marine giant viruses reveals their interplay with eukaryotes and ecological functions. *Nature ecology & evolution*, *4*(12), pp.1639-1649.
13. Bratbak, G., Egge, J.K. and Heldal, M., 1993. Viral mortality of the marine alga Emiliania huxleyi (Haptophyceae) and termination of algal blooms. *Marine Ecology Progress Series*, pp.39-48.
14. Tarutani, K., Nagasaki, K., Itakura, S. and Yamaguchi, M., 2001. Isolation of a virus infecting the novel shellfish-killing dinoflagellate Heterocapsa circularisquama. *Aquatic Microbial Ecology*, *23*(2), pp.103-111.
15. Fischer, M.G., Allen, M.J., Wilson, W.H. and Suttle, C.A., 2010. Giant virus with a remarkable complement of genes infects marine zooplankton. *Proceedings of the National Academy of Sciences*, *107*(45), pp.19508-19513.
16. Needham, D.M., Yoshizawa, S., Hosaka, T., Poirier, C., Choi, C.J., Hehenberger, E., Irwin, N.A., Wilken, S., Yung, C.M., Bachy, C. and Kurihara, R., 2019. A distinct lineage of giant viruses brings a rhodopsin photosystem to unicellular marine predators. *Proceedings of the National Academy of Sciences*, *116*(41), pp.20574-20583.
17. Mihara, T., Koyano, H., Hingamp, P., Grimsley, N., Goto, S. and Ogata, H., 2018. Taxon richness of “Megaviridae” exceeds those of bacteria and archaea in the ocean. *Microbes and Environments*, *33*(2), pp.162-171.
18. Li, Y., Hingamp, P., Watai, H., Endo, H., Yoshida, T. and Ogata, H., 2018. Degenerate PCR primers to reveal the diversity of giant viruses in coastal waters. *Viruses*, *10*(9), p.496.
19. Philippe, N., Legendre, M., Doutre, G., Couté, Y., Poirot, O., Lescot, M., Arslan, D., Seltzer, V., Bertaux, L., Bruley, C. and Garin, J., 2013. Pandoraviruses: amoeba viruses with genomes up to 2.5 Mb reaching that of parasitic eukaryotes. *Science*, *341*(6143), pp.281-286.
20. Schvarcz, C.R. and Steward, G.F., 2018. A giant virus infecting green algae encodes key fermentation genes. *Virology*, *518*, pp.423-433.
21. Monier, A., Chambouvet, A., Milner, D.S., Attah, V., Terrado, R., Lovejoy, C., Moreau, H., Santoro, A.E., Derelle, É. and Richards, T.A., 2017. Host-derived viral transporter protein for nitrogen uptake in infected marine phytoplankton. *Proceedings of the National Academy of Sciences*, *114*(36), pp.E7489-E7498.
22. Schulz, F., Roux, S., Paez-Espino, D., Jungbluth, S., Walsh, D.A., Denef, V.J., McMahon, K.D., Konstantinidis, K.T., Eloe-Fadrosh, E.A., Kyrpides, N.C. and Woyke, T., 2020. Giant virus diversity and host interactions through global metagenomics. *Nature*, *578*(7795), pp.432-436.

bioRxiv preprint doi: https://doi.org/10.1101/2024.08.12.607631; this version posted August 15, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

23. Yutin, N. and Koonin, E.V., 2012. Proteorhodopsin genes in giant viruses. *Biology direct*, *7*, pp.1-6.
24. Philosof, A. and Béjà, O., 2013. Bacterial, archaeal and viral□like rhodopsins from the Red Sea. *Environmental Microbiology Reports*, *5*(3), pp.475-482.
25. Needham, David M., Camille Poirier, Elisabeth Hehenberger, Valeria Jiménez, Jarred E. Swalwell, Alyson E. Santoro, and Alexandra Z. Worden. "Targeted metagenomic recovery of four divergent viruses reveals shared and distinctive characteristics of giant viruses of marine eukaryotes." *Philosophical Transactions of the Royal Society B* 374, no. 1786 (2019): 20190086.
26. Moniruzzaman, M., Martinez-Gutierrez, C.A., Weinheimer, A.R. and Aylward, F.O., 2020. Dynamic genome evolution and complex virocell metabolism of globally-distributed giant viruses. *Nature communications*, *11*(1), p.1710.
27. Ha, A.D., Moniruzzaman, M. and Aylward, F.O., 2021. High transcriptional activity and diverse functional repertoires of hundreds of giant viruses in a coastal marine system. *MSystems*, *6*(4), pp.10-1128.
28. Monier, A., Claverie, J.M. and Ogata, H., 2008. Taxonomic distribution of large DNA viruses in the sea. *Genome biology*, *9*, pp.1-15.
29. Hingamp, P., Grimsley, N., Acinas, S.G., Clerissi, C., Subirana, L., Poulain, J., Ferrera, I., Sarmento, H., Villar, E., Lima-Mendez, G. and Faust, K., 2013. Exploring nucleo-cytoplasmic large DNA viruses in Tara Oceans microbial metagenomes. *The ISME journal*, *7*(9), pp.1678-1695.
30. Clerissi, C., Desdevises, Y., Romac, S., Audic, S., de Vargas, C., Acinas, S.G., Casotti, R., Poulain, J., Wincker, P., Hingamp, P. and Ogata, H., 2015. Deep sequencing of amplified Prasinovirus and host green algal genes from an Indian Ocean transect reveals interacting trophic dependencies and new genotypes. *Environmental microbiology reports*, *7*(6), pp.979-989.
31. Li, Y., Endo, H., Gotoh, Y., Watai, H., Ogawa, N., Blanc-Mathieu, R., Yoshida, T. and Ogata, H., 2019. The earth is small for “Leviathans”: long-distance dispersal of giant viruses across aquatic environments. *Microbes and environments*, *34*(3), pp.334-339.
32. Vincent, F., Gralka, M., Schleyer, G., Schatz, D., Cabrera-Brufau, M., Kuhlisch, C., Sichert, A., Vidal-Melgosa, S., Mayers, K., Barak-Gavish, N. and Flores, J.M., 2023. Viral infection switches the balance between bacterial and eukaryotic recyclers of organic matter during coccolithophore blooms. *Nature communications*, *14*(1), p.510.
33. Kuhlisch, C., Shemi, A., Barak-Gavish, N., Schatz, D. and Vardi, A., 2024. Algal blooms in the ocean: hot spots for chemically mediated microbial interactions. *Nature Reviews Microbiology*, *22*(3), pp.138-154.
34. Kuhlisch, C., Schleyer, G., Shahaf, N., Vincent, F., Schatz, D. and Vardi, A., 2021. Viral infection of algal blooms leaves a unique metabolic footprint on the dissolved organic matter in the ocean. *Science advances*, *7*(25), p.eabf4680.

bioRxiv preprint doi: https://doi.org/10.1101/2024.08.12.607631; this version posted August 15, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

35. Thamatrakoln, K., Talmy, D., Haramaty, L., Maniscalco, C., Latham, J.R., Knowles, B., Natale, F., Coolen, M.J., Follows, M.J. and Bidle, K.D., 2019. Light regulation of coccolithophore host–virus interactions. *New Phytologist*, *221*(3), pp.1289-1302.
36. Vardi, A., Haramaty, L., Van Mooy, B.A., Fredricks, H.F., Kimmance, S.A., Larsen, A. and Bidle, K.D., 2012. Host–virus dynamics and subcellular controls of cell fate in a natural coccolithophore population. *Proceedings of the National Academy of Sciences*, *109*(47), pp.19327-19332.
37. Sharoni, S., Trainic, M., Schatz, D., Lehahn, Y., Flores, M.J., Bidle, K.D., Ben-Dor, S., Rudich, Y., Koren, I. and Vardi, A., 2015. Infection of phytoplankton by aerosolized marine viruses. *Proceedings of the National Academy of Sciences*, *112*(21), pp.6643-6647.
38. Laber, C.P., Hunter, J.E., Carvalho, F., Collins, J.R., Hunter, E.J., Schieler, B.M., Boss, E., More, K., Frada, M., Thamatrakoln, K. and Brown, C.M., 2018. Coccolithovirus facilitation of carbon export in the North Atlantic. *Nature microbiology*, *3*(5), pp.537-547.
39. Gran-Stadniczeñko, S., Krabberød, A.K., Sandaa, R.A., Yau, S., Egge, E. and Edvardsen, B., 2019. Seasonal dynamics of algae-infecting viruses and their inferred interactions with protists. *Viruses*, *11*(11), p.1043.
40. Needham, D.M., Fichot, E.B., Wang, E., Berdjeb, L., Cram, J.A., Fichot, C.G. and Fuhrman, J.A., 2018. Dynamics and interactions of highly resolved marine plankton via automated high-frequency sampling. *The ISME journal*, *12*(10), pp.2417-2432.
41. Needham, D.M., Sachdeva, R. and Fuhrman, J.A., 2017. Ecological dynamics and co-occurrence among marine phytoplankton, bacteria and myoviruses shows microdiversity matters. *The ISME journal*, *11*(7), pp.1614-1629.
42. Marquardt, M., Vader, A., Stübner, E.I., Reigstad, M. and Gabrielsen, T.M., 2016. Strong seasonality of marine microbial eukaryotes in a high-Arctic fjord (Isfjorden, in West Spitsbergen, Norway). *Applied and environmental microbiology*, *82*(6), pp.1868-1880.
43. Lambert, S., Tragin, M., Lozano, J.C., Ghiglione, J.F., Vaulot, D., Bouget, F.Y. and Galand, P.E., 2019. Rhythmicity of coastal marine picoeukaryotes, bacteria and archaea despite irregular environmental perturbations. *The ISME Journal*, *13*(2), pp.388-401.
44. Ignacio-Espinoza, J.C., Ahlgren, N.A. and Fuhrman, J.A., 2020. Long-term stability and Red Queen-like strain dynamics in marine viruses. *Nature Microbiology*, *5*(2), pp.265-271.
45. Dart, E., Fuhrman, J.A. and Ahlgren, N.A., 2023. Diverse marine T4-like cyanophage communities are primarily comprised of low-abundance species including species with distinct seasonal, persistent, occasional, or sporadic dynamics. *Viruses*, *15*(2), p.581.
46. Yeh, Y.C., McNichol, J., Needham, D.M., Fichot, E.B., Berdjeb, L. and Fuhrman, J.A., 2021. Comprehensive single□PCR 16S and 18S rRNA community analysis validated with mock communities, and estimation of sequencing bias against 18S. *Environmental Microbiology*, *23*(6), pp.3240-3250.
47. Bushnell, B., 2018. BBTools: a suite of fast, multithreaded bioinformatics tools designed for analysis of DNA and RNA sequence data. *Joint Genome Institute*.

bioRxiv preprint doi: https://doi.org/10.1101/2024.08.12.607631; this version posted August 15, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

48. Li, H., 2015. BFC: correcting Illumina sequencing errors. *Bioinformatics*, *31*(17), pp.2885-2887.
49. Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D. and Pyshkin, A.V., 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of computational biology*, *19*(5), pp.455-477.
50. Martin, M., 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal*, *17*(1), pp.10-12.
51. Guillou, L., Bachar, D., Audic, S., Bass, D., Berney, C., Bittner, L., Boutte, C., Burgaud, G., de Vargas, C., Decelle, J. and Del Campo, J., 2012. The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote small sub-unit rRNA sequences with curated taxonomy. *Nucleic acids research*, *41*(D1), pp.D597-D604.
52. Callahan, B.J., McMurdie, P.J., Rosen, M.J., Han, A.W., Johnson, A.J.A. and Holmes, S.P., 2016. DADA2: High-resolution sample inference from Illumina amplicon data. *Nature methods*, *13*(7), pp.581-583.
53. Hall, M. and Beiko, R.G., 2018. 16S rRNA gene analysis with QIIME2. *Microbiome analysis: methods and protocols*, pp.113-129.
54. Hyatt, D., Chen, G.L., LoCascio, P.F., Land, M.L., Larimer, F.W. and Hauser, L.J., 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC bioinformatics*, *11*, pp.1-11.
55. Eddy, S.R., 2011. Accelerated profile HMM searches. *PLoS computational biology*, *7*(10), p.e1002195.
56. Fu, L., Niu, B., Zhu, Z., Wu, S. and Li, W., 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, *28*(23), pp.3150-3152.
57. Li, W. and Godzik, A., 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, *22*(13), pp.1658-1659.
58. Shen, W., Le, S., Li, Y. and Hu, F., 2016. SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PloS one*, *11*(10), p.e0163962.
59. Boyd, J.A., Woodcroft, B.J. and Tyson, G.W., 2018. GraftM: a tool for scalable, phylogenetically informed classification of genes within metagenomes. *Nucleic Acids Research*, *46*(10), pp.e59-e59.
60. Li, H., 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997*.
61. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. and 1000 Genome Project Data Processing Subgroup, 2009. The sequence alignment/map format and SAMtools. *bioinformatics*, *25*(16), pp.2078-2079.
62. Li, H., 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, *27*(21), pp.2987-2993.

bioRxiv preprint doi: https://doi.org/10.1101/2024.08.12.607631; this version posted August 15, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

63. Kang, D.D., Li, F., Kirton, E., Thomas, A., Egan, R., An, H. and Wang, Z., 2019. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ*, *7*, p.e7359.
64. Hugoson, E., Lam, W.T. and Guy, L., 2020. miComplete: weighted quality evaluation of assembled microbial genomes. *Bioinformatics*, *36*(3), pp.936-937.
65. Aylward, F.O. and Moniruzzaman, M., 2021. ViralRecall—a flexible command-line tool for the detection of giant virus signatures in 'Omic Data. *Viruses*, *13*(2), p.150.
66. Olm, M.R., Brown, C.T., Brooks, B. and Banfield, J.F., 2017. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *The ISME journal*, *11*(12), pp.2864-2868.
67. Huerta-Cepas, J., Forslund, K., Coelho, L.P., Szklarczyk, D., Jensen, L.J., Von Mering, C. and Bork, P., 2017. Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. *Molecular biology and evolution*, *34*(8), pp.2115-2122.
68. Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S.K., Cook, H., Mende, D.R., Letunic, I., Rattei, T., Jensen, L.J. and von Mering, C., 2019. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic acids research*, *47*(D1), pp.D309-D314.
69. Ahlgren, N.A., Ren, J., Lu, Y.Y., Fuhrman, J.A. and Sun, F., 2017. Alignment-free oligonucleotide frequency dissimilarity measure improves prediction of hosts from metagenomically-derived viral sequences. *Nucleic acids research*, *45*(1), pp.39-53.
70. Keeling, P.J., Burki, F., Wilcox, H.M., Allam, B., Allen, E.E., Amaral-Zettler, L.A., Armbrust, E.V., Archibald, J.M., Bharti, A.K., Bell, C.J. and Beszteri, B., 2014. The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. *PLoS biology*, *12*(6), p.e1001889.
71. Johnson, L.K., Alexander, H. and Brown, C.T., 2019. Re-assembly, quality evaluation, and annotation of 678 microbial eukaryotic reference transcriptomes. *Gigascience*, *8*(4), p.giy158.
72. Tackmann, J., Rodrigues, J.F.M. and von Mering, C., 2019. Rapid inference of direct interactions in large-scale ecological networks from heterogeneous microbial sequencing data. *Cell systems*, *9*(3), pp.286-296.
73. Meng, L., Endo, H., Blanc-Mathieu, R., Chaffron, S., Hernández-Velázquez, R., Kaneko, H. and Ogata, H., 2021. Quantitative assessment of nucleocytoplasmic large DNA virus and host interactions predicted by co-occurrence analyses. *MSphere*, *6*(2), pp.10-1128.
74. Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B. and Ideker, T., 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, *13*(11), pp.2498-2504.
75. Katoh, K. and Standley, D.M., 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution*, *30*(4), pp.772-780.

bioRxiv preprint doi: https://doi.org/10.1101/2024.08.12.607631; this version posted August 15, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

76. Katoh, K., Misawa, K., Kuma, K.I. and Miyata, T., 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic acids research*, *30*(14), pp.3059-3066.
77. Capella-Gutiérrez, S., Silla-Martínez, J.M. and Gabaldón, T., 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, *25*(15), pp.1972-1973.
78. Stamatakis, A., 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*(9), pp.1312-1313.
79. McMurdie, P.J. and Holmes, S., 2013. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PloS one*, *8*(4), p.e61217.
80. Ahdesmäki, M., Lähdesmäki, H., Pearson, R., Huttunen, H. and Yli-Harja, O., 2005. Robust detection of periodic time series measured from biological systems. *BMC bioinformatics*, *6*, pp.1-18.
81. MacKinnon, J.G., 2000, March. Computing numerical distribution functions in econometrics. In *High performance computing systems and applications* (pp. 455-471). Boston, MA: Springer US.
82. Laurinec, P. and Lucká, M., 2018. Clustering-based forecasting method for individual consumers electricity load using time series representations. *Open Computer Science*, *8*(1), pp.38-50.
83. Laurinec, P., 2018. TSrepr R package: time series representations. *Journal of Open Source Software*, *3*(23), p.577.
84. Davies, D. and Bouldin, D., 1979. A cluster separation measure, IEEE transactions on patter analysis and machine intelligence. Vol.
85. Wickham, H., 2006. An introduction to ggplot: An implementation of the grammar of graphics in R. *Statistics*, pp.1-8.
86. Yu, G., Smith, D.K., Zhu, H., Guan, Y. and Lam, T.T.Y., 2017. ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution*, *8*(1), pp.28-36.
87. Kassambara, A., 2018. ggpubr:'ggplot2'based publication-ready plots. *R package version*, p.2.
88. Ihaka, R. and Gentleman, R., 1996. R: a language for data analysis and graphics. *Journal of computational and graphical statistics*, *5*(3), pp.299-314.
89. Scola, B.L., Audic, S., Robert, C., Jungang, L., de Lamballerie, X., Drancourt, M., Birtles, R., Claverie, J.M. and Raoult, D., 2003. A giant virus in amoebae. *Science*, *299*(5615), pp.2033-2033.
90. Claverie, J.M., Ogata, H., Audic, S., Abergel, C., Suhre, K. and Fournier, P.E., 2006. Mimivirus and the emerging concept of "giant" virus. *Virus research*, *117*(1), pp.133-144.
91. Abergel, C., Legendre, M. and Claverie, J.M., 2015. The rapidly expanding universe of giant viruses: Mimivirus, Pandoravirus, Pithovirus and Mollivirus. *FEMS microbiology reviews*, *39*(6), pp.779-796.

bioRxiv preprint doi: https://doi.org/10.1101/2024.08.12.607631; this version posted August 15, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

92. Yeh, Y.C. and Fuhrman, J.A., 2022. Contrasting diversity patterns of prokaryotes and protists over time and depth at the San-Pedro Ocean Time series. *ISME communications*, *2*(1), p.36.
93. Needham, D.M. and Fuhrman, J.A., 2016. Pronounced daily succession of phytoplankton, archaea and bacteria following a spring bloom. *Nature microbiology*, *1*(4), pp.1-7.
94. Monné, M., Robinson, A.J., Boes, C., Harbour, M.E., Fearnley, I.M. and Kunji, E.R., 2007. The mimivirus genome encodes a mitochondrial carrier that transports dATP and dTTP. *Journal of virology*, *81*(7), pp.3181-3186.
95. Monier, A., Welsh, R.M., Gentemann, C., Weinstock, G., Sodergren, E., Armbrust, E.V., Eisen, J.A. and Worden, A.Z., 2012. Phosphate transporters in marine phytoplankton and their viruses: cross□domain commonalities in viral□host gene exchanges. *Environmental Microbiology*, *14*(1), pp.162-176.
96. Moniruzzaman, M., Wurch, L.L., Alexander, H., Dyhrman, S.T., Gobler, C.J. and Wilhelm, S.W., 2017. Virus-host relationships of marine single-celled eukaryotes resolved from metatranscriptomics. *Nature Communications*, *8*(1), p.16054.
97. De Luca, D., Kooistra, W.H., Sarno, D., Gaonkar, C.C. and Piredda, R., 2019. Global distribution and diversity of Chaetoceros (Bacillariophyta, Mediophyceae): integration of classical and novel strategies. *PeerJ*, *7*, p.e7410.
98. Rines, J.E. and Theriot, E.C., 2003. Systematics of Chaetocerotaceae (Bacillariophyceae). I. A phylogenetic analysis of the family. *Phycological Research*, *51*(2), pp.83-98.
99. Scoble, J.M. and Cavalier-Smith, T., 2014. Scale evolution in Paraphysomonadida (Chrysophyceae): sequence phylogeny and revised taxonomy of Paraphysomonas, new genus Clathromonas, and 25 new species. *European journal of protistology*, *50*(5), pp.551-592.
100. Not, F., Latasa, M., Marie, D., Cariou, T., Vaulot, D. and Simon, N., 2004. A single species, Micromonas pusilla (Prasinophyceae), dominates the eukaryotic picoplankton in the Western English Channel. *Applied and Environmental Microbiology*, *70*(7), pp.4064-4072.
101. Ogata, H., Toyoda, K., Tomaru, Y., Nakayama, N., Shirai, Y., Claverie, J.M. and Nagasaki, K., 2009. Remarkable sequence similarity between the dinoflagellate-infecting marine girus and the terrestrial pathogen African swine fever virus. *Virology Journal*, *6*, pp.1-8.
102. Klug, M.J. and Reddy, C.A. eds., 1984. *Current Perspectives in Microbial Ecology: Proceedings of the Third International Symposium on Microbial Ecology, Michigan State University, 7-12 August 1983* (Vol. 51). American Society for Microbiology.
103. Fuhrman, J.A., Hewson, I., Schwalbach, M.S., Steele, J.A., Brown, M.V. and Naeem, S., 2006. Annually reoccurring bacterial communities are predictable from ocean conditions. *Proceedings of the National Academy of Sciences*, *103*(35), pp.13104-13109.

bioRxiv preprint doi: https://doi.org/10.1101/2024.08.12.607631; this version posted August 15, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

104. Ahlgren, N.A., Perelman, J.N., Yeh, Y.C. and Fuhrman, J.A., 2019. Multi□year dynamics of fine□scale marine cyanobacterial populations are more strongly explained by phage interactions than abiotic, bottom□up factors. *Environmental Microbiology*, *21*(8), pp.2948-2963.

105. Chow, C.E.T. and Fuhrman, J.A., 2012. Seasonality and monthly dynamics of marine myovirus communities. *Environmental microbiology*, *14*(8), pp.2171-2183.

106. Dutkiewicz, S., Scott, J.R. and Follows, M.J., 2013. Winners and losers: Ecological and biogeochemical changes in a warming ocean. *Global Biogeochemical Cycles*, *27*(2), pp.463-477.

107. Marston, M.F., Pierciey Jr, F.J., Shepard, A., Gearin, G., Qi, J., Yandava, C., Schuster, S.C., Henn, M.R. and Martiny, J.B., 2012. Rapid diversification of coevolving marine Synechococcus and a virus. *Proceedings of the National Academy of Sciences*, *109*(12), pp.4544-4549.

108. Muir, P., Li, S., Lou, S., Wang, D., Spakowicz, D.J., Salichos, L., Zhang, J., Weinstock, G.M., Isaacs, F., Rozowsky, J. and Gerstein, M., 2016. The real cost of sequencing: scaling computation to keep pace with data generation. *Genome biology*, *17*, pp.1-9.

109. Evans, S., Birch, J., Breier, J., Jakuba, M., Saito, M. and Robidart, J., 2019. Ocean robots uncover microbial secrets. *Microbiology Today*, *46*(1), pp.22-25.

bioRxiv preprint doi: https://doi.org/10.1101/2024.08.12.607631; this version posted August 15, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

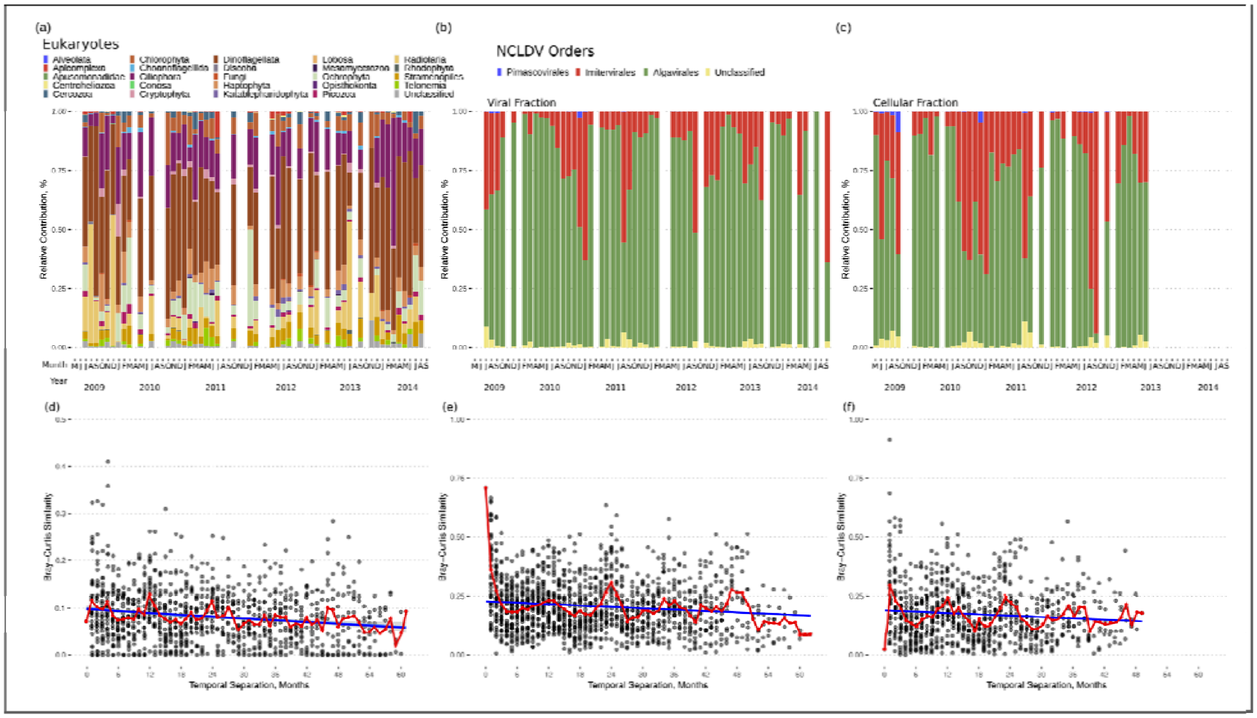


Figure 1. NCLDV and host communities at SPOT vary seasonally in their composition. (a,d) The eukaryotic community assayed via 18S rRNA amplicon sequencing in terms of (a) the overall composition and (d) pairwise community similarity calculated between all sample pairs. (b,e) The relative abundance of NCLDV orders in the viral size fraction (0.02-0.2μm) was assayed via reconstructed PolB sequences. (c,f) The same was done for the cellular size fraction (0.2-1.2μm). In panels (d-f) red lines indicate the mean Bray-Curtis similarity for a given temporal separation.

bioRxiv preprint doi: https://doi.org/10.1101/2024.08.12.607631; this version posted August 15, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

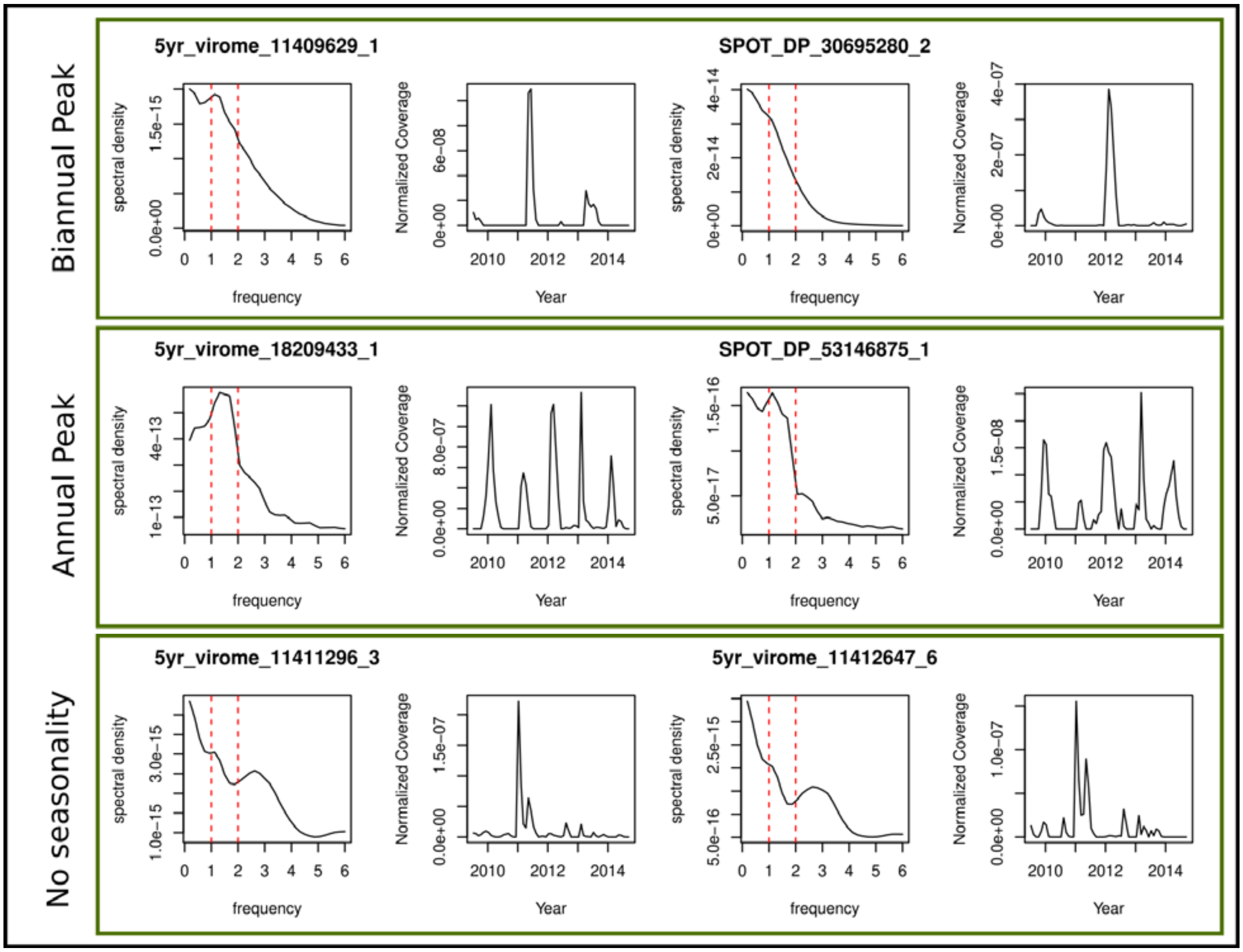


Figure 2: Periodograms to detect seasonality in NCLDV time series. Periodograms represent a decomposition of the time series into a set of periodic functions whose frequencies are represented in the periodogram. High spectral density indicates the dominant frequencies in a time series, expressed in units of 1/years. The class of statistical tests applied in the main text calculates various summary statistics based on these periodograms to assess the significance of seasonal patterns. Here we show representative periodograms along with corresponding normalized coverage for viruses with biannual peaks, annual peaks, and no seasonal peaks.

bioRxiv preprint doi: https://doi.org/10.1101/2024.08.12.607631; this version posted August 15, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

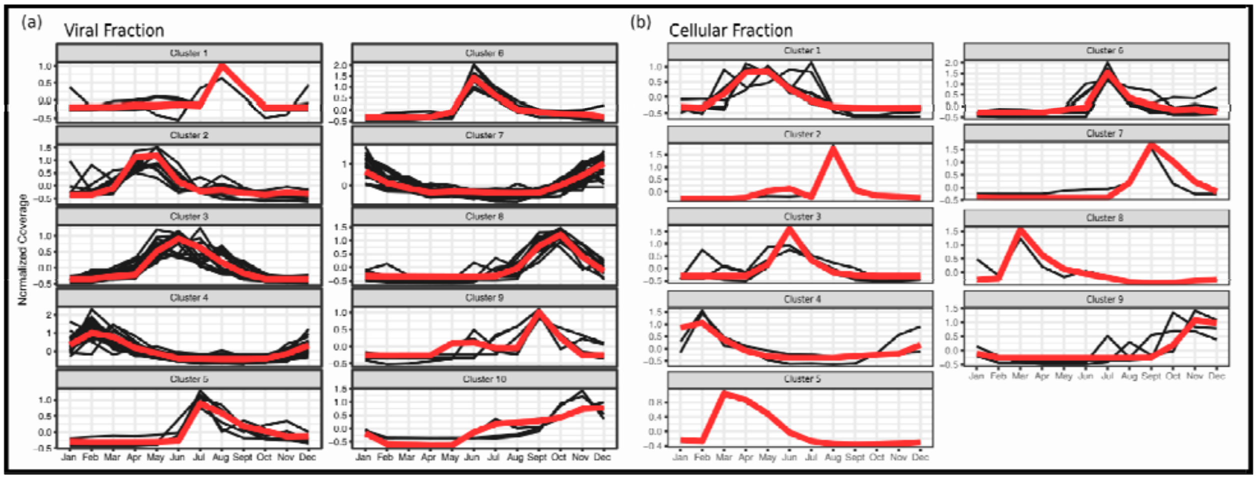


Figure 3. NCLDV phylotypes fall into distinct seasonal clusters. Distinct clusters for the (a) viral and (b) cellular fractions were calculated for significantly seasonal NCLDV phylotypes using k-medoids clustering and the minimum Davies-Boudin index. Dark lines indicate normalized mean coverage of an individual phylotype in the cluster and the medoid of each cluster is shown in red.

bioRxiv preprint doi: https://doi.org/10.1101/2024.08.12.607631; this version posted August 15, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

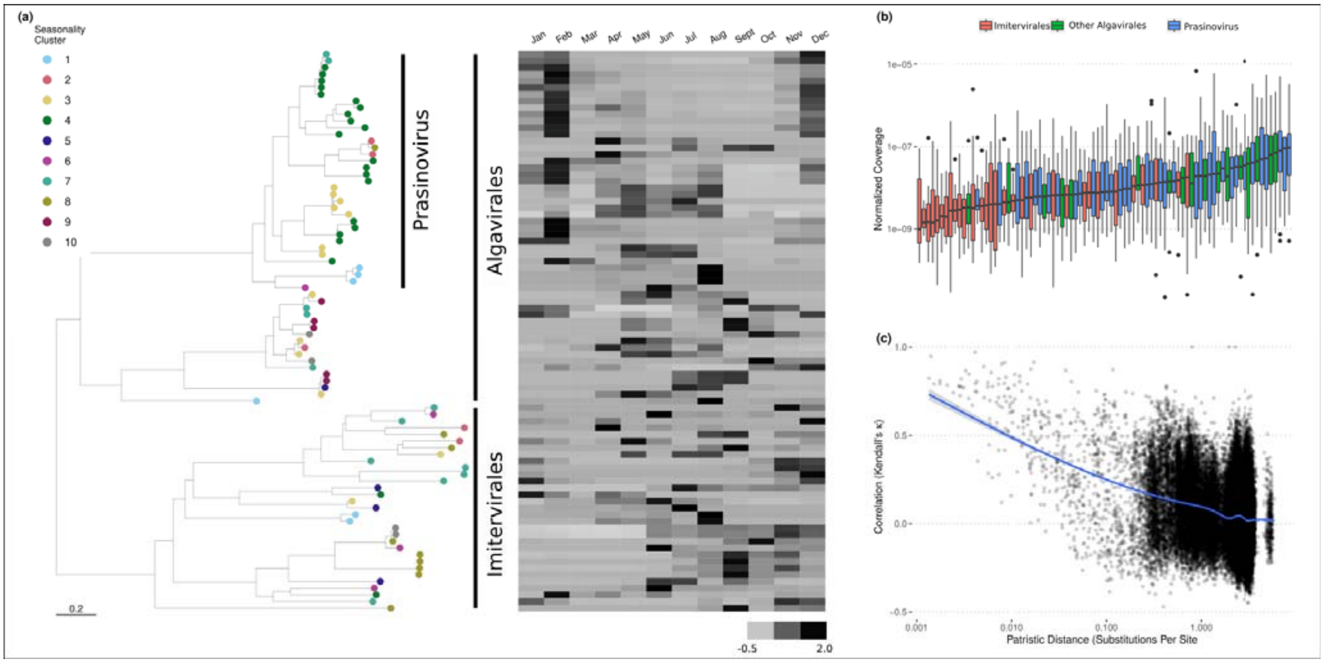


Figure 4. Closely related phylotypes tend to have similar seasonal dynamics (viral size fraction; 0.02-0.2µm). (a) Phylogeny of NCLDV PolB genes and mean annual abundance profiles for each gene. (b) Abundance of each phylotype across samples. Note that Prasinoviruses are particularly abundant in this system. (c) Pairwise correlations in phylotype abundances across time plotted against patristic distances computed from the phylogeny in (a). An expanded figure with phylogenetic tree leaf labels is available in Figure S3.

bioRxiv preprint doi: https://doi.org/10.1101/2024.08.12.607631; this version posted August 15, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

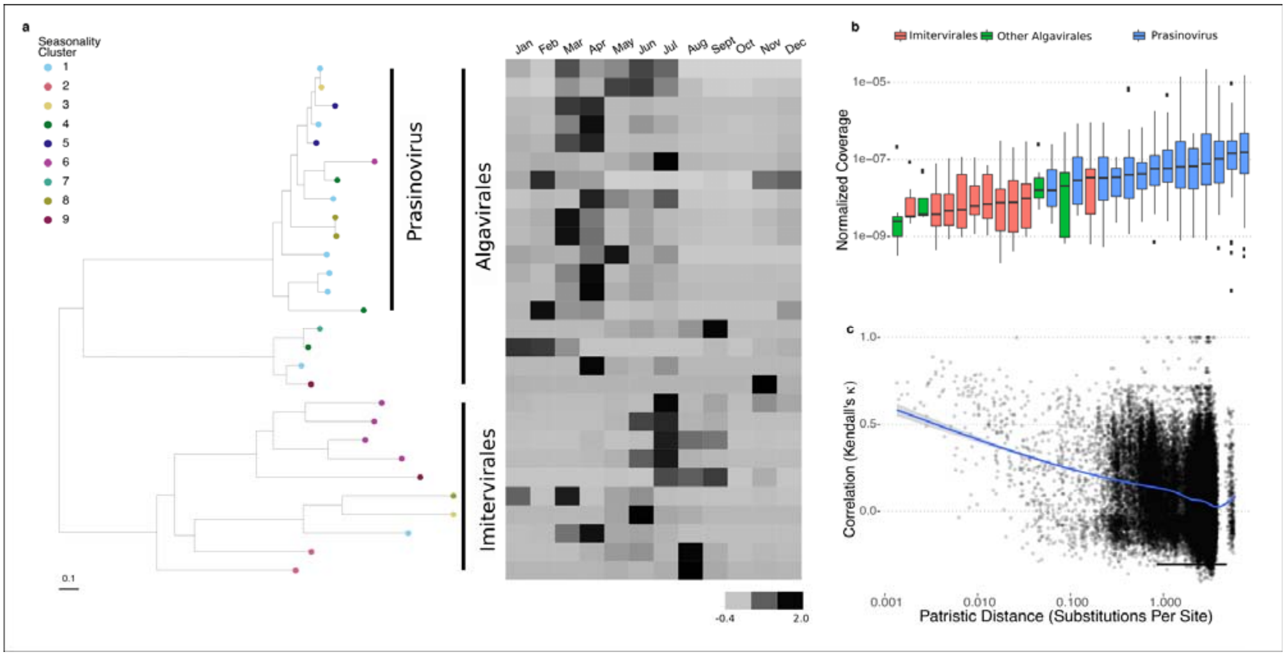


Figure 5. Closely related phylotypes tend to have similar seasonal dynamics (cellular size fraction; 0.2-1.2µm). (a) Phylogeny of NCLDV PolB genes and mean annual abundance profiles for each gene. Observe that peak abundances for Prasinoviruses are shifted later in the year relative to those shown in the viral fraction in Fig 4. (b) The abundance of each phylotype across samples. (c) Pairwise correlations in phylotype abundances across time plotted against patristic distances computed from the phylogeny in (a). An expanded figure with phylogenetic tree leaf labels is avaliable in Figure S4.

bioRxiv preprint doi: https://doi.org/10.1101/2024.08.12.607631; this version posted August 15, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

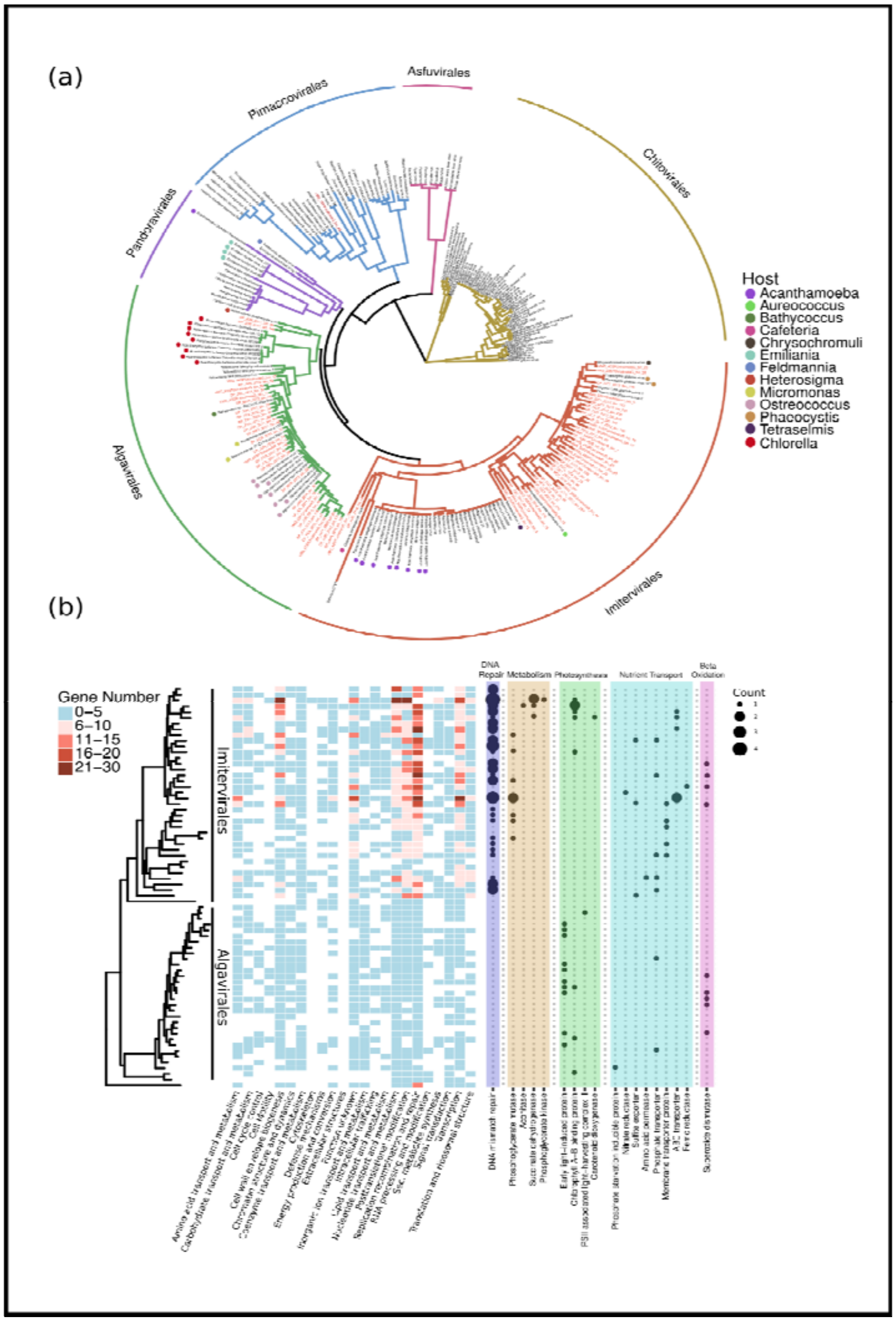


Figure 6: Phylogenetic and functional potential of NCLDV MAGs. (a) Multigene Maximum likelihood phylogenetic tree from a concatenated alignment of five conserved NCLDV genes from reference genomes (black) and MAGs (red). The node color indicates the reported host of reference NCLDVs and the branch color indicates NCLDV order membership. (b) The tree from (a) subset to include only MAGs generated here. The heatmap (center) indicates the number of genes within each COG functional group and the bubble plot (right) indicates the number of genes involved in select metabolic functions.

bioRxiv preprint doi: https://doi.org/10.1101/2024.08.12.607631; this version posted August 15, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

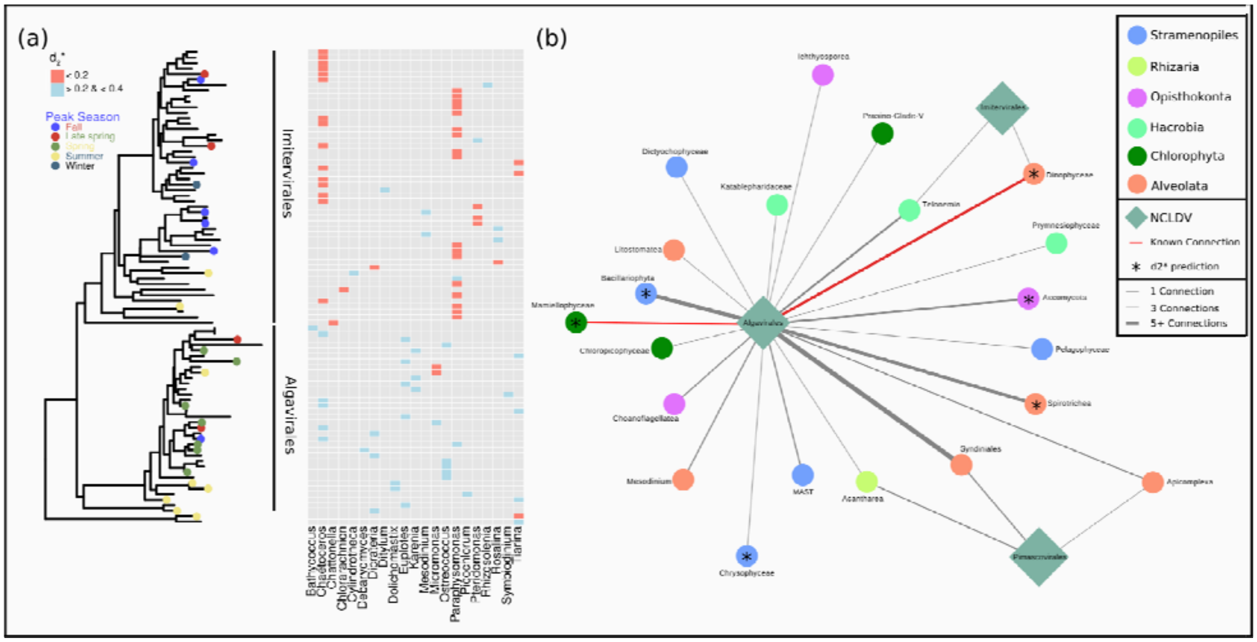
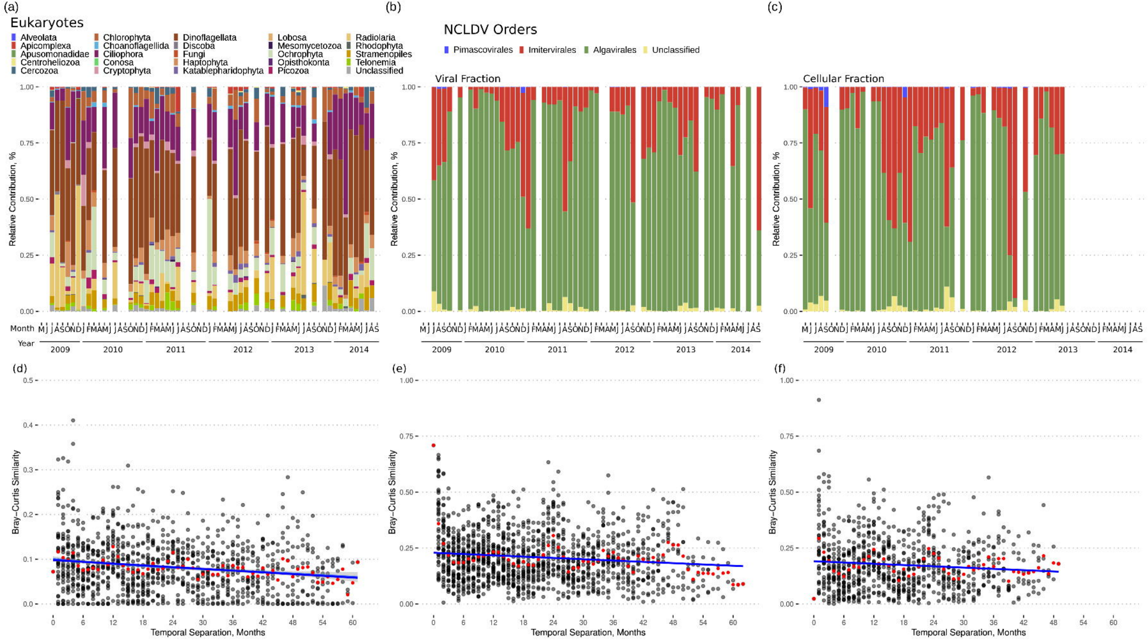
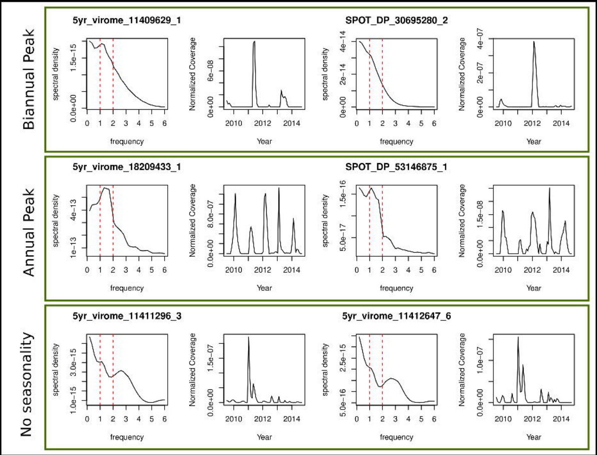
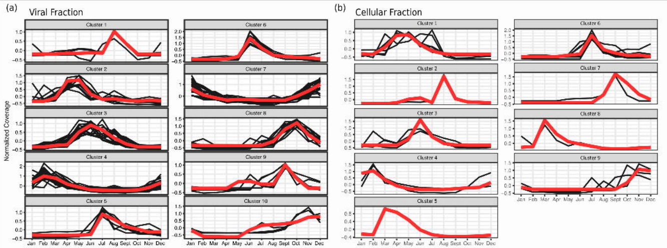
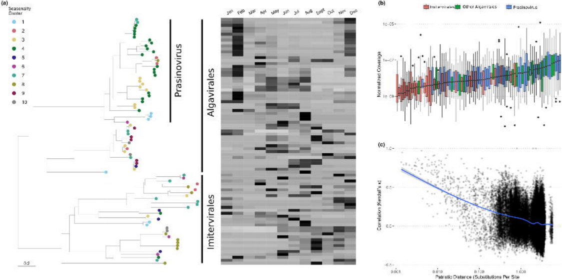


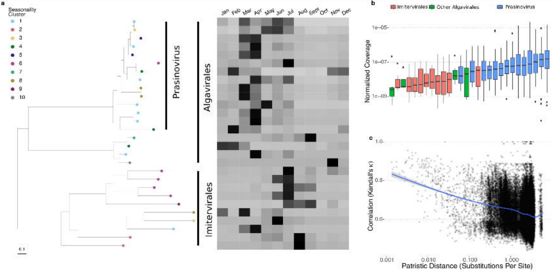
Figure 7: Reconstructed NCLDV MAGs are predicted to infect diverse host taxa. (a) Phylogenetic tree of NCLDV MAGs and their predicted host taxa based on k-mer profile similarity (d2∗) with MMETSP reference genomes. High-confidence (d2∗ < 0.2) host predictions are shown in red, while lower-confidence host predictions are shown in blue. Peak season is also shown for MAGs with significant seasonality. (b) Co-occurrence network of NCLDV PolB phylotypes and eukaryote OTUs using 18S data. A weight cutoff of 0.4 was used to filter connections, and the number of connections is shown by the edge thickness. Known NCLDV and eukaryotic host pairs from cultured isolates are shown in red and connections match and d2* results are shown with a star (*). An expanded network and phylogenetic tree are available in Figure S7.


(a)
Eukaryotes
Alveolata
Apicomplexa
Apusomonadidae
Centroheliozoa
Cercozoa
Chlorophyta
Choanoflagellida
Ciliophora
Conosa
Cryptophyta
Dinoflagellata
Discoba
Fungi
Haptophyta
Katablepharidophyta
Lobosa
Mesomycetozoa
Ochrophyta
Opisthokonta
Picozoa
Radiolaria
Rhodophyta
Stramenopiles
Telonemia
Unclassified
Relative Contribution, %
1.00
0.75
0.50
0.25
0.00
Month
Year
2009
2010
2011
2012
2013
2014
(b)
NCLDV Orders
Pimascovirales
Imitervirales
Algavirales
Unclassified
Viral Fraction
(c)
Cellular Fraction
(d)
Bray−Curtis Similarity
0.5
0.4
0.3
0.2
0.1
0.0
Temporal Separation, Months
0
6
12
18
24
30
36
42
48
54
60
(e)
1.00
0.75
0.50
0.25
0.00
(f)

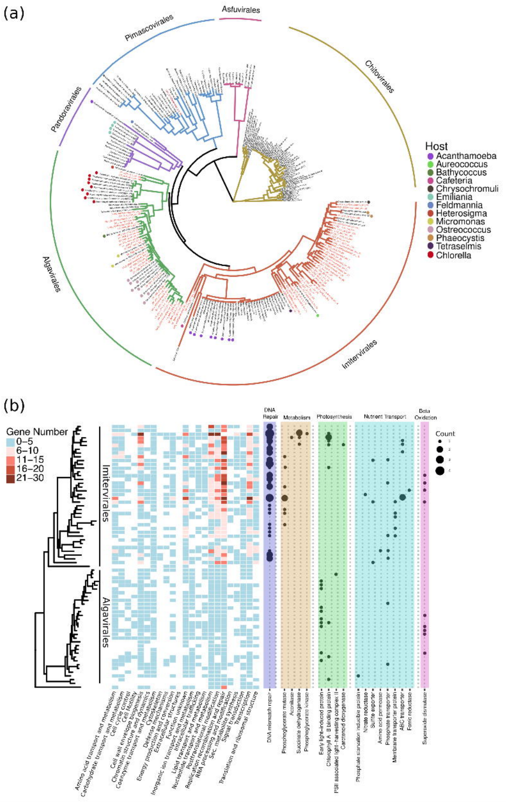

Biannual Peak
5yr_virome_11409629_1
spectral density
0.0e+00
1.5e-15
frequency
0
1
2
3
4
5
6
Normalized Coverage
0e+00
6e-08
Year
2010
2012
2014
SPOT_DP_30695280_2
spectral density
0e+00
2e-14
4e-14
frequency
Normalized Coverage
0e+00
2e-07
4e-07
Year
Annual Peak
5yr_virome_18209433_1
spectral density
1e-13
4e-13
frequency
Normalized Coverage
0.0e+00
8.0e-07
Year
SPOT_DP_53146875_1
spectral density
5.0e-17
1.5e-16
frequency
Normalized Coverage
0.0e+00
1.5e-08
Year
No seasonality
5yr_virome_11411296_3
spectral density
1.0e-15
3.0e-15
frequency
Normalized Coverage
0.0e+00
1.5e-07
Year
5yr_virome_11412647_6
spectral density
5.0e-16
2.5e-15
frequency
Normalized Coverage
0.0e+00
1.0e-07
Year

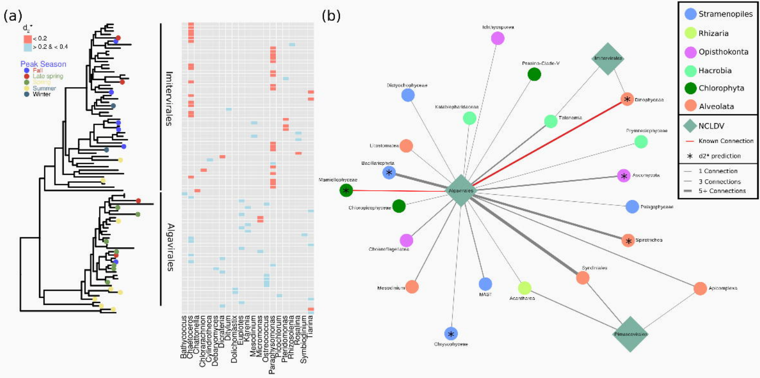

(a)
Viral Fraction
Normalized Coverage
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Cluster 6
Cluster 7
Cluster 8
Cluster 9
Cluster 10
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sept
Oct
Nov
Dec
(b)
Cellular Fraction
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Cluster 6
Cluster 7
Cluster 8
Cluster 9
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sept
Oct
Nov
Dec


(a)
Seasonality Cluster
1
2
3
4
5
6
7
8
9
10
Prasinovirus
Algavirales
Imitervirales
0.2
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sept
Oct
Nov
Dec
(b)
Imitervirales
Other Algavirales
Prasinovirus
1e-05
1e-07
1e-09
Normalized Coverage
(c)
1.0
0.5
0.0
-0.5
Correlation (Kendall's τ)
0.001
0.010
0.100
1.000
Patristic Distance (Substitutions Per Site



a
Seasonality Cluster
1
2
3
4
5
6
7
8
9
10
Prasinovirus
Algavirales
Imitervirales
0.1
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sept
Oct
Nov
Dec
b
Imitervirales
Other Algavirales
Prasinovirus
Normalized Coverage
1e−05
1e−07
1e−09
c
Correlation (Kendall's κ)
1.0
0.5
0.0
0.001
0.010
0.100
1.000
Patristic Distance (Substitutions Per Site)


(a)
Asfuvirales
Pimascovirales
Chitovirales
Pandoravirales
Algavirales
Imitervirales
Host
Acanthamoeba
Aureococcus
Bathycoccus
Cafeteria
Chrysochromuli
Emiliania
Feldmannia
Heterosigma
Micromonas
Ostreococcus
Phaeocystis
Tetraselmis
Chlorella
(b)
Gene Number
0–5
6–10
11–15
16–20
21–30
Imitervirales
Algavirales
Amino acid transport and metabolism
Carbohydrate transport and metabolism
Cell cycle control
Cell Motility
Cell wall envelope biogenesis
Chromatin structure and dynamics
Coenzyme transport and metabolism
Cytoskeleton
Defense mechanisms
Energy production and conversion
Extracellular structures
Function unknown
Inorganic ion transport and metabolism
Intracellular trafficking
Lipid transport and metabolism
Nucleotide transport and metabolism
Posttranslational modification
Replication recombination and repair
RNA processing and modification
Sec. metabolite synthesis
Signal transduction
Transcription
Translation and ribosomal structure
DNA Repair
Metabolism
Photosynthesis
Nutrient Transport
Beta Oxidation
Count
1
2
3
4
DNA mismatch repair
Phosphoglycerate mutase
Aconitase
Succinate dehydrogenase
Phosphoglycerate kinase
Early light-induced protein
Chlorophyll A–B binding protein
PSII associated light-harvesting complex II
Carotenoid dioxygenase
Phosphate starvation inducible protein
Nitrate reductase
Sulfite exporter
Amino acid permease
Phosphate transporter
Membrane transporter protein
ABC transporter
Ferric reductase
Superoxide dismutase


(a)
d_2*
< 0.2
> 0.2 & < 0.4
Peak Season
Fall
Late spring
Spring
Summer
Winter
Imitervirales
Algavirales
Bathycoccus
Chaetoceros
Chattonella
Chlorarachnion
Cylindrotheca
Debaryomyces
Dictyocha
Ditylum
Dolichomastix
Euplotes
Karenia
Mesodinium
Micromonas
Ostreococcus
Paraphysomonas
Picochlorum
Pteridomonas
Rhizosolenia
Rosalina
Symbiodinium
Tiarina
(b)
Ichthyosporea
Prasino-Clade-V
Imitervirales
Dictyochophyceae
Katablepharidaceae
Dinophyceae
Telonemia
Prymnesiophyceae
Litostomatea
Bacillariophyta
Mamiellophyceae
Algavirales
Ascomycota
Chloropicophyceae
Pelagophyceae
Choanoflagellatea
Spirotrichea
Syndiniales
Mesodinium
MAST
Acantharea
Apicomplexa
Pimascovirales
Chrysophyceae
Stramenopiles
Rhizaria
Opisthokonta
Hacrobia
Chlorophyta
Alveolata
NCLDV
Known Connection
d2* prediction
1 Connection
3 Connections
5+ Connections