Biogeography of marine giant viruses reveals their interplay with eukaryotes and ecological functions

Hisashi Endo[1], Romain Blanc-Mathieu[1,2], Yanze Li[1], Guillem Salazar[3], Nicolas Henry[4,5], Karine Labadie[6], Colomban de Vargas[4,5], Matthew B. Sullivan[7,8], Chris Bowler[9,10], Patrick Wincker[10,11], Lee Karp-Boss[12], Shinichi Sunagawa[3], Hiroyuki Ogata[1,*]

Affiliations:

1. Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto, 611-0011, Japan
2. Laboratoire de Physiologie Cellulaire & Végétale, CEA, Univ. Grenoble Alpes, CNRS, INRA, IRIG, Grenoble, France
3. Department of Biology, Institute of Microbiology and Swiss Institute of Bioinformatics, ETH Zürich, Zürich 8093, Switzerland
4. CNRS, UMR 7144, Station Biologique de Roscoff, Place Georges Teissier, 29680 Roscoff, France.
5. Sorbonne Universités, UPMC Université Paris 06, UMR 7144, Station Biologique de Roscoff, Place Georges Teissier, 29680 Roscoff, France.
6. Genoscope, Institut de Biologie François-Jacob, Commissariat à l'Énergie Atomique (CEA), Université Paris-Saclay, Évry, France.
7. Department of Microbiology, The Ohio State University, Columbus, OH 43210, USA
8. Department of Civil, Environmental and Geodetic Engineering, The Ohio State University, Columbus, OH 43210, USA
9. Institut de Biologie de l'ENS (IBENS), Département de biologie, École normale supérieure, CNRS, INSERM, Université PSL, Paris 75005, France
10. Research Federation for the study of Global Ocean Systems Ecology and Evolution, FR2022/Tara Oceans GOSEE, 3 rue Michel-Ange, 75016 Paris, France
11. Génomique Métabolique, Genoscope, Institut de Biologie François Jacob, Commissariat à l'Énergie Atomique (CEA), CNRS, Université Évry, Université Paris-Saclay, Évry, France.

12. School of Marine Sciences, University of Maine, Orono, ME, USA

***Corresponding author**:

H. Ogata, E-mail: ogata@kuicr. kyoto-u.ac.jp, Phone: +81-774-38-3270

Abstract

Nucleocytoplasmic large DNA viruses (NCLDVs) are ubiquitous in marine environments and infect diverse eukaryotes. However, little is known about their biogeography and ecology in the ocean. By leveraging the *Tara* Oceans pole-to-pole metagenomic data set, we investigated the distribution of NCLDVs across size fractions, depths and biomes, as well as their associations with eukaryotic communities. Our analyses revealed a heterogeneous distribution of NCLDVs across oceans, with an elevated uniqueness in polar biomes. The community structures of NCLDV families were correlated with specific eukaryotic lineages including many photosynthetic groups. NCDLV communities were generally distinct between surface and mesopelagic zones, but at some locations, they exhibited a high similarity between the two depths. This vertical similarity was correlated to surface phytoplankton biomass but not to physical mixing processes, suggesting the potential role of vertical export in structuring mesopelagic NCLDV communities. These results underscore the importance of the coupling between NCLDVs and eukaryotes in biogeochemical processes in the ocean.

Introduction

The photic zone is the most productive layer of the ocean, containing a wide variety of microorganisms such as bacteria, autotrophic and heterotrophic protists and multicellular organisms. The population dynamics of these organisms determine the flows of energy and materials through marine food webs, playing a fundamental role in ecosystem functioning and biogeochemical cycles in the ocean[1,2]. Viruses exert a top-down control on marine organisms and release material to the pools of particulate and dissolved organic matter[3]. This material and remineralized inorganic nutrients are utilized by autotrophic and mixotrophic phytoplankton[4]. The recycling of nutrients in the surface layer potentially reduces the transfer of fixed organic carbon to higher trophic levels and the deep sea[5,6]. However, it is also possible that viruses enhance downward carbon flux by facilitating cell aggregation and producing carbon-enriched materials from infected cells[7-9].

Nucleocytoplasmic large DNA viruses (NCLDVs or so-called “giant viruses”) represent a monophyletic group of viruses that infect a variety of eukaryotic lineages[10-12]. Studies focusing on conserved marker genes such as family B DNA polymerase (*polB*) have revealed that NCLDVs are highly diverse and abundant in aquatic environments[13-16]. The diversity of a family of NCLDVs, namely *Mimiviridae*, exceeds that of bacteria and archaea in the ocean[17] and their richness in a few liters of seawater can reach more than 5,000 operational taxonomic units[18]. More recently, several thousand draft genomes (i.e., metagenome-assembled genomes; MAGs) of NCLDVs were constructed from environmental sequences, thanks to the development of high-throughput sequencing and bioinformatics technologies[19,20]. However, the global biogeography of marine NCLDVs still remains under-explored.

A growing number of marine eukaryotes have been reported as host organisms of NCLDVs, particularly phytoplankton groups such as haptophytes, chlorophytes and dinoflagellates[21-23]. Other eukaryotic lineages, including non-photosynthetic organisms such as bicosoecids and choanoflagellates, have also been reported as host organisms of

NCLDVs in marine environments[24,25]. These studies collectively suggest the ecological importance of NCLDVs in the ocean via top-down effects on eukaryotic communities. However, our knowledge of NCLDV-host relationships is highly limited, given the large phylogenetic diversities of NCLDVs and microeukaryotes.

Here we reveal patterns in the global biogeography of NCLDVs using the metagenomic data from the *Tara* Oceans project. The metagenomic data cover varying geographic regions including polar and deep-sea ecosystems, in which NCLDVs are under-researched[26-28]. We constructed NCLDV taxonomic abundance profiles for 283 samples, representing two viral size fractions, three ocean depth ranges (surface, deep chlorophyll maximum and mesopelagic), and four biomes (coastal, trades, westerlies and polar). The global biogeography of NCLDVs derived from these data reveals strong associations between NCLDVs and eukaryotic microorganisms. Furthermore, vertical connectivity of NCLDV communities indicates a possible mechanism for how mesopelagic NCLDV communities are structured with respect to ocean biogeochemical processes.

Results

NCLDV phylotypes detected in *Tara* Oceans metagenomes

We detected 6,818 PolBs affiliated with NCLDVs in the second version of the Ocean Microbial Reference Gene Catalog (OM-RGC.v2)[28] using the pplacer phylogenetic placement method[29] (see methods for details). The OM-RGC.v2 was built based on 370 *Tara* Oceans metagenomes from femto- (<0.2 µm; 151 samples), pico- (0.22–1.6 or 0.22–3.0 µm; 180 samples) and other (39 samples) size fractions. After removing 32 samples with a low NCLDV frequency and 55 samples from non-target size fractions and depths, the remaining 283 samples contained 6,783 NCLDV PolB sequences. The pplacer classified these PolBs into nine NCLDV families/lineages. The number of phylotypes (distinct *polB* at 95% nucleotide sequence identity) was the largest in *Mimiviridae* (5,091 phylotypes), followed by *Phycodnaviridae* (981 phylotypes). The number of phylotypes taxonomically assigned to *Iridoviridae*, *Medusavirus* and *Asfarviridae*, were 239, 120 and 109, respectively. We also detected PolBs assigned to *Pithoviridae* (93), *Ascoviridae* (78), *Poxviridae* (51) and *Marseilleviridae* (21). However, *Poxviridae* was omitted from our discussion as the environmental gene sequences were distantly related to known *Poxviridae*. Rarefaction analysis showed that, at the end of sampling, the number of NCLDV phylotypes increased by less than 0.01% per sample for all samples, and ranged from 0.02% to 0.32% when samples were divided into different size fractions, depths and biomes (Extended Data Fig. 1).

To examine detailed phylogenetic affiliation and to visualize the dispersal characteristics of each NCLDV phylotypes detected by pplacer, we constructed a phylogenetic tree using selected PolB sequences (Extended Data Figs. 2–4). Among the *Mimiviridae* family, genes closely related to the algal-infecting subfamily, recently proposed as “Mesomimivirinae” (e.g., AaV, CeV, pkV, PgV, PoV and TetV)[30], which infect pelagophytes (the genus *Aureococcus*), haptophytes (the genera *Haptolina*, *Prymnesium* and *Phaeocystis*), and chlorophytes (the genera *Pyramimonas* and *Tetraselmis*), were relatively abundant. On the other hand, only a few sequences were affiliated with the subfamilies “Megamimivirinae” and “Klosneuvirinae” except the *Cafeteria roenbergensis virus* (CroV), which is the only member of “Megamimivirinae” isolated from the marine environment[24]. Among *Phycodnaviridae,* the genus *Prasinovirus* (e.g., BpV, MpV, OtV and OlV), which infect chlorophyte genera such as *Bathycoccus*, *Micromonas* and *Ostreococcus*, showed the highest richness.

Heterogeneity in NCLDV community structure across size, depth and biomes

The dominant NCLDV taxa detected from all sample locations and depths in the pico-size fraction were *Mimiviridae* and *Phycodnaviridae*, with average contributions of 64.6% and 25.4%, respectively (Fig. 1A). The dominant groups of NCLDVs varied widely among sites and depths in samples from the femto-size fraction (Fig. 1B). In this fraction, *Phycodnaviridae* and *Asfarviridae* had relatively high contributions to the total

NCLDVs with the mean values of 29.7% and 19.9%, respectively. *Mimiviridae* and *Ascoviridae* were also important contributors with mean values of 12.2% and 11.1%, respectively.

A non-metric multidimensional scaling (NMDS) analysis showed that NCLDV assemblages clustered according to size fraction, depth and biome (Fig. 2A–2C). Significant differences in NCLDV community composition were detected among all categories (PERMANOVA, p <0.01), and size fraction, depth and biome explained 5.5%, 4.3% and 10.9% of the total variance, respectively.

Taxonomic richness (i.e., number of phylotypes) and Shannon's diversity index were used to investigate variation in NCLDV community diversity. In this study, we analyzed the samples from all depths and size fractions to compare diversity differences among depth ranges, although latitudinal trend in Shannon's diversity for pico-sized communities from the surface was reported previously[31]. In the pico-size fraction, mean values for NCLDV richness at the surface and in the DCM layer were about 1.7 times higher than that in the mesopelagic layer (Kruskal-Wallis and Dunn's post hoc test, p <0.01) (Extended Data Fig. 5A). In the femto-size fraction, NCLDV richness was significantly higher at the surface and MES layer than in the DCM layer (Dunn's test, p = 0.04–0.05), although the differences were small and not consistent with the pico-size fraction.

High uniqueness of NCLDV phylotypes in the Arctic Ocean

We analyzed the overlap and uniqueness of NCLDV phylotypes across different ecological zones (i.e., size fraction, depth and biome) to evaluate their ability to disperse across different environments. Each ecological category was divided into two major groups (i.e., pico- and femto-sizes, euphotic and mesopelagic zones, and polar and non-polar biomes), because the NCLDV community in mesopelagic zone or polar biome was separated most significantly from other depths or biomes (Fig. 2). We found 4,003 (59.0% to the total NCLDVs) shared NCLDV phylotypes across size fractions, 4,737 (69.8%)

shared phylotypes across depth ranges, and 1,950 (28.7%) shared phylotypes across biomes (Fig. 3A). Only twelve unique phylotypes were detected in the femto-size fraction, whereas 2,768 unique phylotypes were identified in the pico-size fraction. The euphotic zone (surface and DCM) harbored 1,986 unique phylotypes, whereas the aphotic mesopelagic zone had only 60 unique phylotypes. The polar biome (the Arctic and the Southern Ocean) included 620 unique NCLDV phylotypes, whereas 4,213 unique NCLDVs were detected in non-polar biomes (i.e., trades, westerlies and coastal).

To further characterize regional differences in the NCLDV community, we investigated the total and unique NCLDV phylotypes observed in nine geographic regions and the phylotypes shared among regions. The total number of phylotypes was relatively high in the Atlantic, Pacific and Indian Oceans and in the Mediterranean Sea, with values of between 3,665 and 4,685 (Fig. 3B). Lower numbers of NCLDV phylotypes were identified from the Red Sea (2,653) and the Arctic Ocean (2,467). The Southern Ocean presented the lowest number of NCLDV phylotypes (561), although this was based on only 5 samples. The Arctic Ocean samples displayed a high number of unique NCLDV phylotypes (551), which corresponded to 22.3% of the total phylotypes detected in this region. In contrast, the number of unique phylotypes from other regions ranged from 0 to 134 (0.0% to 3.4%).

There was no linear or saturation trend in the number of total or unique NCLDV phylotypes with increasing sample size (Fig. 3C). The high proportion of unique phylotypes in the Arctic Ocean was not a function of sample size, although the number of total phylotypes detected in the Southern Ocean may be limited by the low number of samples. The phylogenetic positions of unique NCLDVs from the polar biome were dispersed across most of the NCLDV families (Fig. 4)

NCLDV distributions correlate with eukaryotic communities

A partial Mantel test was conducted to assess community associations among the NCLDV families/lineages and major eukaryotic lineages. The pairwise partial correlation coefficients (Spearman's ρ) varied from –0.17 to 0.76 (Fig. 5A), and 93.6% of the examined pairs (225 out of 234 for the pico-size fraction and 213 out of 234 for the femto-size fraction) showed statistically significant correlations (p <0.01, permutation test) after false discovery rate (FDR) correction. Pairs from pico-sized NCLDV communities with a correlation coefficient ≥0.53 were considered to represent strong positive associations, because 8 out of 9 known marine virus-host lineage associations were recovered by this criterion (Figs. 5A and 5B). Using this threshold, 30 out of 234 NCLDV-eukaryote lineage pairs were found to have strong linkages (Fig. 5C). The NCLDV families/lineages were generally highly correlated with the known host groups among autotrophic and mixotrophic microalgae (haptophytes, chlorophytes, dinophytes, pelagophytes and raphidophytes) (ρ = 0.54–0.67). Interestingly, *Mimiviridae* was strongly correlated with chrysophyte microalgae (ρ = 0.65), which are not currently known as NCLDV hosts. Other than algal lineages, a strong positive correlation was found between *Mimiviridae* and heterotrophic eukaryote choanoflagellates (ρ = 0.76), which are a known lineage of *Mimiviridae*. A group of non-photosynthetic heterokonts bicosoecids are also a known host of the *Mimiviridae* species CroV in marine environments, but this group was not highly correlated with *Mimiviridae* (ρ = 0.30).

Potential chrysophyte viruses constitute novel clades of *Mimiviridae*

To explore possible associations between NCLDVs and chrysophytes as indicated by the Mantel's regression analysis (Fig. 5C), we tested for chrysophyte-derived genes in the metagenome-assembled genomes (MAGs) of NCLDVs generated by Schultz et al. (2020)[19] and Moniruzzaman et al. (2020)[20]. The results showed that 89 (82 after removing redundancy) out of 2,263 MAGs contained genes closely related to the transcripts of the chrysophytes (Supplementary Data 1). Comparisons between PolB sequences revealed 27 PolBs from the OM-RGC.v2 that were closely related to the NCLDV MAGs with chrysophyte homologs. Most of these PolBs constituted novel clades within the branches of *Mimiviridae* (Fig. 4; Extended Data Fig. 4). We confirmed that other genes in the contigs that contained chrysophyte homologs are highly similar to the *Mimiviridae* or *Phycodnaviridae* sequences in many cases (Extended Data Fig. 6).

Vertical connectivity of NCLDV communities

The vertical connectivity of NCLDV communities was investigated using Bray-Curtis community similarity measures to compare between epipelagic (surface or DCM) and mesopelagic samples at individual sampling locations. The Bray-Curtis similarities were less than 0.10 for about half of the tested locations (20 out of 36 surface sites and 13 out of 26 DCM sites; Fig. 6A; Extended Data Fig. 7A). All sites in the Arctic Ocean and several sites in tropical and subtropical regions showed relatively high similarities between the two depth (0.15 to 0.60). The NCLDV community similarity value was positively correlated with the chlorophyll *a* concentration in the epipelagic layer (Spearman's $\rho = 0.52$, $p < 0.01$, asymptotic *t* approximation, n = 36 for surface; $\rho = 0.44$, $p = 0.02$, n = 25 for DCM) and NCLDV richness in the mesopelagic layer ($\rho = 0.82$, $p < 0.01$, n = 36 for surface; $\rho = 0.70$, $p < 0.01$, n = 26 for DCM) (Figs. 6B and 6C; Extended Data Figs. 7B and 7C). We also evaluated relationships between NCLDV vertical similarity and physical environmental factors including: the sampling depth of mesopelagic water, the mixed layer depth, and the temperature difference between epipelagic and mesopelagic waters. No significant correlations were detected among these parameters ($p > 0.05$, n = 32–36 for surface samples and n = 25–26 for DCM samples) (Figs. 6D–F; Extended Data Figs. 7D–F).

We plotted correlations among the relative contributions of NCLDV phylotypes between the euphotic and aphotic zones at all sampling locations (Extended Data Figs. 8 and 9). Where there was a strong similarity in the NCLDV community found at different depths, *Phycodnaviridae* generally contributed highly to samples from the Arctic Ocean (e.g., TARA stations 158, 201 and 209), and both *Mimiviridae* and *Phycodnaviridae* contributed strongly in tropical and subtropical regions (e.g., stations 72, 110 and 122).

Discussion

We investigated the diversity and community structure of NCLDVs based on metagenomic PolB sequences collected from the world oceans. NCLDV communities differed substantially between pico- and femto- size fractions (Fig. 1). NCLDV communities in the pico-size fractions were dominated by *Mimiviridae* and *Phycodnaviridae*, regardless of sampling location or depth (Fig. 1A). In marine environments, species from the haptophytes (the genera *Prymnesium*, *Haptolina*, and *Phaeocystis*), chlorophytes (*Pyramimonas*), pelagophytes (*Aureococcus*), bicosoecids (*Cafeteria*) and choanoflagellates (*Bicosta*) are known hosts of *Mimiviridae*, while species of haptophytes (*Emiliania*), chlorophytes (*Ostreococcus*, *Micromonas* and *Bathycoccus*) and raphidophytes (*Heterosigma*) have been reported as *Phycodnaviridae* hosts (Virus-Host DB)[32]. Although the dominance of *Mimiviridae* and *Phycodnaviridae* have been reported in previous studies, mainly from coastal seawater[13,14], our results demonstrate the ubiquitous nature of these protist-infecting viruses across world ocean biomes. It is worth noting that most of the NCLDVs (99.7%) detected from the femto-size fraction were also present in the pico-size fraction (Fig. 3A), despite the large differences in relative abundance between two size fractions at each location. Therefore, the abundance information can be important for characterizing the differences of NCLDV communities. A proportion of the NCLDVs in the pico-size fraction were present within infected cells, because cell sizes of some host species such as *Aureococcus anophagefferens* and *Micromonas pusilla* are less than 3 μm. Thus, the abundance of these lineages in the pico-size fraction may be partly enriched by the viruses replicating inside their hosts.

In addition to *Phycodnaviridae* and *Mimiviridae*, *Asfarviridae* also contribute an important proportion of NCLDVs in the femto-size fraction of most euphotic zones (Fig. 1B). Although very limited information is available regarding the natural hosts for this group, a representative *Asfarviridae*-like species in marine environments is *Heterocapsa circularisquama* DNA virus (HcDNAV), which infects the red-tide-forming dinoflagellate *H. circularisquama*[33]. In the terrestrial ecosystem, this viral family is known to infect a wide variety of organisms such as amoebozoa, arthropods and mammals[32,34]. Given the broad range of host species for this viral lineage, there may be an unknown but wide-spread host taxa for *Asfarviridae* in the ocean.

Our study revealed a heterogeneous pattern in the distribution of NCLDVs across the oceans of the world (Fig. 2C). Although there are limited studies available on the factors controlling the large-scale distribution of viruses, it is widely accepted that both deterministic (environmental factors and inter-specific interactions) and stochastic processes (e.g., immigration and speciation) are important in making up microbial assemblages[35-37]. The distribution and diversity of viruses would not be directly affected by environmental variables such as temperature and nutrient availability, but is directly influenced by the geographic ranges of their host species[3,38]. Recent work with cyanophages demonstrated that a significant number of free-living viruses are locally produced through active infection rather than from migration[39]. Therefore, we expect that viral community structure will reflect host distribution as well as infectious activity.

Despite significant differences in community composition across oceanic biomes, we found that most NCLDV phylotypes are dispersed throughout tropical and temperate regions (Figs. 3A and 3B), presumably following their host community composition, which is primarily determined by temperature[40]. However, the polar biome (mainly the Arctic Ocean) constitutes a “hotspot” of unique NCLDV phylotypes from a wide variety of families, despite having a low total richness in comparison to other regions (Figs. 3B and 3C). We revealed that NCLDVs unique to non-polar biome were also abundant (Fig. 4), indicating a strong separation of NCLDV communities between polar and non-polar biomes. A geographical barrier and steep environmental gradients may underlie this distinct ecosystem structure (i.e., different host communities and their productivity) in the Arctic Ocean[27,28,31]. Moreover, the Arctic Ocean is characterized by high amounts of river discharge, contributing more than 10% to global runoff flux[41]. Consequently, biological processes in the Arctic may be influenced by river inputs from terrestrial ecosystems.

These factors may collectively contribute to the remarkable number of unique NCLDV phylotypes found in the Arctic, that were undetectable in other regions. The biogeography of NCLDVs on a global scale implies a tight link between the NCLDVs and the distribution of their hosts, which is strongly influenced by physicochemical and biological factors.

Tight coupling between NCLDVs and their hosts was further corroborated by our partial Mantel statistics, which described both known virus-host interactions and additional but currently unrecognized associations between viruses and eukaryotic lineages at the community level. Using the pico-sized NCLDV community, we detected almost all known virus-host interactions, except for those involving Bicoecea (Fig. 5C). This demonstrates that distance-based correlation analysis using global ocean samples is useful for detecting virus-host interplay in natural environments, although the validations of the previously unknown associations remain to be further explored. Strong positive relationships between NCLDVs and eukaryotes involved many phytoplankton lineages including haptophytes, chlorophytes, dinophytes, pelagophytes and raphidophytes, all of which include known host lineages of NCLDVs (Fig. 5C). Strong correlations were also detected with heterotrophic choanoflagellates, which have recently been identified as a novel host of *Mimiviridae*[25]. Some NCLDVs, especially *Mimiviridae*, had strong correlations with chrysophytes, although no host species have yet been reported for this lineage. Many environmental NCLDV genomes were found to encode genes that are likely to be derived from marine chrysophytes (Supplementary Data 1–3). Taxonomic analyses based on PolB phylogeny and homology search revealed that most of these phylotypes represent previously unknown clades of the *Mimiviridae* tree (Extended Data 4 and 6; Supplementary Data 4), suggesting that chrysophytes may be an important host lineage of *Mimiviridae* in the ocean.

The global distribution of NCLDVs are determined by the geographic ranges of their host organisms. Therefore, the virus-eukaryote associations that we detected likely arose under these constraints. On the other hand, it is expected that NCLDVs influence the abundance of eukaryotes at a local scale. Previous studies show that bacterial viruses have an important role in determining bacterial mortality, because they substantially outnumber their hosts and have highly specific infection mechanisms[42]. Similarly, NCLDVs are reported to be more abundant than their host cells and have high infection specificity[11,14,43]. For example, *Emiliania huxleyi* viruses (EhVs) of the *Phycodnaviridae* family are responsible for almost all of the mortality of the haptophyte *E. huxleyi* during blooms[22,44,45]. Another field study suggests that viral lysis can explain a greater proportion of phytoplankton mortality than grazing by zooplankton[6]. These studies, combined with the global associations that were detected in this study, emphasize the potential importance of NCLDVs in structuring eukaryotic communities.

Our results indicate that marine phytoplankton lineages could represent one of the most important host groups of NCLDVs. Therefore, NCLDVs could be involved in the regulation of biogeochemical processes mediated by phytoplankton. We investigated this by assessing the vertical connectivity of viral communities. The NMDS analysis showed clear differences between the NCLDV community composition of epipelagic (euphotic) and mesopelagic (aphotic) zones at most sampling sites (Fig. 2B). Similar results were also reported for phage communities in the Pacific Ocean[46]. The vertical separation of viral communities may be caused by the stable stratification below the mixed layers (typically above 200 m depth), which severely inhibits vertical water exchange. Despite this limitation, mesopelagic ecosystems shared a significant number (98.7%) of NCLDV phylotypes with the upper epipelagic layers (Fig. 3A), suggesting the vertical connectivity of NCLDVs and their local adaptation. Indeed, some mesopelagic NCLDV communities were very similar to surface communities (Fig. 6A and Extended Data Fig. 7A). This implies that the surface and mesopelagic NCLDV communities may be connected at some locations. The major source of energy and materials in the mesopelagic layer is the gravitational export of organic particles from the surface layer (i.e., the biological carbon pump)[47-49]. Therefore, some surface viruses may be exported to mesopelagic layers with sinking aggregated phytoplankton cells[50-52].

A significant positive correlation existed between surface phytoplankton biomass and NCLDV community similarity across depths (Fig. 6B and Extended Data Fig. 7B). Since highly productive areas are likely to have a greater flux of settling particles to the deep layers, this result supports the idea that NCLDVs are transported with the sinking particles. High vertical connectivity was consistently associated with an increase in NCLDV richness in the mesopelagic zone (Fig. 6C and Extended Data Fig. 7C). Previous studies showed that sinking particles can transfer bacterial and phage populations to the deep layer[52,53]. Mestre et al.[52] demonstrated that particle-attached prokaryotes had higher capacity for immigration than free-living ones. Based on the particle-driven vertical dispersion model, we can expect that NCLDVs, inside or attached to their host cells or cell debris, might be preferentially exported into the deep sea. Numerous studies based on sediment trap measurement have shown that larger phytoplankton, such as diatoms, contribute strongly to vertical flux because of their high sinking velocities[54,55]. However, recent studies show that smaller phytoplankton including haptophytes and chlorophytes, known hosts of marine NCLDVs, also contribute greatly to downward carbon export[8,9,56]. The high vertical connectivity of NCLDVs was not affected by the extent of the depth range nor by proxies for vertical mixing (Figs. 6D–F and Extended Data Figs. 7D–F), indicating that the migration of NCLDVs occurred regardless of physical processes such as upwelling, turbulent mixing, and convection. This result suggests that sinking export is a major source of a variety of NCLDVs to deeper waters, where NCLDV diversity is relatively low without this effect. A recent study revealed that some *Phycodnaviridae* and *Mimiviridae* potentially accelerate biological carbon export from the productive surface layer to deep layers, presumably by promoting cell death and aggregation of their host species[57]. *Phycodnaviridae* and *Mimiviridae* also contributed strongly to high vertical connectivity in our study (Extended Data Figs. 8 and 9). The infection of the coccolithophore by the *Phycodnaviridae* EhV was observed to facilitate the sinking of host cells, likely by enhancing the production of transparent exopolymer particles and subsequent aggregation[9]. Therefore, the high vertical connectivity of NCLDVs detected in our analysis may be partly associated with enhanced vertical export of their infected hosts.

The present study expands our knowledge of marine NCLDV biogeography. Most NCLDV phylotypes are ubiquitously distributed over the oceans of the globe, although a high proportion of unique NCLDVs was detected in the Arctic Ocean. Our comparison of community distribution patterns highlighted the tight interplay between NCLDVs and microeukaryotes. As marine ecological and biogeochemical processes are governed primarily by microbes, NCLDVs would have an important influence on the dynamics of marine systems. We also identified unexpected similarity of NCLDV communities between surface and deep waters at some locations. This supports the idea that viral activity may be related to the strength of the biological carbon pump, because the efficiency and sinking rate of export production depends largely on surface phytoplankton composition and their infection status[8,9,55,58]. Our findings underscore the importance of NCLDVs as a component of marine microbial communities, and contribute to refine our knowledge of marine ecosystems, a key regulator of the Earth's climate.

Methods

Sample collection

Metagenomic datasets were generated from samples collected by the *Tara* Oceans expeditions from 2009 to 2013[26-28,31,59]. The second version of the Ocean Microbial Reference Gene Catalog (OM-RGC.v2) is a non-redundant gene catalog constructed from 370 metagenomic samples from the *Tara* Oceans project[28] (https://www.ocean-microbiome.org). The catalog includes 46,775,154 genes in total, and the gene abundance profiles are expressed as the sum of within-reads aligned base pairs normalized by gene length, in *Tara* Oceans samples[28].

Recruitment of NCLDV marker genes from the OM-RGC.v2

To assess the community composition of NCLDVs, we used family B DNA polymerase (*polB*) as a marker gene of NCLDVs. Initially, amino acid sequences of the OM-RGC.v2 were searched against an in-house profile hidden Markov model (HMM) of NCLDV PolB sequences using the software HMMER, hmmsearch (version 3.1)[60] with a threshold E-value <1×10^{-5}. Consequently, 29,315 PolB sequences were obtained from the OM-RGC.v2, although this collection included sequences other than NCLDVs. To remove the sequences not derived from NCLDVs and classify the taxonomic identity of each NCLDV sequence, phylogenetic mapping was performed within known PolB sequences. A maximum-likelihood (ML) reference phylogenetic tree was built based on 211 PolB reference protein sequences from eukaryotes, bacteria, archaea, phages and NCLDVs. These sequences were aligned using the default settings of the multiple sequence alignment program MAFFT-linsi (version 7)[61] and ML tree was constructed with the use of randomized axelerated maximum likelihood (RAxML) program (version 7.2.8)[62]. In the reference trees, we included sequences from eight proposed families of NCLDVs[63]: *Mimiviridae* (synonymous with *Megaviridae*), *Phycodnaviridae*, *Pithoviridae*, *Marseilleviridae*, *Ascoviridae*, *Iridoviridae*, *Asfarviridae*, and *Poxviridae* (Extended Data Figs. 2–4). A sequence from a novel NCLDV clade *Medusavirus* was also included as a reference[64]. Query sequences were aligned against the reference alignment using the MAFFT 'addfragments' option, and then mapped onto the reference tree using the software program pplacer[29].

Abundance profiling of NCLDVs

We used the abundance profile of NCLDV genes from the OM-RGC.v2 to evaluate the relative frequency and diversity of NCLDVs. In the abundance matrix, we only included samples from the pico-size (0.22–1.6 or 0.22–3.0 µm) and femto-size (<0.22 µm) fractions. Samples used in the analysis were from three depth ranges: the surface (2–9 m), the deep chlorophyll maximum (DCM, 15–180 m) and the mesopelagic (MES, 250–1,000 m). The sum of length-normalized PolB abundances ranged from 5.3 to 22,847.5 across samples. The samples containing low PolB abundances tended to yield lower diversity estimates (i.e., number of phylotypes and Shannon's entropy) (Extended Data Fig. 10). To avoid bias due to the low sequencing effort, samples for which the sum of length-normalized PolB abundance was less than 50 (set as a proxy for low NCLDV frequency) were removed from the analysis. The abundance matrix was then standardized by the sample with the lowest sum of length-normalized PolB abundance value. The minimum value of PolB abundance among NCLDV phylotypes in the sample having the lowest sum of length normalized PolB was set as the cutoff threshold. For each sample, NCLDV phylotypes with a length-normalized abundance of less than this threshold were treated as absent. A sample of a femto-size fraction of surface water from station 155 was also removed, because it contained only one NCLDV PolB after standardization. Consequently, our dataset was comprised of 283 samples (172 pico-fraction samples and 111 femto-fraction samples), covering 88 sampling sites. These sites were categorized into four biomes (coastal, trades, westerlies and polar biomes) according to latitude or distance from the shore, and nine oceanic regions, as defined by Longhurst[65] (Supplementary Table 1).

Phylogenetic tree construction

To construct a phylogenetic tree, the NCLDV-derived PolB sequences obtained from the OM-RGC.v2 were filtered by length (≥700 amino acid sequences) because the inclusion of short sequences yields unreliable phylogenies. Amino acid sequences from the resulting 911 genes were aligned with known NCLDV sequences using the *linsi* option from the MAFFT. The ML tree was constructed using RAxML with the use of a known NCLDV sequence tree as a backbone constraint. We confirmed the validity of the pplacer family assignment for 905 out of 911 selected sequences. The remaining six sequences that were incorrectly placed within the phylogenetic tree were removed. The ML tree was visualized using the program iTOL[66].

Prediction of potential chrysophyte viruses using metagenomic assembled genomes

To explore the genomic contents of environmental NCLDVs, we made use of two sets of metagenome-assembled genomes (MAGs) of NCLDVs (GVMAGs high and medium quality[19]; MoMAGs[20]), which were generated from environmental metagenomic datasets collected on global scales. Gene prediction was made for all MAGs using the program GeneMarkS[67], then the predicted genes were searched using BLASTP against a database that combines the NCBI Reference Sequence database (RefSeq release 90) and the marine microbial eukaryote transcriptomes project (MMETSP) database[68]. We identified MAGs whose genes exhibited the best hit to transcripts of chrysophytes with >50% amino acid identity and >100 alignment length (Supplementary Data 1). For these MAGs, we checked the redundancy between the MoMAG and GVMAG datasets using average nucleotide identity of ≥95% and an alignment fraction of ≥50% with FastANI (version 1.3)[69]. Although seven MAGs were found to be overlapped between the two datasets (Supplementary Data 1), all of the MAGs were retained for downstream analyses as these had different contig structures. The chrysophyte-related genes were considered potential candidates for horizontal gene transfer between chrysophytes and NCLDVs, and were BLASTP searched against the RefSeq database for additional functional annotation (Supplementary Data 2). We then extracted PolB sequences from the NCLDV MAGs which had a chrysophyte-related gene using the HMMER hmmsearch program. These PolBs were BLASTP searched against the NCLDV PolBs from the OM-RGC.v2. MAG-derived PolBs aligned with over 700 amino acid sequences with >90% identity were assigned to the PolB phylotypes derived from the OM-RGC.v2 (Supplementary Data 3). Phylogenetic affiliations of PolB from the chrysophyte-related MAGs were confirmed using a phylogenetic tree. To further test the credibility of our analysis, we checked other genes on the contigs that harbored the chrysophyte homologs using BLASTP against the RefSeq database (Supplementary Data 4; Extended Data Fig. 6).

Diversity analyses

Diversity and multivariate analyses were performed using the statistical software R (version 3.6.2) (https://www.r-project.org/). To evaluate the diversity of each sample, the number of NCLDVs (richness) and Shannon's entropy were assessed by the package 'vegan' (https://cran.r-project.org/web/packages/vegan). NCLDV richness among sizes and depths were compared using a Kruskal-Wallis test followed by Dunn's multiple comparison. Compositional variation among samples was assessed with a non-metric multidimensional scaling (NMDS) ordination based on Bray-Curtis dissimilarity. Statistical significance of differences among the sample groups (size, depth and biomes) was tested using a permutational multivariate analysis of variance (PERMANOVA)[70] with 9,999 permutations.

Partial Mantel test

A partial Mantel test was performed to assess the correlation between two multivariate matrices while controlling the potential effects of geographic distance (spatial autocorrelation) using the R package 'vegan'. Abundance matrices for the NCLDV and eukaryotic lineages were constructed from the integrated abundance tables, and the total abundance at each site was normalized to 1. The eukaryote abundance table was constructed based on 18S rRNA gene metabarcoding[71]. Data for NLCDVs were obtained from pico- (0.22–1.6/3.0 µm) or femto-size (<0.2 µm) fractions and for the eukaryotic community from the pico- to meso-size fraction (0.8–2,000 µm). There were 84 overlapping sampling events between pico-size NCLDVs and eukaryotic communities and 55 overlapping sampling events between femto-size NCLDVs and eukaryotic communities. All overlapping samples were derived from the surface or DCM depth layers. Distance matrices for viruses and eukaryotes were calculated using the Bray-Curtis measure. Geographic distances among sample sites were also measured using Haversine distance and were used as a third distance matrix. Partial Mantel correlations were computed between all pairs of distance matrices of eukaryotic communities and NCLDVs with 9,999 permutations for each comparison. The false discovery rate (FDR) was computed using the Benjamini-Hochberg method[72].

Statistical test

Two-sided test was applied for all statistical tests.

Data availability

The complete sequence data of the OM-RGC.v2 and the abundance profile can be downloaded from https://www.ocean-microbiome.org. All sequences of 18S rRNA gene metabarcoding have been deposited at European Nucleotide Archive (ENA) under the BioProject ID PRJEB6610 and PRJEB9737. Environmental metadata are archived at https://doi.pangaea.de/10.1594/PANGAEA.875582. Files used for recruiting NCLDV PolB genes as well as processed abundance profiles of eukaryotes and NCLDVs with corresponding environmental data are available at the GenomeNet FTP: ftp://ftp.genome.jp/pub/db/community/tara/Biogeography/.

Code availability

Custom scripts developed for this study are available at GitHub: https://github.com/HisashiENDO/NCLDV_Biogeography.

Figure legends

Figure 1 Latitudinal patterns in NCLDV community composition. Relative contributions of NCLDV families at each depth range of (A) pico- and (B) femto-size fractions. The number of phylotypes detected in each sample is also indicated with a white circle. Sampling stations were arranged in rows from south to north, and color-coded based on biome (for a map of the sampling stations, please see Salazar et al., 2019[28]).

Figure 2 Community characteristics of NCLDVs. Non-metric multidimensional scaling (NMDS) ordination based on the NCLDV community showing results for all samples (A) and separately for pico- and femto-size fractions (B and C). Sample groups are color-coded by size fraction (A), depth (B) and biome (C). Ellipses represent 90% confidence levels for each group. All group categories are significantly different from each other as analyzed using PERMANOVA (*p* <0.01). Sample sizes for the test are noted in Supplementary Table 1.

Figure 3 Structural differentiation of NLCDV community across ecological zones. (A) Venn diagrams showing the numbers of shared or unique NLCDVs phylotypes across size fractions (left), depths (center) and biomes (right). (B) Map showing the number of total, unique and shared NCLDVs across nine oceanic regions. The map was drawn using the R package 'maps' (https://cran.r-project.org/web/packages/maps). (C) Relationships among sample size and total or unique NCLDVs detected in each region. Abbreviations: SO: Southern Ocean; RS: Red Sea; MS: Mediterranean Sea; NPO: North Pacific Ocean; NAO: North Atlantic Ocean; SAO: South Atlantic Ocean; SPO: South Pacific Ocean; IO: Indian Ocean; AO: Arctic Ocean.

Figure 4 Phylogenetic affiliations of environmental NCLDVs and their dispersal characteristics. Phylogenetic tree constructed from 905 long (≥700 amino acid) PolB sequences from the OM-RGC.v2 and 67 known NCLDV sequences (see also Extended Data Figs. 2–4 for details). The first six layers indicate the occurrence of NCLDVs unique to each size fraction, depth and biome. The outside layer denotes phylogenetic positions of known sequences (color code as in the legend) and the phylotypes closely related (>90% amino acid identity) to those of NCLDV MAGs having chrysophyte homologs (indicated in yellow). Abbreviations: OLPV-2: *Organic Lake phycodnavirus* 2; OLPV-1: *Organic Lake phycodnavirus* 1; CeV: *Chrysochromulina ericina* virus 1; PgV: *Phaeocystis globosa* virus 16T; HeV: *Haptolina ericina* virus RF02; PkV-2; *Prymnesium kappa* virus RF02; TetV-1: *Tetraselmis* virus 1; PoV: *Pyramimonas orientalis* virus 1; AaV: *Aureococcus anophagefferens* virus BtV-01; PkV-1; *Prymnesium kappa* virus RF01; ChoanoV: ChoanoVirus; CroV: *Cafeteria roenbergensis* virus BV-PW1; MpV-1: *Micromonas* sp.

RCC1109 virus MpV1; OlV-1: *Ostreococcus lucimarinus* virus 1; Otv-1: *Ostreococcus tauri* virus 1; Otv-2: *Ostreococcus tauri* virus 2; MpV-12T: *Micromonas pusilla* virus 12T: BpV-1: *Bathycoccus* sp. RCC1105 virus; BCV-FR483: *Paramecium bursaria Chlorella* virus FR-483; ACTV-1: *Acanthocystis turfacea Chlorella* virus 1; PBCV-1: *Paramecium bursaria Chlorella* virus 1; EhV-86: *Emiliania huxleyi* virus 86; FsV: *Feldmannia species* virus; EsV-1: *Ectocampus siliculou* virus 1; *P. salinus*: *Pandoravirus salinus*; *P. dulcis*: *Pandoravirus dulcis*; HaV-1: *Heterosigma akashiwo* virus 1.

Figure 5 **Associations between NCLDVs and eukaryotic communities.** (A) Partial Mantel correlation coefficients (Spearman's *ρ*) between NCLDVs and eukaryotic communities. Each plot shows the value of *ρ* computed based on pico- (x-axis) and femto-sized (y-axis) NCLDV communities. Known virus-host associations are shown as red dots. (B) Histogram and density estimates showing the distribution of *ρ* values in known (red) and unknown (gray) pairs. (C) Pairwise comparisons of the partial Mantel correlation coefficients between NCLDV and eukaryotic lineages. Correlation coefficients *ρ* >0.53 based on pico-size NCLDV communities are drawn as edges. Known virus-host associations are shown in red, whereas unknown associations are shown in gray.

Figure 6 **Vertical linkage of NCLDV communities between the surface and mesopelagic layers.** (A) Latitudinal trend in NCLDV community similarity between two depths (with the station numbers). Relationship between NCLDV vertical similarity and (B) the surface chlorophyll *a* biomass, (C) NCLDV richness in the mesopelagic layer, (D) sampling depth of mesopelagic seawater, (E) the mixed layer depth and (F) temperature difference between epipelagic and mesopelagic samples. All NCLDV data were generated based on the pico-size fraction. Shaded areas represent 90% confidence intervals.

References

1 Field, C. B., Behrenfeld, M. J., Randerson, J. T. & Falkowski, P. Primary production of the biosphere: integrating terrestrial and oceanic components. *Science* **281**, 237-240, doi:10.1126/science.281.5374.237 (1998).

2 Worden, A. Z. *et al.* Environmental science. Rethinking the marine carbon cycle: factoring in the multifarious lifestyles of microbes. *Science* **347**, 1257594, doi:10.1126/science.1257594 (2015).

3 Brum, J. R. & Sullivan, M. B. Rising to the challenge: accelerated pace of discovery transforms marine virology. *Nat Rev Microbiol* **13**, 147-159, doi:10.1038/nrmicro3404 (2015).

4 Selosse, M.-A., Charpin, M. & Not, F. Mixotrophy everywhere on land and in water: the grand écart hypothesis. *Ecology Letters* **20**, 246-263, doi:10.1111/ele.12714 (2017).

5 Weitz, J. S. *et al.* A multitrophic model to quantify the effects of marine viruses on microbial food webs and ecosystem processes. *Isme j* **9**, 1352-1364, doi:10.1038/ismej.2014.220 (2015).

6 Mojica, K. D., Huisman, J., Wilhelm, S. W. & Brussaard, C. P. Latitudinal variation in virus-induced mortality of phytoplankton across the North Atlantic Ocean. *Isme j* **10**, 500-513, doi:10.1038/ismej.2015.130 (2016).

7 Suttle, C. A. Marine viruses--major players in the global ecosystem. *Nat Rev Microbiol* **5**, 801-812, doi:10.1038/nrmicro1750 (2007).

8 Guidi, L. *et al.* Plankton networks driving carbon export in the oligotrophic ocean. *Nature* **532**, 465-470, doi:10.1038/nature16942 (2016).

9 Laber, C. P. *et al.* Coccolithovirus facilitation of carbon export in the North Atlantic. *Nat Microbiol* **3**, 537-547, doi:10.1038/s41564-018-0128-4 (2018).

10 Colson, P. *et al.* "Megavirales", a proposed new order for eukaryotic nucleocytoplasmic large DNA viruses. *Arch Virol* **158**, 2517-2521, doi:10.1007/s00705-013-1768-6 (2013).

11 Fischer, M. G. Giant viruses come of age. *Curr Opin Microbiol* **31**, 50-57, doi:10.1016/j.mib.2016.03.001 (2016).

12 Koonin, E. V. & Yutin, N. Evolution of the Large Nucleocytoplasmic DNA Viruses of Eukaryotes and Convergent Origins of Viral Gigantism. *Adv Virus Res* **103**, 167-202, doi:10.1016/bs.aivir.2018.09.002 (2019).

13 Monier, A., Claverie, J. M. & Ogata, H. Taxonomic distribution of large DNA viruses in the sea. *Genome Biol* **9**, R106, doi:10.1186/gb-2008-9-7-r106 (2008).

14 Hingamp, P. *et al.* Exploring nucleo-cytoplasmic large DNA viruses in Tara Oceans microbial metagenomes. *ISME J* **7**, 1678-1695, doi:10.1038/ismej.2013.59 (2013).

15 Clerissi, C. *et al.* Deep sequencing of amplified Prasinovirus and host green algal genes from an Indian Ocean transect reveals interacting trophic dependencies and new genotypes. *Environ Microbiol Rep* **7**, 979-989, doi:10.1111/1758-2229.12345 (2015).

16 Li, Y. *et al.* The Earth Is Small for "Leviathans": Long Distance Dispersal of Giant Viruses across Aquatic Environments. *Microbes Environ* **34**, 334-339, doi:10.1264/jsme2.ME19037 (2019).

17 Mihara, T. *et al.* Taxon Richness of "Megaviridae" Exceeds those of Bacteria and Archaea in the Ocean. *Microbes Environ* **33**, 162-171, doi:10.1264/jsme2.ME17203 (2018).

18 Li, Y. *et al.* Degenerate PCR Primers to Reveal the Diversity of Giant Viruses in Coastal Waters. *Viruses* **10**, 496, doi:10.3390/v10090496 (2018).

19 Schulz, F. *et al.* Giant virus diversity and host interactions through global metagenomics. *Nature*, doi:10.1038/s41586-020-1957-x (2020).

20 Moniruzzaman, M., Martinez-Gutierrez, C. A., Weinheimer, A. R. & Aylward, F. O. Dynamic genome evolution and complex virocell metabolism of globally-distributed giant viruses. *Nat Commun* **11**, 1710, doi:10.1038/s41467-020-15507-2 (2020).

21 Cottrell, M. T. & Suttle, C. A. Wide-spread occurrence and clonal variation in viruses which cause lysis of a cosmopolitan, eukaryotic marine phytoplankter, Micromonas pusilla. *Mar Ecol Prog Ser* **78** (1991).

22 Bratbak, G., Egge, J. K. & Heldal, M. Viral mortality of the marine alga Emiliania huxleyi (Haptophyceae) and termination of algal blooms. *Marine Ecology Progress Series* **93**, 39-48 (1993).

23 Kenji, T., Keizo, N., Shigeru, I. & Mineo, Y. Isolation of a virus infecting the novel shellfish-killing dinoflagellate Heterocapsa circularisquama. *Aquatic Microbial Ecology* **23**, 103-111 (2001).

24 Fischer, M. G., Allen, M. J., Wilson, W. H. & Suttle, C. A. Giant virus with a remarkable complement of genes infects marine zooplankton. *Proc Natl Acad Sci U S A* **107**, 19508-19513, doi:10.1073/pnas.1007615107 (2010).

25 Needham, D. M. *et al.* A distinct lineage of giant viruses brings a rhodopsin photosystem to unicellular marine predators. *Proc Natl Acad Sci U S A* **116**, 20574-20583, doi:10.1073/pnas.1907517116 (2019).

26 Pesant, S. *et al.* Open science resources for the discovery and analysis of Tara Oceans data. *Sci Data* **2**, 150023, doi:10.1038/sdata.2015.23 (2015).

27 Gregory, A. C. *et al.* Marine DNA Viral Macro- and Microdiversity from Pole to Pole. *Cell* **177**, 1109-1123 e1114, doi:10.1016/j.cell.2019.03.040 (2019).

28 Salazar, G. *et al.* Gene Expression Changes and Community Turnover Differentially Shape the Global Ocean Metatranscriptome. *Cell* **179**, 1068-1083 e1021, doi:10.1016/j.cell.2019.10.014 (2019).

29 Matsen, F. A., Kodner, R. B. & Armbrust, E. V. pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics* **11**, 538, doi:10.1186/1471-2105-11-538 (2010).

30 Gallot-Lavallee, L., Blanc, G. & Claverie, J. M. Comparative Genomics of Chrysochromulina Ericina Virus and Other Microalga-Infecting Large DNA

Viruses Highlights Their Intricate Evolutionary Relationship with the Established Mimiviridae Family. *J Virol* **91**, doi:10.1128/jvi.00230-17 (2017).

31 Ibarbalz, F. M. *et al.* Global Trends in Marine Plankton Diversity across Kingdoms of Life. *Cell* **179**, 1084-1097 e1021, doi:10.1016/j.cell.2019.10.008 (2019).

32 Mihara, T. *et al.* Linking Virus Genomes with Host Taxonomy. *Viruses* **8**, 66, doi:10.3390/v8030066 (2016).

33 Ogata, H. *et al.* Remarkable sequence similarity between the dinoflagellate-infecting marine girus and the terrestrial pathogen African swine fever virus. *Virol J* **6**, 178, doi:10.1186/1743-422X-6-178 (2009).

34 Andreani, J. *et al.* Pacmanvirus, a New Giant Icosahedral Virus at the Crossroads between Asfarviridae and Faustoviruses. *J Virol* **91**, doi:10.1128/JVI.00212-17 (2017).

35 Barton, A. D., Dutkiewicz, S., Flierl, G., Bragg, J. & Follows, M. J. Patterns of diversity in marine phytoplankton. *Science* **327**, 1509-1511, doi:10.1126/science.1184961 (2010).

36 Lima-Mendez, G. *et al.* Ocean plankton. Determinants of community structure in the global plankton interactome. *Science* **348**, 1262073, doi:10.1126/science.1262073 (2015).

37 Zhou, J. & Ning, D. Stochastic Community Assembly: Does It Matter in Microbial Ecology? *Microbiol Mol Biol Rev* **81**, doi:10.1128/mmbr.00002-17 (2017).

38 Chow, C. E. & Suttle, C. A. Biogeography of Viruses in the Sea. *Annu Rev Virol* **2**, 41-66, doi:10.1146/annurev-virology-031413-085540 (2015).

39 Yoshida, T. *et al.* Locality and diel cycling of viral production revealed by a 24 h time course cross-omics analysis in a coastal region of Japan. *Isme j* **12**, 1287-1295, doi:10.1038/s41396-018-0052-x (2018).

40 Sunagawa, S. *et al.* Ocean plankton. Structure and function of the global ocean

microbiome. *Science* **348**, 1261359, doi:10.1126/science.1261359 (2015).

41 Syed, T. H., Famiglietti, J. S., Zlotnicki, V. & Rodell, M. Contemporary estimates of Pan-Arctic freshwater discharge from GRACE and reanalysis. *Geophysical Research Letters* **34**, doi:10.1029/2007gl031254 (2007).

42 Wommack, K. E. & Colwell, R. R. Virioplankton: viruses in aquatic ecosystems. *Microbiol Mol Biol Rev* **64**, 69-114, doi:10.1128/mmbr.64.1.69-114.2000 (2000).

43 Bellec, L. *et al.* Cophylogenetic interactions between marine viruses and eukaryotic picophytoplankton. *BMC Evol Biol* **14**, 59, doi:10.1186/1471-2148-14-59 (2014).

44 Brussaard, C. P. D., Kempers, R. S., Kop, A. J., Riegman, R. & Heldal, M. Virus-like particles in a summer bloom of Emiliania huxleyi in the North Sea. *Aquatic Microbial Ecology* **10**, 105-113 (1996).

45 Stephan, J. *et al.* Flow cytometric analysis of an Emiliana huxleyi bloom terminated by viral infection. *Aquatic Microbial Ecology* **27**, 111-124 (2002).

46 Hurwitz, B. L., Westveld, A. H., Brum, J. R. & Sullivan, M. B. Modeling ecological drivers in marine viral communities using comparative metagenomics and network analyses. *Proc Natl Acad Sci U S A* **111**, 10714-10719, doi:10.1073/pnas.1319778111 (2014).

47 Herndl, G. J. & Reinthaler, T. Microbial control of the dark end of the biological pump. *Nat Geosci* **6**, 718-724, doi:10.1038/ngeo1921 (2013).

48 Giering, S. L. *et al.* Reconciliation of the carbon budget in the ocean's twilight zone. *Nature* **507**, 480-483, doi:10.1038/nature13123 (2014).

49 Boyd, P. W., Claustre, H., Levy, M., Siegel, D. A. & Weber, T. Multi-faceted particle pumps drive carbon sequestration in the ocean. *Nature* **568**, 327-335, doi:10.1038/s41586-019-1098-2 (2019).

50 Janice, E. L. & Curtis, A. S. Effect of viral infection on sinking rates of Heterosigma akashiwo and its implications for bloom termination. *Aquatic Microbial Ecology* **37**, 1-7 (2004).

51 Close, H. G. *et al.* Export of submicron particulate organic matter to mesopelagic depth in an oligotrophic gyre. *Proc Natl Acad Sci U S A* **110**, 12565-12570, doi:10.1073/pnas.1217514110 (2013).

52 Mestre, M. *et al.* Sinking particles promote vertical connectivity in the ocean microbiome. *Proc Natl Acad Sci U S A* **115**, E6799-E6807, doi:10.1073/pnas.1802470115 (2018).

53 Hurwitz, B. L., Brum, J. R. & Sullivan, M. B. Depth-stratified functional and taxonomic niche specialization in the 'core' and 'flexible' Pacific Ocean Virome. *Isme j* **9**, 472-484, doi:10.1038/ismej.2014.143 (2015).

54 Sancetta, C., Villareal, T. & Falkowski, P. Massive fluxes of rhizosolenid diatoms: A common occurrence? *Limnology and Oceanography* **36**, 1452-1457, doi:10.4319/lo.1991.36.7.1452 (1991).

55 Kawakami, H. & Honda, M. C. Time-series observation of POC fluxes estimated from 234Th in the northwestern North Pacific. *Deep Sea Research Part I: Oceanographic Research Papers* **54**, 1070-1090, doi:10.1016/j.dsr.2007.04.005 (2007).

56 Richardson, T. L. & Jackson, G. A. Small phytoplankton and carbon export from the surface ocean. *Science* **315**, 838-840, doi:10.1126/science.1133471 (2007).

57 Blanc-Mathieu, R. *et al.* Viruses of the eukaryotic plankton are predicted to increase carbon export efficiency in the global sunlit ocean. *bioRxiv*, 710228, doi:10.1101/710228 (2019).

58 Iversen, M. H. & Ploug, H. Ballast minerals and the sinking carbon flux in the ocean: carbon-specific respiration rates and sinking velocity of marine snow aggregates. *Biogeosciences* **7**, 2613-2624, doi:10.5194/bg-7-2613-2010 (2010).

59 Alberti, A. *et al.* Viral to metazoan marine plankton nucleotide sequences from the Tara Oceans expedition. *Sci Data* **4**, 170093, doi:10.1038/sdata.2017.93 (2017).

60 Eddy, S. R. Profile hidden Markov models. *Bioinformatics* **14**, 755-763,

doi:10.1093/bioinformatics/14.9.755 (1998).

61 Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* **30**, 772-780, doi:10.1093/molbev/mst010 (2013).

62 Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**, 2688-2690, doi:10.1093/bioinformatics/btl446 (2006).

63 Koonin, E. V. & Yutin, N. Multiple evolutionary origins of giant viruses. *F1000Res* **7**, doi:10.12688/f1000research.16248.1 (2018).

64 Yoshikawa, G. *et al.* Medusavirus, a Novel Large DNA Virus Discovered from Hot Spring Water. *J Virol* **93**, doi:10.1128/JVI.02130-18 (2019).

65 Longhurst, A. R. in *Ecological Geography of the Sea (Second Edition)* (ed Alan R. Longhurst) 89-102 (Academic Press, 2007).

66 Letunic, I. & Bork, P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res* **44**, W242-245, doi:10.1093/nar/gkw290 (2016).

67 Besemer, J., Lomsadze, A. & Borodovsky, M. GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res* **29**, 2607-2618, doi:10.1093/nar/29.12.2607 (2001).

68 Keeling, P. J. *et al.* The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. *PLoS Biol* **12**, e1001889, doi:10.1371/journal.pbio.1001889 (2014).

69 Jain, C., Rodriguez, R. L., Phillippy, A. M., Konstantinidis, K. T. & Aluru, S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat Commun* **9**, 5114, doi:10.1038/s41467-018-07641-9 (2018).

70 Anderson, M. J. A new method for non-parametric multivariate analysis of

variance. *Austral Ecology* **26**, 32-46, doi:10.1111/j.1442-9993.2001.01070.pp.x (2001).

71 de Vargas, C. *et al.* Ocean plankton. Eukaryotic plankton diversity in the sunlit ocean. *Science* **348**, 1261605, doi:10.1126/science.1261605 (2015).

72 Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)* **57**, 289-300, doi:10.1111/j.2517-6161.1995.tb02031.x (1995).

Acknowledgement

This work was supported by JSPS/KAKENHI (Nos. 26430184, 18H02279, and 19H05667 to H.O. and Nos. 19K15895 and 19H04263 to H.E.), Scientific Research on Innovative Areas from the Ministry of Education, Culture, Science, Sports and Technology (MEXT) of Japan (Nos. 16H06429, 16K21723, and 16H06437 to H.O.), Kyoto University Research Coordination Alliance (funding to H.E.), and the Collaborative Research Program of the Institute for Chemical Research, Kyoto University (Nos. 2019-30 and 2020-27). Computational time was provided by the SuperComputer System, Institute for Chemical Research, Kyoto University. We further thank the *Tara* Oceans consortium, the projects OCEANOMICS (ANR-11-BTBR-0008) and France Genomique (ANR-10-INBS-09), and the people and sponsors who supported *Tara* Oceans. *Tara* Oceans (that includes both the *Tara* Oceans and *Tara* Oceans Polar Circle expeditions) would not exist without the leadership of the *Tara* Expeditions Foundation and the continuous support of 23 institutes (https://oceans.taraexpeditions.org). This article is contribution number 108 of *Tara* Oceans.

Author contributions

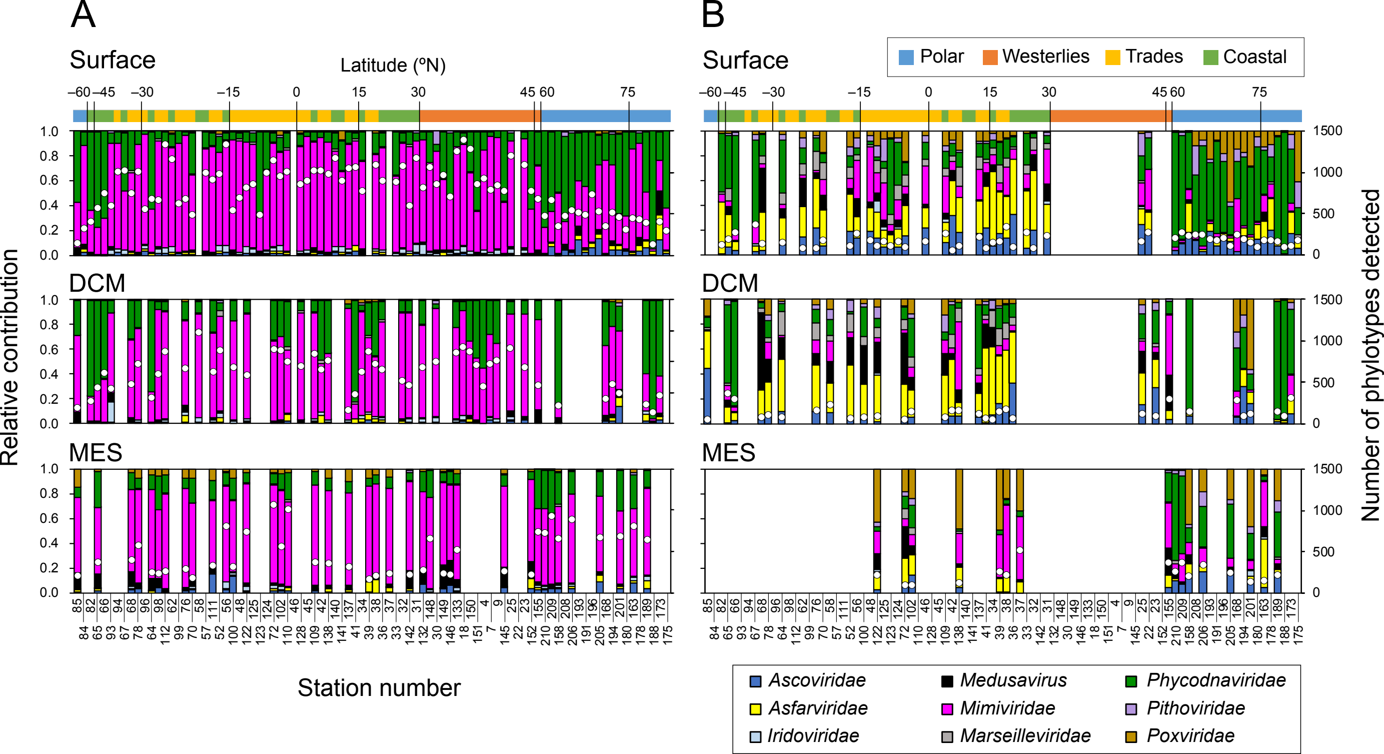
HE and HO designed the study. HE performed most of the bioinformatics analysis. RB-M and YL contributed to the bioinformatics analysis. GS, NH, KL, CdV, MBS, CB, PW, LK-B, and SS contributed to the generation of primary data. CdV, MBS, CB, PW, LK-B, SS, and HO coordinated *Tara* Oceans. All authors contributed to the writing of the manuscript.

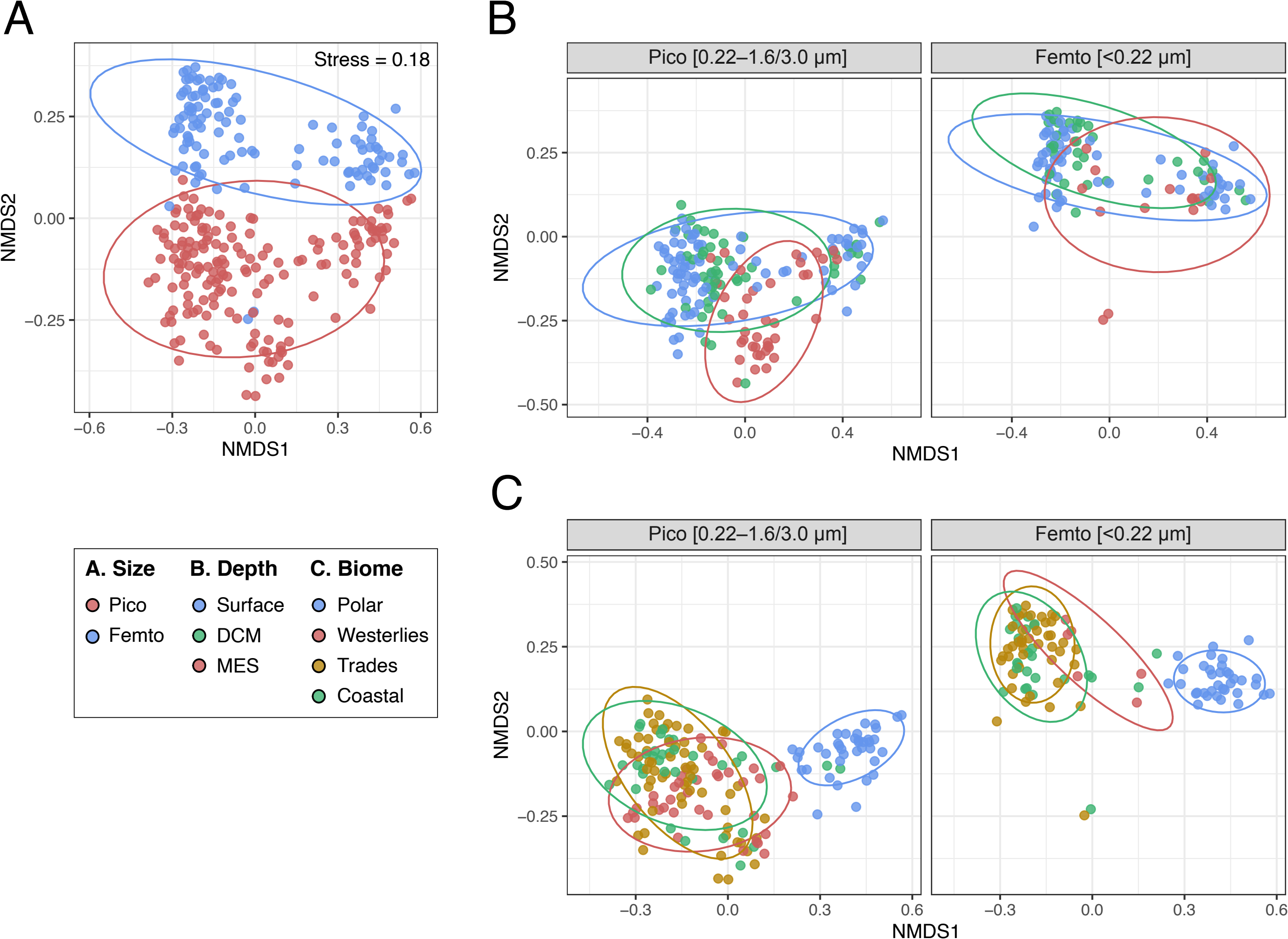
Materials & Correspondence

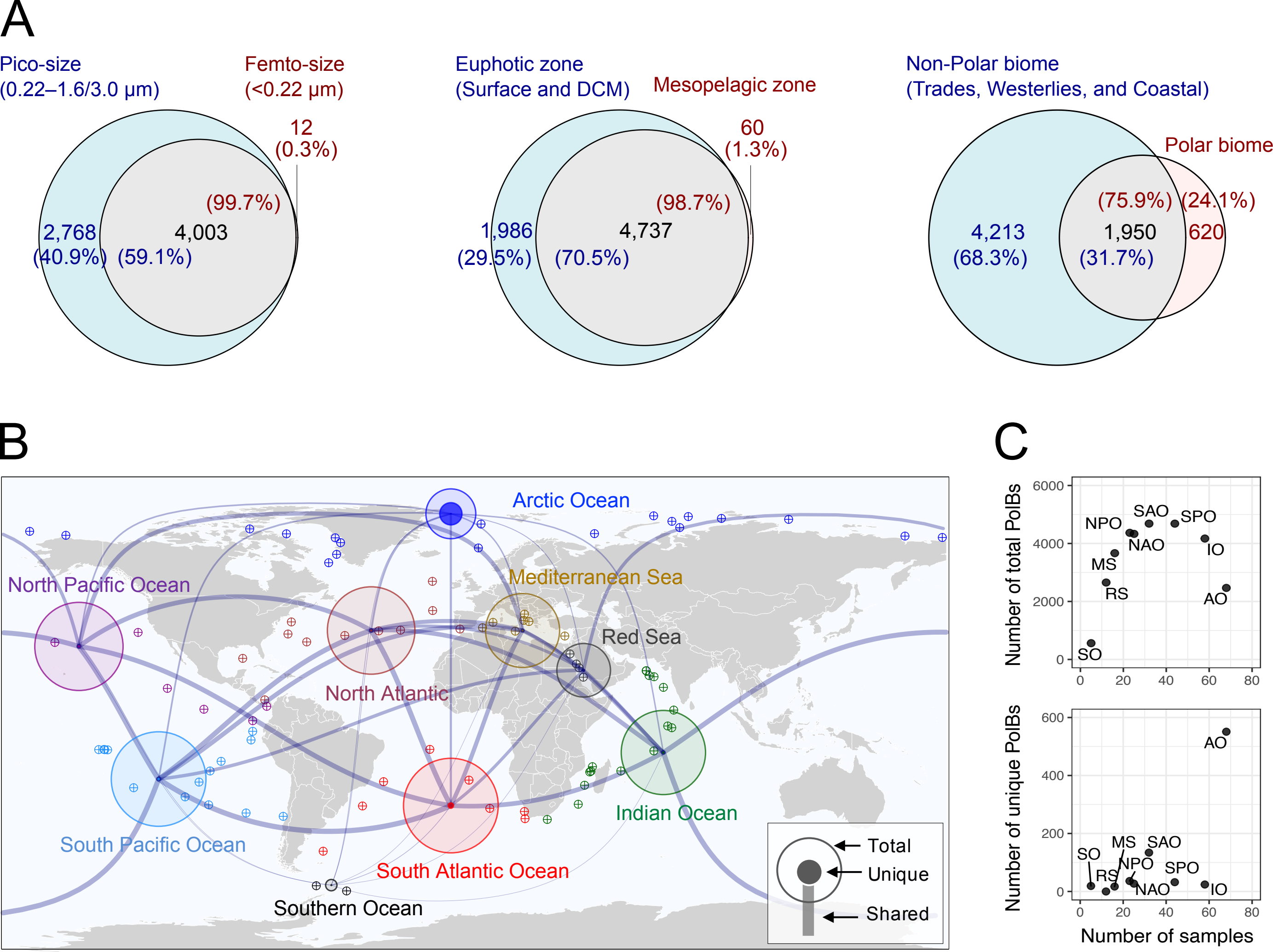
Correspondence and material requests should be addressed to HO (email: ogata@kuicr.kyoto-u.ac.jp).

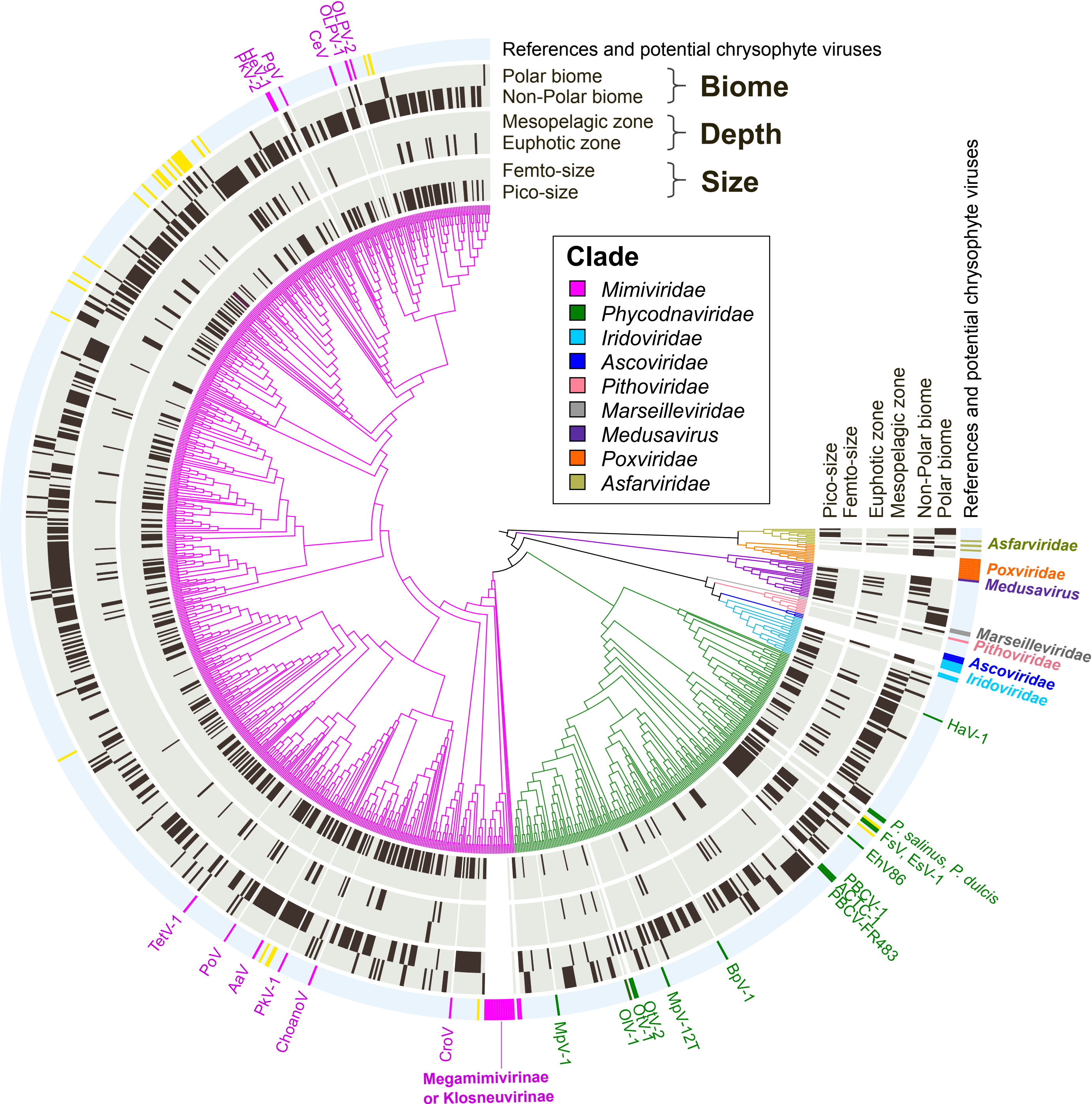
Competing financial interests

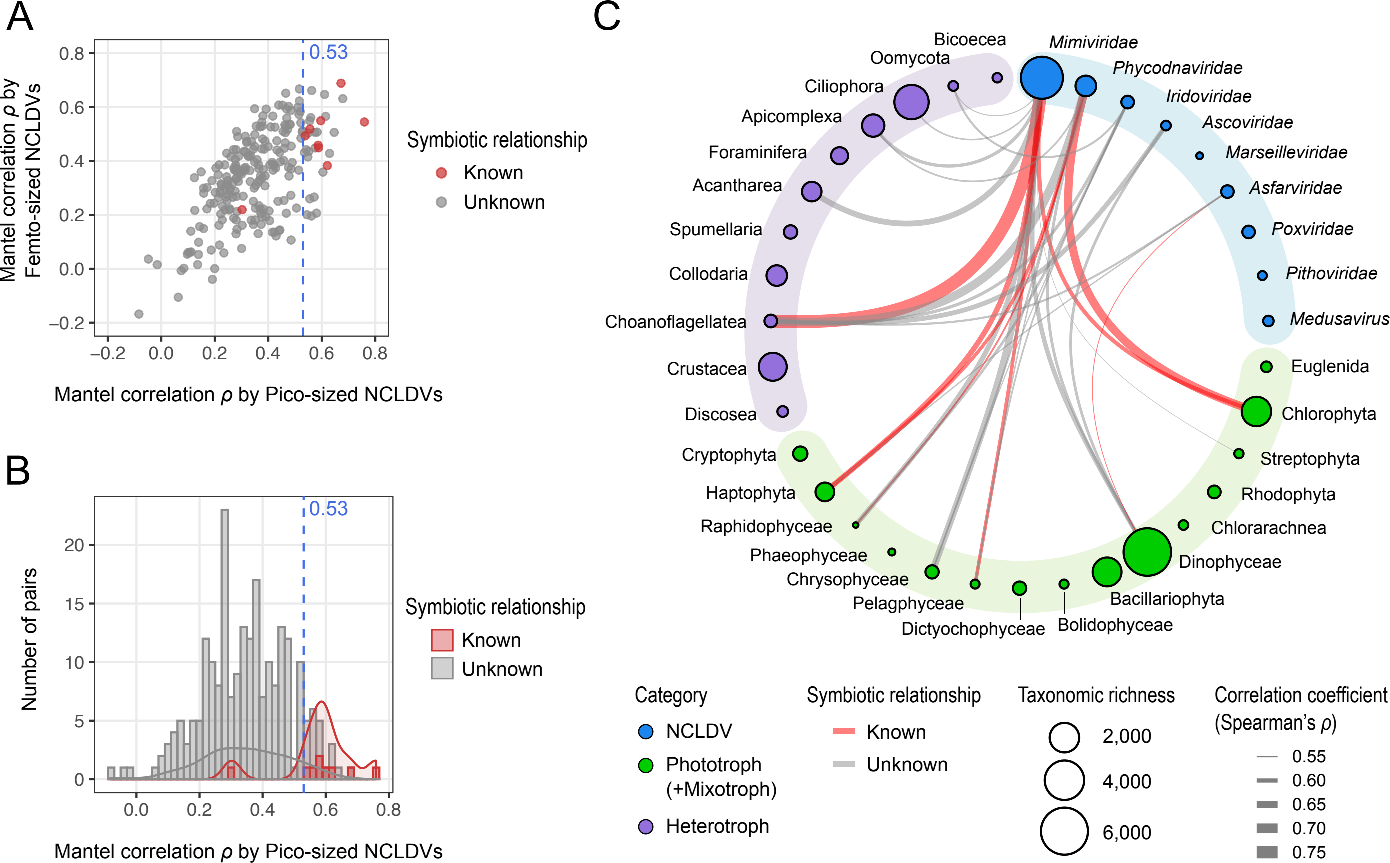
The authors declare no competing financial interests.

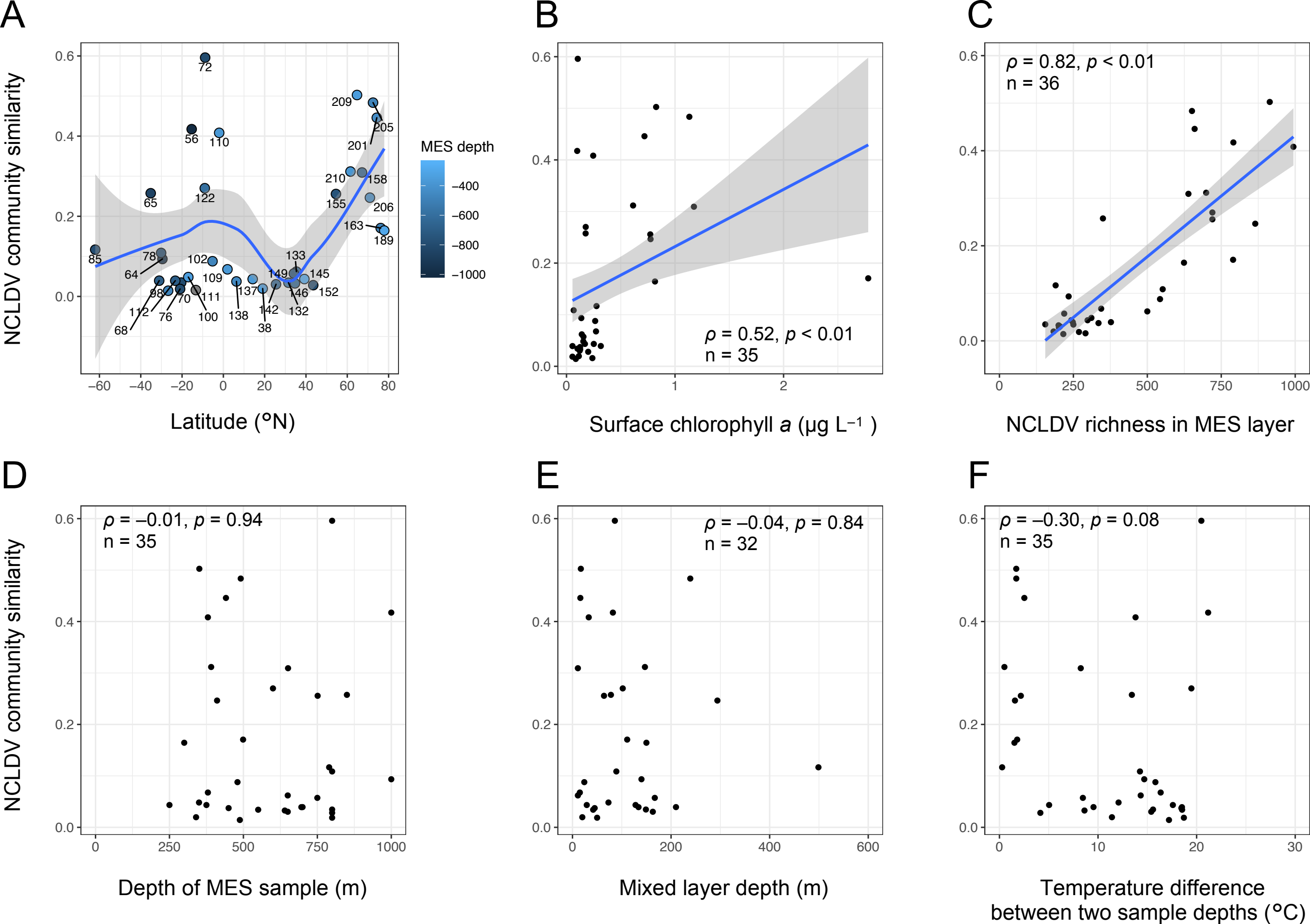

A
B
Surface
DCM
MES
Latitude (ºN)
–60 –45 –30 –15 0 15 30 45 60 75
Polar
Westerlies
Trades
Coastal
Relative contribution
0.0 0.2 0.4 0.6 0.8 1.0
Number of phylotypes detected
0 500 1000 1500
85 82 66 94 68 96 98 62 76 58 111 56 48 125 124 102 46 45 42 140 137 34 38 37 32 31 148 149 133 150 4 9 25 23 155 209 208 193 196 168 201 163 189 173
84 65 93 67 78 64 112 99 70 57 52 100 122 123 72 110 128 109 138 141 41 39 36 33 142 132 30 146 18 151 7 145 22 152 210 158 206 191 205 194 180 178 188 175
Station number
Ascoviridae
Asfarviridae
Iridoviridae
Medusavirus
Mimiviridae
Marseilleviridae
Phycodnaviridae
Pithoviridae
Poxviridae


A
Stress = 0.18
0.25
0.00
−0.25
NMDS2
−0.6
−0.3
0.0
0.3
0.6
NMDS1
B
Pico [0.22–1.6/3.0 µm]
Femto [<0.22 µm]
0.25
0.00
−0.25
−0.50
NMDS2
−0.4
0.0
0.4
−0.4
0.0
0.4
NMDS1
C
Pico [0.22–1.6/3.0 µm]
Femto [<0.22 µm]
0.50
0.25
0.00
−0.25
NMDS2
−0.3
0.0
0.3
0.6
−0.3
0.0
0.3
0.6
NMDS1
A. Size
● Pico
● Femto
B. Depth
● Surface
● DCM
● MES
C. Biome
● Polar
● Westerlies
● Trades
● Coastal


A
Pico-size
(0.22–1.6/3.0 µm)
Femto-size
(<0.22 µm)
12
(0.3%)
(99.7%)
2,768
(40.9%)
4,003
(59.1%)
Euphotic zone
(Surface and DCM)
Mesopelagic zone
60
(1.3%)
(98.7%)
1,986
(29.5%)
4,737
(70.5%)
Non-Polar biome
(Trades, Westerlies, and Coastal)
Polar biome
(75.9%)
(24.1%)
4,213
(68.3%)
1,950
(31.7%)
620
B
Arctic Ocean
North Pacific Ocean
Mediterranean Sea
Red Sea
North Atlantic
South Pacific Ocean
South Atlantic Ocean
Indian Ocean
Southern Ocean
Total
Unique
Shared
C
Number of total PolBs
6000
4000
2000
0
0
20
40
60
80
SAO
SPO
NPO
NAO
IO
MS
RS
AO
SO
Number of unique PolBs
600
400
200
0
0
20
40
60
80
AO
SO
MS
SAO
RS
NPO
SPO
NAO
IO
Number of samples


References and potential chrysophyte viruses
Polar biome
Non-Polar biome
Biome
Mesopelagic zone
Euphotic zone
Depth
Femto-size
Pico-size
Size
Clade
Mimiviridae
Phycodnaviridae
Iridoviridae
Ascoviridae
Pithoviridae
Marseilleviridae
Medusavirus
Poxviridae
Asfarviridae
Pico-size
Femto-size
Euphotic zone
Mesopelagic zone
Non-Polar biome
Polar biome
References and potential chrysophyte viruses
Asfarviridae
Poxviridae
Medusavirus
Marseilleviridae
Pithoviridae
Ascoviridae
Iridoviridae
HaV-1
P. salinus, P. dulcis
FsV, EsV-1
EhV86
PBCV-1
ACTV-1
PBCV-FR483
BpV-1
MpV-12T
OtV-2
OtV-1
OlV-1
MpV-1
Megamimivirinae
or Klosneuvirinae
CroV
ChoanoV
PkV-1
AaV
PoV
TetV-1
PgV
HeV-1
PkV-2
CeV
OLPV-1
OLPV-2


A
0.53
Mantel correlation ρ by Femto-sized NCLDVs
Mantel correlation ρ by Pico-sized NCLDVs
0.8
0.6
0.4
0.2
0.0
−0.2
Symbiotic relationship
Known
Unknown
B
0.53
Number of pairs
20
15
10
5
0
0.0
0.2
0.4
0.6
0.8
Mantel correlation ρ by Pico-sized NCLDVs
Symbiotic relationship
Known
Unknown
C
Bicoecea
Oomycota
Ciliophora
Apicomplexa
Foraminifera
Acantharea
Spumellaria
Collodaria
Choanoflagellatea
Crustacea
Discosea
Mimiviridae
Phycodnaviridae
Iridoviridae
Ascoviridae
Marseilleviridae
Asfarviridae
Poxviridae
Pithoviridae
Medusavirus
Euglenida
Chlorophyta
Streptophyta
Rhodophyta
Chlorarachnea
Dinophyceae
Bacillariophyta
Bolidophyceae
Dictyochophyceae
Pelagphyceae
Chrysophyceae
Phaeophyceae
Raphidophyceae
Haptophyta
Cryptophyta
Category
NCLDV
Phototroph (+Mixotroph)
Heterotroph
Symbiotic relationship
Known
Unknown
Taxonomic richness
2,000
4,000
6,000
Correlation coefficient (Spearman's ρ)
0.55
0.60
0.65
0.70
0.75


A
B
C
D
E
F
NCLDV community similarity
NCLDV community similarity
0.6
0.4
0.2
0.0
72
209
205
201
56
110
210
158
65
122
155
206
163
189
85
78
64
102
133
149
145
109
146
152
98
70
111
137
142
132
112
76
100
138
38
68
−60
−40
−20
0
20
40
60
80
Latitude (°N)
MES depth
−400
−600
−800
−1000
ρ = 0.52, p < 0.01
n = 35
Surface chlorophyll a (μg L−1)
ρ = 0.82, p < 0.01
n = 36
NCLDV richness in MES layer
ρ = −0.01, p = 0.94
n = 35
Depth of MES sample (m)
ρ = −0.04, p = 0.84
n = 32
Mixed layer depth (m)
ρ = −0.30, p = 0.08
n = 35
Temperature difference
between two sample depths (°C)
0
1
2
250
500
750
1000
200
400
600
10
20
30