



TITLE:

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

AUTHOR(S):

Endo, Hisashi; Blanc-Mathieu, Romain; Li, Yanze; Salazar, Guillem; Henry, Nicolas; Labadie, Karine; de Vargas, Colomban; ... Karp-Boss, Lee; Sunagawa, Shinichi; Ogata, Hiroyuki

CITATION:

Endo, Hisashi ...[et al]. Biogeography of marine giant viruses reveals their interplay with eukaryotes and ecological functions. *Nature Ecology & Evolution* 2020, 4(12): 1639-1649

ISSUE DATE:

2020-12

URL:

<http://hdl.handle.net/2433/259797>

RIGHT:

This is a post-peer-review, pre-copyedit version of an article published in 'Nature Ecology & Evolution'. The final authenticated version is available online at: <https://doi.org/10.1038/s41559-020-01288-w>.; The full-text file will be made open to the public on 07 September 2021 in accordance with publisher's 'Terms and Conditions for Self-Archiving'; この論文は出版社版ではありません。引用の際には出版社版をご確認ご利用ください。; This is not the published version. Please cite only the published version.

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

3

4 Hisashi Endo¹, Romain Blanc-Mathieu^{1,2}, Yanze Li¹, Guillem Salazar³, Nicolas Henry^{4,5},
5 Karine Labadie⁶, Colomban de Vargas^{4,5}, Matthew B. Sullivan^{7,8}, Chris Bowler^{9,10},
6 Patrick Wincker^{10,11}, Lee Karp-Boss¹², Shinichi Sunagawa³, Hiroyuki Ogata^{1,*}

7

8 **Affiliations:**

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

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

33

34 **Corresponding author:*

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

36

37 **Abstract**

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

52

53

54 Introduction

55 The photic zone is the most productive layer of the ocean, containing a wide variety
56 of microorganisms such as bacteria, autotrophic and heterotrophic protists and
57 multicellular organisms. The population dynamics of these organisms determine the
58 flows of energy and materials through marine food webs, playing a fundamental role in
59 ecosystem functioning and biogeochemical cycles in the ocean^{1,2}. Viruses exert a top-
60 down control on marine organisms and release material to the pools of particulate and
61 dissolved organic matter³. This material and remineralized inorganic nutrients are utilized
62 by autotrophic and mixotrophic phytoplankton⁴. The recycling of nutrients in the surface
63 layer potentially reduces the transfer of fixed organic carbon to higher trophic levels and
64 the deep sea^{5,6}. However, it is also possible that viruses enhance downward carbon flux
65 by facilitating cell aggregation and producing carbon-enriched materials from infected
66 cells⁷⁻⁹.

67 Nucleocytoplasmic large DNA viruses (NCLDVs or so-called “giant viruses”)
68 represent a monophyletic group of viruses that infect a variety of eukaryotic lineages¹⁰⁻¹².
69 Studies focusing on conserved marker genes such as family B DNA polymerase (*polB*)
70 have revealed that NCLDVs are highly diverse and abundant in aquatic environments<sup>13-
71 16</sup>. The diversity of a family of NCLDVs, namely *Mimiviridae*, exceeds that of bacteria
72 and archaea in the ocean¹⁷ and their richness in a few liters of seawater can reach more
73 than 5,000 operational taxonomic units¹⁸. More recently, several thousand draft genomes
74 (i.e., metagenome-assembled genomes; MAGs) of NCLDVs were constructed from
75 environmental sequences, thanks to the development of high-throughput sequencing and
76 bioinformatics technologies^{19,20}. However, the global biogeography of marine NCLDVs
77 still remains under-explored.

78 A growing number of marine eukaryotes have been reported as host organisms of
79 NCLDVs, particularly phytoplankton groups such as haptophytes, chlorophytes and
80 dinoflagellates²¹⁻²³. Other eukaryotic lineages, including non-photosynthetic organisms
81 such as bicosoecids and choanoflagellates, have also been reported as host organisms of

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

86 Here we reveal patterns in the global biogeography of NCLDVs using the
87 metagenomic data from the *Tara* Oceans project. The metagenomic data cover varying
88 geographic regions including polar and deep-sea ecosystems, in which NCLDVs are
89 under-researched²⁶⁻²⁸. We constructed NCLDV taxonomic abundance profiles for 283
90 samples, representing two viral size fractions, three ocean depth ranges (surface, deep
91 chlorophyll maximum and mesopelagic), and four biomes (coastal, trades, westerlies and
92 polar). The global biogeography of NCLDVs derived from these data reveals strong
93 associations between NCLDVs and eukaryotic microorganisms. Furthermore, vertical
94 connectivity of NCLDV communities indicates a possible mechanism for how
95 mesopelagic NCLDV communities are structured with respect to ocean biogeochemical
96 processes.

97

98 **Results**

99 **NCLDV phylotypes detected in *Tara* Oceans metagenomes**

100 We detected 6,818 PolBs affiliated with NCLDVs in the second version of the Ocean
101 Microbial Reference Gene Catalog (OM-RGC.v2)²⁸ using the pplacer phylogenetic
102 placement method²⁹ (see methods for details). The OM-RGC.v2 was built based on 370
103 *Tara* Oceans metagenomes from femto- (<0.2 μm ; 151 samples), pico- (0.22–1.6 or 0.22–
104 3.0 μm ; 180 samples) and other (39 samples) size fractions. After removing 32 samples
105 with a low NCLDV frequency and 55 samples from non-target size fractions and depths,
106 the remaining 283 samples contained 6,783 NCLDV PolB sequences. The pplacer
107 classified these PolBs into nine NCLDV families/lineages. The number of phylotypes
108 (distinct *polB* at 95% nucleotide sequence identity) was the largest in *Mimiviridae* (5,091
109 phylotypes), followed by *Phycodnaviridae* (981 phylotypes). The number of phylotypes

110 taxonomically assigned to *Iridoviridae*, *Medusavirus* and *Asfarviridae*, were 239, 120
 111 and 109, respectively. We also detected PolBs assigned to *Pithoviridae* (93), *Ascoviridae*
 112 (78), *Poxviridae* (51) and *Marseilleviridae* (21). However, *Poxviridae* was omitted from
 113 our discussion as the environmental gene sequences were distantly related to known
 114 *Poxviridae*. Rarefaction analysis showed that, at the end of sampling, the number of
 115 NCLDV phylotypes increased by less than 0.01% per sample for all samples, and ranged
 116 from 0.02% to 0.32% when samples were divided into different size fractions, depths and
 117 biomes (Extended Data Fig. 1).

118 To examine detailed phylogenetic affiliation and to visualize the dispersal
 119 characteristics of each NCLDV phylotypes detected by pplacer, we constructed a
 120 phylogenetic tree using selected PolB sequences (Extended Data Figs. 2–4). Among the
 121 *Mimiviridae* family, genes closely related to the algal-infecting subfamily, recently
 122 proposed as “Mesomimivirinae” (e.g., AaV, CeV, pkV, PgV, PoV and TetV)³⁰, which
 123 infect pelagophytes (the genus *Aureococcus*), haptophytes (the genera *Haptolina*,
 124 *Prymnesium* and *Phaeocystis*), and chlorophytes (the genera *Pyramimonas* and
 125 *Tetraselmis*), were relatively abundant. On the other hand, only a few sequences were
 126 affiliated with the subfamilies “Megamimivirinae” and “Klosneuvirinae” except the
 127 *Cafeteria roenbergensis virus* (CroV), which is the only member of “Megamimivirinae”
 128 isolated from the marine environment²⁴. Among *Phycodnaviridae*, the genus
 129 *Prasinovirus* (e.g., BpV, MpV, OtV and OIV), which infect chlorophyte genera such as
 130 *Bathycoccus*, *Micromonas* and *Ostreococcus*, showed the highest richness.

131

132 **Heterogeneity in NCLDV community structure across size, depth and biomes**

133 The dominant NCLDV taxa detected from all sample locations and depths in the pico-
 134 size fraction were *Mimiviridae* and *Phycodnaviridae*, with average contributions of
 135 64.6% and 25.4%, respectively (Fig. 1A). The dominant groups of NCLDVs varied
 136 widely among sites and depths in samples from the femto-size fraction (Fig. 1B). In this
 137 fraction, *Phycodnaviridae* and *Asfarviridae* had relatively high contributions to the total

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

141 A non-metric multidimensional scaling (NMDS) analysis showed that NCLDV
 142 assemblages clustered according to size fraction, depth and biome (Fig. 2A–2C).
 143 Significant differences in NCLDV community composition were detected among all
 144 categories (PERMANOVA, $p < 0.01$), and size fraction, depth and biome explained 5.5%,
 145 4.3% and 10.9% of the total variance, respectively.

146 Taxonomic richness (i.e., number of phylotypes) and Shannon's diversity index were
 147 used to investigate variation in NCLDV community diversity. In this study, we analyzed
 148 the samples from all depths and size fractions to compare diversity differences among
 149 depth ranges, although latitudinal trend in Shannon's diversity for pico-sized
 150 communities from the surface was reported previously³¹. In the pico-size fraction, mean
 151 values for NCLDV richness at the surface and in the DCM layer were about 1.7 times
 152 higher than that in the mesopelagic layer (Kruskal-Wallis and Dunn's post hoc test, p
 153 < 0.01) (Extended Data Fig. 5A). In the femto-size fraction, NCLDV richness was
 154 significantly higher at the surface and MES layer than in the DCM layer (Dunn's test, p
 155 $= 0.04–0.05$), although the differences were small and not consistent with the pico-size
 156 fraction.

157

158 **High uniqueness of NCLDV phylotypes in the Arctic Ocean**

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

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

173 To further characterize regional differences in the NCLDV community, we
174 investigated the total and unique NCLDV phylotypes observed in nine geographic regions
175 and the phylotypes shared among regions. The total number of phylotypes was relatively
176 high in the Atlantic, Pacific and Indian Oceans and in the Mediterranean Sea, with values
177 of between 3,665 and 4,685 (Fig. 3B). Lower numbers of NCLDV phylotypes were
178 identified from the Red Sea (2,653) and the Arctic Ocean (2,467). The Southern Ocean
179 presented the lowest number of NCLDV phylotypes (561), although this was based on
180 only 5 samples. The Arctic Ocean samples displayed a high number of unique NCLDV
181 phylotypes (551), which corresponded to 22.3% of the total phylotypes detected in this
182 region. In contrast, the number of unique phylotypes from other regions ranged from 0 to
183 134 (0.0% to 3.4%).

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

190

191 **NCLDV distributions correlate with eukaryotic communities**

192 A partial Mantel test was conducted to assess community associations among the
193 NCLDV families/lineages and major eukaryotic lineages. The pairwise partial correlation

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

211

212 **Potential chrysophyte viruses constitute novel clades of *Mimiviridae***

213 To explore possible associations between NCLDVs and chrysophytes as indicated by
 214 the Mantel's regression analysis (Fig. 5C), we tested for chrysophyte-derived genes in
 215 the metagenome-assembled genomes (MAGs) of NCLDVs generated by Schultz et al.
 216 (2020)¹⁹ and Moniruzzaman et al. (2020)²⁰. The results showed that 89 (82 after removing
 217 redundancy) out of 2,263 MAGs contained genes closely related to the transcripts of the
 218 chrysophytes (Supplementary Data 1). Comparisons between PolB sequences revealed
 219 27 PolBs from the OM-RGC.v2 that were closely related to the NCLDV MAGs with
 220 chrysophyte homologs. Most of these PolBs constituted novel clades within the branches
 221 of *Mimiviridae* (Fig. 4; Extended Data Fig. 4). We confirmed that other genes in the

222 contigs that contained chrysophyte homologs are highly similar to the *Mimiviridae* or
 223 *Phycodnaviridae* sequences in many cases (Extended Data Fig. 6).

224

225 **Vertical connectivity of NCLDV communities**

226 The vertical connectivity of NCLDV communities was investigated using Bray-Curtis
 227 community similarity measures to compare between epipelagic (surface or DCM) and
 228 mesopelagic samples at individual sampling locations. The Bray-Curtis similarities were
 229 less than 0.10 for about half of the tested locations (20 out of 36 surface sites and 13 out
 230 of 26 DCM sites; Fig. 6A; Extended Data Fig. 7A). All sites in the Arctic Ocean and
 231 several sites in tropical and subtropical regions showed relatively high similarities
 232 between the two depth (0.15 to 0.60). The NCLDV community similarity value was
 233 positively correlated with the chlorophyll *a* concentration in the epipelagic layer
 234 (Spearman's $\rho = 0.52$, $p < 0.01$, asymptotic *t* approximation, $n = 36$ for surface; $\rho = 0.44$,
 235 $p = 0.02$, $n = 25$ for DCM) and NCLDV richness in the mesopelagic layer ($\rho = 0.82$, p
 236 < 0.01 , $n = 36$ for surface; $\rho = 0.70$, $p < 0.01$, $n = 26$ for DCM) (Figs. 6B and 6C; Extended
 237 Data Figs. 7B and 7C). We also evaluated relationships between NCLDV vertical
 238 similarity and physical environmental factors including: the sampling depth of
 239 mesopelagic water, the mixed layer depth, and the temperature difference between
 240 epipelagic and mesopelagic waters. No significant correlations were detected among
 241 these parameters ($p > 0.05$, $n = 32$ – 36 for surface samples and $n = 25$ – 26 for DCM
 242 samples) (Figs. 6D–F; Extended Data Figs. 7D–F).

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

249

250 Discussion

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

273 In addition to *Phycodnaviridae* and *Mimiviridae*, *Asfarviridae* also contribute an
 274 important proportion of NCLDVs in the femto-size fraction of most euphotic zones (Fig.
 275 1B). Although very limited information is available regarding the natural hosts for this
 276 group, a representative *Asfarviridae*-like species in marine environments is *Heterocapsa*
 277 *circularisquama* DNA virus (HcDNAV), which infects the red-tide-forming

278 dinoflagellate *H. circularisquama*³³. In the terrestrial ecosystem, this viral family is
279 known to infect a wide variety of organisms such as amoebzoa, arthropods and
280 mammals^{32,34}. Given the broad range of host species for this viral lineage, there may be
281 an unknown but wide-spread host taxa for *Asfarviridae* in the ocean.

282 Our study revealed a heterogeneous pattern in the distribution of NCLDV across the
283 oceans of the world (Fig. 2C). Although there are limited studies available on the factors
284 controlling the large-scale distribution of viruses, it is widely accepted that both
285 deterministic (environmental factors and inter-specific interactions) and stochastic
286 processes (e.g., immigration and speciation) are important in making up microbial
287 assemblages³⁵⁻³⁷. The distribution and diversity of viruses would not be directly affected
288 by environmental variables such as temperature and nutrient availability, but is directly
289 influenced by the geographic ranges of their host species^{3,38}. Recent work with
290 cyanophages demonstrated that a significant number of free-living viruses are locally
291 produced through active infection rather than from migration³⁹. Therefore, we expect that
292 viral community structure will reflect host distribution as well as infectious activity.

293 Despite significant differences in community composition across oceanic biomes, we
294 found that most NCLDV phlotypes are dispersed throughout tropical and temperate
295 regions (Figs. 3A and 3B), presumably following their host community composition,
296 which is primarily determined by temperature⁴⁰. However, the polar biome (mainly the
297 Arctic Ocean) constitutes a “hotspot” of unique NCLDV phlotypes from a wide variety
298 of families, despite having a low total richness in comparison to other regions (Figs. 3B
299 and 3C). We revealed that NCLDVs unique to non-polar biome were also abundant (Fig.
300 4), indicating a strong separation of NCLDV communities between polar and non-polar
301 biomes. A geographical barrier and steep environmental gradients may underlie this
302 distinct ecosystem structure (i.e., different host communities and their productivity) in the
303 Arctic Ocean^{27,28,31}. Moreover, the Arctic Ocean is characterized by high amounts of river
304 discharge, contributing more than 10% to global runoff flux⁴¹. Consequently, biological
305 processes in the Arctic may be influenced by river inputs from terrestrial ecosystems.

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

311 Tight coupling between NCLDVs and their hosts was further corroborated by our
312 partial Mantel statistics, which described both known virus-host interactions and
313 additional but currently unrecognized associations between viruses and eukaryotic
314 lineages at the community level. Using the pico-sized NCLDV community, we detected
315 almost all known virus-host interactions, except for those involving Bicoecia (Fig. 5C).
316 This demonstrates that distance-based correlation analysis using global ocean samples is
317 useful for detecting virus-host interplay in natural environments, although the validations
318 of the previously unknown associations remain to be further explored. Strong positive
319 relationships between NCLDVs and eukaryotes involved many phytoplankton lineages
320 including haptophytes, chlorophytes, dinophytes, pelagophytes and raphidophytes, all of
321 which include known host lineages of NCLDVs (Fig. 5C). Strong correlations were also
322 detected with heterotrophic choanoflagellates, which have recently been identified as a
323 novel host of *Mimiviridae*²⁵. Some NCLDVs, especially *Mimiviridae*, had strong
324 correlations with chrysophytes, although no host species have yet been reported for this
325 lineage. Many environmental NCLDV genomes were found to encode genes that are
326 likely to be derived from marine chrysophytes (Supplementary Data 1–3). Taxonomic
327 analyses based on PolB phylogeny and homology search revealed that most of these
328 phylotypes represent previously unknown clades of the *Mimiviridae* tree (Extended Data
329 4 and 6; Supplementary Data 4), suggesting that chrysophytes may be an important host
330 lineage of *Mimiviridae* in the ocean.

331 The global distribution of NCLDVs are determined by the geographic ranges of their
332 host organisms. Therefore, the virus-eukaryote associations that we detected likely arose
333 under these constraints. On the other hand, it is expected that NCLDVs influence the

334 abundance of eukaryotes at a local scale. Previous studies show that bacterial viruses have
 335 an important role in determining bacterial mortality, because they substantially
 336 outnumber their hosts and have highly specific infection mechanisms⁴². Similarly,
 337 NCLDV are reported to be more abundant than their host cells and have high infection
 338 specificity^{11,14,43}. For example, *Emiliania huxleyi* viruses (EhVs) of the *Phycodnaviridae*
 339 family are responsible for almost all of the mortality of the haptophyte *E. huxleyi* during
 340 blooms^{22,44,45}. Another field study suggests that viral lysis can explain a greater proportion
 341 of phytoplankton mortality than grazing by zooplankton⁶. These studies, combined with
 342 the global associations that were detected in this study, emphasize the potential
 343 importance of NCLDV in structuring eukaryotic communities.

344 Our results indicate that marine phytoplankton lineages could represent one of the
 345 most important host groups of NCLDV. Therefore, NCLDV could be involved in the
 346 regulation of biogeochemical processes mediated by phytoplankton. We investigated this
 347 by assessing the vertical connectivity of viral communities. The NMDS analysis showed
 348 clear differences between the NCLDV community composition of epipelagic (euphotic)
 349 and mesopelagic (aphotic) zones at most sampling sites (Fig. 2B). Similar results were
 350 also reported for phage communities in the Pacific Ocean⁴⁶. The vertical separation of
 351 viral communities may be caused by the stable stratification below the mixed layers
 352 (typically above 200 m depth), which severely inhibits vertical water exchange. Despite
 353 this limitation, mesopelagic ecosystems shared a significant number (98.7%) of NCLDV
 354 phylotypes with the upper epipelagic layers (Fig. 3A), suggesting the vertical connectivity
 355 of NCLDV and their local adaptation. Indeed, some mesopelagic NCLDV communities
 356 were very similar to surface communities (Fig. 6A and Extended Data Fig. 7A). This
 357 implies that the surface and mesopelagic NCLDV communities may be connected at some
 358 locations. The major source of energy and materials in the mesopelagic layer is the
 359 gravitational export of organic particles from the surface layer (i.e., the biological carbon
 360 pump)⁴⁷⁻⁴⁹. Therefore, some surface viruses may be exported to mesopelagic layers with
 361 sinking aggregated phytoplankton cells⁵⁰⁻⁵².

362 A significant positive correlation existed between surface phytoplankton biomass and
363 NCLDV community similarity across depths (Fig. 6B and Extended Data Fig. 7B). Since
364 highly productive areas are likely to have a greater flux of settling particles to the deep
365 layers, this result supports the idea that NCLDVs are transported with the sinking particles.
366 High vertical connectivity was consistently associated with an increase in NCLDV
367 richness in the mesopelagic zone (Fig. 6C and Extended Data Fig. 7C). Previous studies
368 showed that sinking particles can transfer bacterial and phage populations to the deep
369 layer^{52,53}. Mestre et al.⁵² demonstrated that particle-attached prokaryotes had higher
370 capacity for immigration than free-living ones. Based on the particle-driven vertical
371 dispersion model, we can expect that NCLDVs, inside or attached to their host cells or
372 cell debris, might be preferentially exported into the deep sea. Numerous studies based
373 on sediment trap measurement have shown that larger phytoplankton, such as diatoms,
374 contribute strongly to vertical flux because of their high sinking velocities^{54,55}. However,
375 recent studies show that smaller phytoplankton including haptophytes and chlorophytes,
376 known hosts of marine NCLDVs, also contribute greatly to downward carbon export^{8,9,56}.
377 The high vertical connectivity of NCLDVs was not affected by the extent of the depth
378 range nor by proxies for vertical mixing (Figs. 6D–F and Extended Data Figs. 7D–F),
379 indicating that the migration of NCLDVs occurred regardless of physical processes such
380 as upwelling, turbulent mixing, and convection. This result suggests that sinking export
381 is a major source of a variety of NCLDVs to deeper waters, where NCLDV diversity is
382 relatively low without this effect. A recent study revealed that some *Phycodnaviridae* and
383 *Mimiviridae* potentially accelerate biological carbon export from the productive surface
384 layer to deep layers, presumably by promoting cell death and aggregation of their host
385 species⁵⁷. *Phycodnaviridae* and *Mimiviridae* also contributed strongly to high vertical
386 connectivity in our study (Extended Data Figs. 8 and 9). The infection of the
387 coccolithophore by the *Phycodnaviridae* EhV was observed to facilitate the sinking of
388 host cells, likely by enhancing the production of transparent exopolymer particles and
389 subsequent aggregation⁹. Therefore, the high vertical connectivity of NCLDVs detected

390 in our analysis may be partly associated with enhanced vertical export of their infected
391 hosts.

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

405

406 **Methods**

407 **Sample collection**

408 Metagenomic datasets were generated from samples collected by the *Tara* Oceans
409 expeditions from 2009 to 2013^{26-28,31,59}. The second version of the Ocean Microbial
410 Reference Gene Catalog (OM-RGC.v2) is a non-redundant gene catalog constructed from
411 370 metagenomic samples from the *Tara* Oceans project²⁸ ([https://www.ocean-](https://www.ocean-microbiome.org)
412 [microbiome.org](https://www.ocean-microbiome.org)). The catalog includes 46,775,154 genes in total, and the gene abundance
413 profiles are expressed as the sum of within-reads aligned base pairs normalized by gene
414 length, in *Tara* Oceans samples²⁸.

415

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

417 To assess the community composition of NCLDVs, we used family B DNA

418 polymerase (*polB*) as a marker gene of NCLDV. Initially, amino acid sequences of the
419 OM-RGC.v2 were searched against an in-house profile hidden Markov model (HMM) of
420 NCLDV PolB sequences using the software HMMER, *hmmsearch* (version 3.1)⁶⁰ with a
421 threshold E-value $<1 \times 10^{-5}$. Consequently, 29,315 PolB sequences were obtained from the
422 OM-RGC.v2, although this collection included sequences other than NCLDVs. To
423 remove the sequences not derived from NCLDVs and classify the taxonomic identity of
424 each NCLDV sequence, phylogenetic mapping was performed within known PolB
425 sequences. A maximum-likelihood (ML) reference phylogenetic tree was built based on
426 211 PolB reference protein sequences from eukaryotes, bacteria, archaea, phages and
427 NCLDVs. These sequences were aligned using the default settings of the multiple
428 sequence alignment program MAFFT-linsi (version 7)⁶¹ and ML tree was constructed
429 with the use of randomized accelerated maximum likelihood (RAxML) program (version
430 7.2.8)⁶². In the reference trees, we included sequences from eight proposed families of
431 NCLDVs⁶³: *Mimiviridae* (synonymous with *Megaviridae*), *Phycodnaviridae*,
432 *Pithoviridae*, *Marseilleviridae*, *Ascoviridae*, *Iridoviridae*, *Asfarviridae*, and *Poxviridae*
433 (Extended Data Figs. 2–4). A sequence from a novel NCLDV clade *Medusavirus* was
434 also included as a reference⁶⁴. Query sequences were aligned against the reference
435 alignment using the MAFFT ‘addfragments’ option, and then mapped onto the reference
436 tree using the software program *pplacer*²⁹.

437

438 **Abundance profiling of NCLDVs**

439 We used the abundance profile of NCLDV genes from the OM-RGC.v2 to evaluate
440 the relative frequency and diversity of NCLDVs. In the abundance matrix, we only
441 included samples from the pico-size (0.22–1.6 or 0.22–3.0 μm) and femto-size (<0.22
442 μm) fractions. Samples used in the analysis were from three depth ranges: the surface (2–
443 9 m), the deep chlorophyll maximum (DCM, 15–180 m) and the mesopelagic (MES, 250–
444 1,000 m). The sum of length-normalized PolB abundances ranged from 5.3 to 22,847.5
445 across samples. The samples containing low PolB abundances tended to yield lower

446 diversity estimates (i.e., number of phylotypes and Shannon's entropy) (Extended Data
447 Fig. 10). To avoid bias due to the low sequencing effort, samples for which the sum of
448 length-normalized PolB abundance was less than 50 (set as a proxy for low NCLDV
449 frequency) were removed from the analysis. The abundance matrix was then standardized
450 by the sample with the lowest sum of length-normalized PolB abundance value. The
451 minimum value of PolB abundance among NCLDV phylotypes in the sample having the
452 lowest sum of length normalized PolB was set as the cutoff threshold. For each sample,
453 NCLDV phylotypes with a length-normalized abundance of less than this threshold were
454 treated as absent. A sample of a femto-size fraction of surface water from station 155 was
455 also removed, because it contained only one NCLDV PolB after standardization.
456 Consequently, our dataset was comprised of 283 samples (172 pico-fraction samples and
457 111 femto-fraction samples), covering 88 sampling sites. These sites were categorized
458 into four biomes (coastal, trades, westerlies and polar biomes) according to latitude or
459 distance from the shore, and nine oceanic regions, as defined by Longhurst⁶⁵
460 (Supplementary Table 1).

461

462 **Phylogenetic tree construction**

463 To construct a phylogenetic tree, the NCLDV-derived PolB sequences obtained from
464 the OM-RGC.v2 were filtered by length (≥ 700 amino acid sequences) because the
465 inclusion of short sequences yields unreliable phylogenies. Amino acid sequences from
466 the resulting 911 genes were aligned with known NCLDV sequences using the *linsi*
467 option from the MAFFT. The ML tree was constructed using RAxML with the use of a
468 known NCLDV sequence tree as a backbone constraint. We confirmed the validity of the
469 pplacer family assignment for 905 out of 911 selected sequences. The remaining six
470 sequences that were incorrectly placed within the phylogenetic tree were removed. The
471 ML tree was visualized using the program iTOL⁶⁶.

472

473 **Prediction of potential chrysophyte viruses using metagenomic assembled genomes**

474 To explore the genomic contents of environmental NCLDV, we made use of two sets
475 of metagenome-assembled genomes (MAGs) of NCLDV (GVMAGs high and medium
476 quality¹⁹; MoMAGs²⁰), which were generated from environmental metagenomic datasets
477 collected on global scales. Gene prediction was made for all MAGs using the program
478 GeneMarkS⁶⁷, then the predicted genes were searched using BLASTP against a database
479 that combines the NCBI Reference Sequence database (RefSeq release 90) and the marine
480 microbial eukaryote transcriptomes project (MMETSP) database⁶⁸. We identified MAGs
481 whose genes exhibited the best hit to transcripts of chrysophytes with >50% amino acid
482 identity and >100 alignment length (Supplementary Data 1). For these MAGs, we
483 checked the redundancy between the MoMAG and GVMAG datasets using average
484 nucleotide identity of $\geq 95\%$ and an alignment fraction of $\geq 50\%$ with FastANI (version
485 1.3)⁶⁹. Although seven MAGs were found to be overlapped between the two datasets
486 (Supplementary Data 1), all of the MAGs were retained for downstream analyses as these
487 had different contig structures. The chrysophyte-related genes were considered potential
488 candidates for horizontal gene transfer between chrysophytes and NCLDV, and were
489 BLASTP searched against the RefSeq database for additional functional annotation
490 (Supplementary Data 2). We then extracted PolB sequences from the NCLDV MAGs
491 which had a chrysophyte-related gene using the HMMER hmmsearch program. These
492 PolBs were BLASTP searched against the NCLDV PolBs from the OM-RGC.v2. MAG-
493 derived PolBs aligned with over 700 amino acid sequences with >90% identity were
494 assigned to the PolB phylotypes derived from the OM-RGC.v2 (Supplementary Data 3).
495 Phylogenetic affiliations of PolB from the chrysophyte-related MAGs were confirmed
496 using a phylogenetic tree. To further test the credibility of our analysis, we checked other
497 genes on the contigs that harbored the chrysophyte homologs using BLASTP against the
498 RefSeq database (Supplementary Data 4; Extended Data Fig. 6).

499

500 **Diversity analyses**

501 Diversity and multivariate analyses were performed using the statistical software R

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

511

512 **Partial Mantel test**

513 A partial Mantel test was performed to assess the correlation between two multivariate
514 matrices while controlling the potential effects of geographic distance (spatial
515 autocorrelation) using the R package 'vegan'. Abundance matrices for the NCLDV and
516 eukaryotic lineages were constructed from the integrated abundance tables, and the total
517 abundance at each site was normalized to 1. The eukaryote abundance table was
518 constructed based on 18S rRNA gene metabarcoding⁷¹. Data for NCLDVs were obtained
519 from pico- (0.22–1.6/3.0 μm) or femto-size (<0.2 μm) fractions and for the eukaryotic
520 community from the pico- to meso-size fraction (0.8–2,000 μm). There were 84
521 overlapping sampling events between pico-size NCLDVs and eukaryotic communities
522 and 55 overlapping sampling events between femto-size NCLDVs and eukaryotic
523 communities. All overlapping samples were derived from the surface or DCM depth
524 layers. Distance matrices for viruses and eukaryotes were calculated using the Bray-
525 Curtis measure. Geographic distances among sample sites were also measured using
526 Haversine distance and were used as a third distance matrix. Partial Mantel correlations
527 were computed between all pairs of distance matrices of eukaryotic communities and
528 NCLDVs with 9,999 permutations for each comparison. The false discovery rate (FDR)
529 was computed using the Benjamini-Hochberg method⁷².

530

531 **Statistical test**

532 Two-sided test was applied for all statistical tests.

533

534 **Data availability**

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

543

544 **Code availability**

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

547

548 Figure legends

- 549 **Figure 1 Latitudinal patterns in NCLDV community composition.** Relative
 550 contributions of NCLDV families at each depth range of (A) pico- and (B)
 551 femto-size fractions. The number of phylotypes detected in each sample is also
 552 indicated with a white circle. Sampling stations were arranged in rows from
 553 south to north, and color-coded based on biome (for a map of the sampling
 554 stations, please see Salazar et al., 2019²⁸).
- 555 **Figure 2 Community characteristics of NCLDVs.** Non-metric multidimensional
 556 scaling (NMDS) ordination based on the NCLDV community showing results
 557 for all samples (A) and separately for pico- and femto-size fractions (B and C).
 558 Sample groups are color-coded by size fraction (A), depth (B) and biome (C).
 559 Ellipses represent 90% confidence levels for each group. All group categories
 560 are significantly different from each other as analyzed using PERMANOVA (p
 561 <0.01). Sample sizes for the test are noted in Supplementary Table 1.
- 562 **Figure 3 Structural differentiation of NCLDV community across ecological zones.**
 563 (A) Venn diagrams showing the numbers of shared or unique NCLDVs
 564 phylotypes across size fractions (left), depths (center) and biomes (right). (B)
 565 Map showing the number of total, unique and shared NCLDVs across nine
 566 oceanic regions. The map was drawn using the R package ‘maps’
 567 (<https://cran.r-project.org/web/packages/maps>). (C) Relationships among
 568 sample size and total or unique NCLDVs detected in each region.
 569 Abbreviations: SO: Southern Ocean; RS: Red Sea; MS: Mediterranean Sea;
 570 NPO: North Pacific Ocean; NAO: North Atlantic Ocean; SAO: South Atlantic
 571 Ocean; SPO: South Pacific Ocean; IO: Indian Ocean; AO: Arctic Ocean.
- 572 **Figure 4 Phylogenetic affiliations of environmental NCLDVs and their dispersal**
 573 **characteristics.** Phylogenetic tree constructed from 905 long (≥ 700 amino
 574 acid) PolB sequences from the OM-RGC.v2 and 67 known NCLDV sequences
 575 (see also Extended Data Figs. 2–4 for details). The first six layers indicate the
 576 occurrence of NCLDVs unique to each size fraction, depth and biome. The
 577 outside layer denotes phylogenetic positions of known sequences (color code
 578 as in the legend) and the phylotypes closely related ($>90\%$ amino acid identity)
 579 to those of NCLDV MAGs having chrysofyte homologs (indicated in yellow).
 580 Abbreviations: OLPV-2: *Organic Lake phycodnavirus 2*; OLPV-1: *Organic*
 581 *Lake phycodnavirus 1*; CeV: *Chrysochromulina ericina* virus 1; PgV:
 582 *Phaeocystis globosa* virus 16T; HeV: *Haptolina ericina* virus RF02; PkV-2;
 583 *Prymnesium kappa* virus RF02; TetV-1: *Tetraselmis* virus 1; PoV:
 584 *Pyramimonas orientalis* virus 1; AaV: *Aureococcus anophagefferens* virus
 585 BtV-01; PkV-1; *Prymnesium kappa* virus RF01; ChoanoV: ChoanoVirus;
 586 CroV: *Cafeteria roenbergensis* virus BV-PW1; MpV-1: *Micromonas* sp.

587 RCC1109 virus MpV1; OIV-1: *Ostreococcus lucimarinus* virus 1; Otv-1:
588 *Ostreococcus tauri* virus 1; Otv-2: *Ostreococcus tauri* virus 2; MpV-12T:
589 *Micromonas pusilla* virus 12T; BpV-1: *Bathycoccus* sp. RCC1105 virus; BCV-
590 FR483: *Paramecium bursaria* *Chlorella* virus FR-483; ACTV-1:
591 *Acanthocystis turfacea* *Chlorella* virus 1; PBCV-1: *Paramecium bursaria*
592 *Chlorella* virus 1; EhV-86: *Emiliana huxleyi* virus 86; FsV: *Feldmannia*
593 *species* virus; EsV-1: *Ectocampus siliculou* virus 1; *P. salinus*: *Pandoravirus*
594 *salinus*; *P. dulcis*: *Pandoravirus dulcis*; HaV-1: *Heterosigma akashiwo* virus
595 1.

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

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

614

615

616 **References**

- 617 1 Field, C. B., Behrenfeld, M. J., Randerson, J. T. & Falkowski, P. Primary
618 production of the biosphere: integrating terrestrial and oceanic components.
619 *Science* **281**, 237-240, doi:10.1126/science.281.5374.237 (1998).
- 620 2 Worden, A. Z. *et al.* Environmental science. Rethinking the marine carbon cycle:
621 factoring in the multifarious lifestyles of microbes. *Science* **347**, 1257594,
622 doi:10.1126/science.1257594 (2015).
- 623 3 Brum, J. R. & Sullivan, M. B. Rising to the challenge: accelerated pace of
624 discovery transforms marine virology. *Nat Rev Microbiol* **13**, 147-159,
625 doi:10.1038/nrmicro3404 (2015).
- 626 4 Selosse, M.-A., Charpin, M. & Not, F. Mixotrophy everywhere on land and in
627 water: the grand écart hypothesis. *Ecology Letters* **20**, 246-263,
628 doi:10.1111/ele.12714 (2017).
- 629 5 Weitz, J. S. *et al.* A multitrophic model to quantify the effects of marine viruses
630 on microbial food webs and ecosystem processes. *Isme j* **9**, 1352-1364,
631 doi:10.1038/ismej.2014.220 (2015).
- 632 6 Mojica, K. D., Huisman, J., Wilhelm, S. W. & Brussaard, C. P. Latitudinal
633 variation in virus-induced mortality of phytoplankton across the North Atlantic
634 Ocean. *Isme j* **10**, 500-513, doi:10.1038/ismej.2015.130 (2016).
- 635 7 Suttle, C. A. Marine viruses--major players in the global ecosystem. *Nat Rev*
636 *Microbiol* **5**, 801-812, doi:10.1038/nrmicro1750 (2007).
- 637 8 Guidi, L. *et al.* Plankton networks driving carbon export in the oligotrophic ocean.
638 *Nature* **532**, 465-470, doi:10.1038/nature16942 (2016).
- 639 9 Laber, C. P. *et al.* Coccolithovirus facilitation of carbon export in the North
640 Atlantic. *Nat Microbiol* **3**, 537-547, doi:10.1038/s41564-018-0128-4 (2018).
- 641 10 Colson, P. *et al.* "Megavirales", a proposed new order for eukaryotic
642 nucleocytoplasmic large DNA viruses. *Arch Virol* **158**, 2517-2521,
643 doi:10.1007/s00705-013-1768-6 (2013).

- 644 11 Fischer, M. G. Giant viruses come of age. *Curr Opin Microbiol* **31**, 50-57,
645 doi:10.1016/j.mib.2016.03.001 (2016).
- 646 12 Koonin, E. V. & Yutin, N. Evolution of the Large Nucleocytoplasmic DNA
647 Viruses of Eukaryotes and Convergent Origins of Viral Gigantism. *Adv Virus Res*
648 **103**, 167-202, doi:10.1016/bs.aivir.2018.09.002 (2019).
- 649 13 Monier, A., Claverie, J. M. & Ogata, H. Taxonomic distribution of large DNA
650 viruses in the sea. *Genome Biol* **9**, R106, doi:10.1186/gb-2008-9-7-r106 (2008).
- 651 14 Hingamp, P. *et al.* Exploring nucleo-cytoplasmic large DNA viruses in Tara
652 Oceans microbial metagenomes. *ISME J* **7**, 1678-1695,
653 doi:10.1038/ismej.2013.59 (2013).
- 654 15 Clerissi, C. *et al.* Deep sequencing of amplified Prasinovirus and host green algal
655 genes from an Indian Ocean transect reveals interacting trophic dependencies and
656 new genotypes. *Environ Microbiol Rep* **7**, 979-989, doi:10.1111/1758-2229.12345
657 (2015).
- 658 16 Li, Y. *et al.* The Earth Is Small for "Leviathans": Long Distance Dispersal of Giant
659 Viruses across Aquatic Environments. *Microbes Environ* **34**, 334-339,
660 doi:10.1264/jsme2.ME19037 (2019).
- 661 17 Mihara, T. *et al.* Taxon Richness of "Megaviridae" Exceeds those of Bacteria and
662 Archaea in the Ocean. *Microbes Environ* **33**, 162-171,
663 doi:10.1264/jsme2.ME17203 (2018).
- 664 18 Li, Y. *et al.* Degenerate PCR Primers to Reveal the Diversity of Giant Viruses in
665 Coastal Waters. *Viruses* **10**, 496, doi:10.3390/v10090496 (2018).
- 666 19 Schulz, F. *et al.* Giant virus diversity and host interactions through global
667 metagenomics. *Nature*, doi:10.1038/s41586-020-1957-x (2020).
- 668 20 Moniruzzaman, M., Martinez-Gutierrez, C. A., Weinheimer, A. R. & Aylward, F.
669 O. Dynamic genome evolution and complex virocell metabolism of globally-
670 distributed giant viruses. *Nat Commun* **11**, 1710, doi:10.1038/s41467-020-15507-
671 2 (2020).

- 672 21 Cottrell, M. T. & Suttle, C. A. Wide-spread occurrence and clonal variation in
673 viruses which cause lysis of a cosmopolitan, eukaryotic marine phytoplankter,
674 *Micromonas pusilla*. *Mar Ecol Prog Ser* **78** (1991).
- 675 22 Bratbak, G., Egge, J. K. & Heldal, M. Viral mortality of the marine alga *Emiliana*
676 *huxleyi* (Haptophyceae) and termination of algal blooms. *Marine Ecology*
677 *Progress Series* **93**, 39-48 (1993).
- 678 23 Kenji, T., Keizo, N., Shigeru, I. & Mineo, Y. Isolation of a virus infecting the
679 novel shellfish-killing dinoflagellate *Heterocapsa circularisquama*. *Aquatic*
680 *Microbial Ecology* **23**, 103-111 (2001).
- 681 24 Fischer, M. G., Allen, M. J., Wilson, W. H. & Suttle, C. A. Giant virus with a
682 remarkable complement of genes infects marine zooplankton. *Proc Natl Acad Sci*
683 *U S A* **107**, 19508-19513, doi:10.1073/pnas.1007615107 (2010).
- 684 25 Needham, D. M. *et al.* A distinct lineage of giant viruses brings a rhodopsin
685 photosystem to unicellular marine predators. *Proc Natl Acad Sci U S A* **116**,
686 20574-20583, doi:10.1073/pnas.1907517116 (2019).
- 687 26 Pesant, S. *et al.* Open science resources for the discovery and analysis of Tara
688 Oceans data. *Sci Data* **2**, 150023, doi:10.1038/sdata.2015.23 (2015).
- 689 27 Gregory, A. C. *et al.* Marine DNA Viral Macro- and Microdiversity from Pole to
690 Pole. *Cell* **177**, 1109-1123 e1114, doi:10.1016/j.cell.2019.03.040 (2019).
- 691 28 Salazar, G. *et al.* Gene Expression Changes and Community Turnover
692 Differentially Shape the Global Ocean Metatranscriptome. *Cell* **179**, 1068-1083
693 e1021, doi:10.1016/j.cell.2019.10.014 (2019).
- 694 29 Matsen, F. A., Kodner, R. B. & Armbrust, E. V. pplacer: linear time maximum-
695 likelihood and Bayesian phylogenetic placement of sequences onto a fixed
696 reference tree. *BMC Bioinformatics* **11**, 538, doi:10.1186/1471-2105-11-538
697 (2010).
- 698 30 Gallot-Lavallee, L., Blanc, G. & Claverie, J. M. Comparative Genomics of
699 *Chrysochromulina ericina* Virus and Other Microalga-Infecting Large DNA

- 700 Viruses Highlights Their Intricate Evolutionary Relationship with the Established
701 Mimiviridae Family. *J Virol* **91**, doi:10.1128/jvi.00230-17 (2017).
- 702 31 Ibarbalz, F. M. *et al.* Global Trends in Marine Plankton Diversity across
703 Kingdoms of Life. *Cell* **179**, 1084-1097 e1021, doi:10.1016/j.cell.2019.10.008
704 (2019).
- 705 32 Mihara, T. *et al.* Linking Virus Genomes with Host Taxonomy. *Viruses* **8**, 66,
706 doi:10.3390/v8030066 (2016).
- 707 33 Ogata, H. *et al.* Remarkable sequence similarity between the dinoflagellate-
708 infecting marine virus and the terrestrial pathogen African swine fever virus. *Virology*
709 **6**, 178, doi:10.1186/1743-422X-6-178 (2009).
- 710 34 Andreani, J. *et al.* Pacmanvirus, a New Giant Icosahedral Virus at the Crossroads
711 between Asfarviridae and Faustoviruses. *J Virol* **91**, doi:10.1128/JVI.00212-17
712 (2017).
- 713 35 Barton, A. D., Dutkiewicz, S., Flierl, G., Bragg, J. & Follows, M. J. Patterns of
714 diversity in marine phytoplankton. *Science* **327**, 1509-1511,
715 doi:10.1126/science.1184961 (2010).
- 716 36 Lima-Mendez, G. *et al.* Ocean plankton. Determinants of community structure in
717 the global plankton interactome. *Science* **348**, 1262073,
718 doi:10.1126/science.1262073 (2015).
- 719 37 Zhou, J. & Ning, D. Stochastic Community Assembly: Does It Matter in
720 Microbial Ecology? *Microbiol Mol Biol Rev* **81**, doi:10.1128/membr.00002-17
721 (2017).
- 722 38 Chow, C. E. & Suttle, C. A. Biogeography of Viruses in the Sea. *Annu Rev Virol*
723 **2**, 41-66, doi:10.1146/annurev-virology-031413-085540 (2015).
- 724 39 Yoshida, T. *et al.* Locality and diel cycling of viral production revealed by a 24 h
725 time course cross-omics analysis in a coastal region of Japan. *ISME J* **12**, 1287-
726 1295, doi:10.1038/s41396-018-0052-x (2018).
- 727 40 Sunagawa, S. *et al.* Ocean plankton. Structure and function of the global ocean

- 728 microbiome. *Science* **348**, 1261359, doi:10.1126/science.1261359 (2015).
- 729 41 Syed, T. H., Famiglietti, J. S., Zlotnicki, V. & Rodell, M. Contemporary estimates
 730 of Pan-Arctic freshwater discharge from GRACE and reanalysis. *Geophysical*
 731 *Research Letters* **34**, doi:10.1029/2007gl031254 (2007).
- 732 42 Wommack, K. E. & Colwell, R. R. Virioplankton: viruses in aquatic ecosystems.
 733 *Microbiol Mol Biol Rev* **64**, 69-114, doi:10.1128/mmbr.64.1.69-114.2000 (2000).
- 734 43 Bellec, L. *et al.* Cophylogenetic interactions between marine viruses and
 735 eukaryotic picophytoplankton. *BMC Evol Biol* **14**, 59, doi:10.1186/1471-2148-
 736 14-59 (2014).
- 737 44 Brussaard, C. P. D., Kempers, R. S., Kop, A. J., Riegman, R. & Heldal, M. Virus-
 738 like particles in a summer bloom of *Emiliana huxleyi* in the North Sea. *Aquatic*
 739 *Microbial Ecology* **10**, 105-113 (1996).
- 740 45 Stephan, J. *et al.* Flow cytometric analysis of an *Emiliana huxleyi* bloom
 741 terminated by viral infection. *Aquatic Microbial Ecology* **27**, 111-124 (2002).
- 742 46 Hurwitz, B. L., Westveld, A. H., Brum, J. R. & Sullivan, M. B. Modeling
 743 ecological drivers in marine viral communities using comparative metagenomics
 744 and network analyses. *Proc Natl Acad Sci U S A* **111**, 10714-10719,
 745 doi:10.1073/pnas.1319778111 (2014).
- 746 47 Herndl, G. J. & Reinthaler, T. Microbial control of the dark end of the biological
 747 pump. *Nat Geosci* **6**, 718-724, doi:10.1038/ngeo1921 (2013).
- 748 48 Giering, S. L. *et al.* Reconciliation of the carbon budget in the ocean's twilight
 749 zone. *Nature* **507**, 480-483, doi:10.1038/nature13123 (2014).
- 750 49 Boyd, P. W., Claustre, H., Levy, M., Siegel, D. A. & Weber, T. Multi-faceted
 751 particle pumps drive carbon sequestration in the ocean. *Nature* **568**, 327-335,
 752 doi:10.1038/s41586-019-1098-2 (2019).
- 753 50 Janice, E. L. & Curtis, A. S. Effect of viral infection on sinking rates of
 754 *Heterosigma akashiwo* and its implications for bloom termination. *Aquatic*
 755 *Microbial Ecology* **37**, 1-7 (2004).

- 756 51 Close, H. G. *et al.* Export of submicron particulate organic matter to mesopelagic
757 depth in an oligotrophic gyre. *Proc Natl Acad Sci U S A* **110**, 12565-12570,
758 doi:10.1073/pnas.1217514110 (2013).
- 759 52 Mestre, M. *et al.* Sinking particles promote vertical connectivity in the ocean
760 microbiome. *Proc Natl Acad Sci U S A* **115**, E6799-E6807,
761 doi:10.1073/pnas.1802470115 (2018).
- 762 53 Hurwitz, B. L., Brum, J. R. & Sullivan, M. B. Depth-stratified functional and
763 taxonomic niche specialization in the 'core' and 'flexible' Pacific Ocean Virome.
764 *Isme j* **9**, 472-484, doi:10.1038/ismej.2014.143 (2015).
- 765 54 Sancetta, C., Villareal, T. & Falkowski, P. Massive fluxes of rhizosolenid diatoms:
766 A common occurrence? *Limnology and Oceanography* **36**, 1452-1457,
767 doi:10.4319/lo.1991.36.7.1452 (1991).
- 768 55 Kawakami, H. & Honda, M. C. Time-series observation of POC fluxes estimated
769 from ²³⁴Th in the northwestern North Pacific. *Deep Sea Research Part I:
770 Oceanographic Research Papers* **54**, 1070-1090, doi:10.1016/j.dsr.2007.04.005
771 (2007).
- 772 56 Richardson, T. L. & Jackson, G. A. Small phytoplankton and carbon export from
773 the surface ocean. *Science* **315**, 838-840, doi:10.1126/science.1133471 (2007).
- 774 57 Blanc-Mathieu, R. *et al.* Viruses of the eukaryotic plankton are predicted to
775 increase carbon export efficiency in the global sunlit ocean. *bioRxiv*, 710228,
776 doi:10.1101/710228 (2019).
- 777 58 Iversen, M. H. & Ploug, H. Ballast minerals and the sinking carbon flux in the
778 ocean: carbon-specific respiration rates and sinking velocity of marine snow
779 aggregates. *Biogeosciences* **7**, 2613-2624, doi:10.5194/bg-7-2613-2010 (2010).
- 780 59 Alberti, A. *et al.* Viral to metazoan marine plankton nucleotide sequences from
781 the Tara Oceans expedition. *Sci Data* **4**, 170093, doi:10.1038/sdata.2017.93
782 (2017).
- 783 60 Eddy, S. R. Profile hidden Markov models. *Bioinformatics* **14**, 755-763,

- 784 doi:10.1093/bioinformatics/14.9.755 (1998).
- 785 61 Kato, K. & Standley, D. M. MAFFT multiple sequence alignment software
786 version 7: improvements in performance and usability. *Mol Biol Evol* **30**, 772-780,
787 doi:10.1093/molbev/mst010 (2013).
- 788 62 Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic
789 analyses with thousands of taxa and mixed models. *Bioinformatics* **22**, 2688-2690,
790 doi:10.1093/bioinformatics/btl446 (2006).
- 791 63 Koonin, E. V. & Yutin, N. Multiple evolutionary origins of giant viruses.
792 *F1000Res* **7**, doi:10.12688/f1000research.16248.1 (2018).
- 793 64 Yoshikawa, G. *et al.* Medusavirus, a Novel Large DNA Virus Discovered from
794 Hot Spring Water. *J Virol* **93**, doi:10.1128/JVI.02130-18 (2019).
- 795 65 Longhurst, A. R. in *Ecological Geography of the Sea (Second Edition)* (ed
796 Alan R. Longhurst) 89-102 (Academic Press, 2007).
- 797 66 Letunic, I. & Bork, P. Interactive tree of life (iTOL) v3: an online tool for the
798 display and annotation of phylogenetic and other trees. *Nucleic Acids Res* **44**,
799 W242-245, doi:10.1093/nar/gkw290 (2016).
- 800 67 Besemer, J., Lomsadze, A. & Borodovsky, M. GeneMarkS: a self-training method
801 for prediction of gene starts in microbial genomes. Implications for finding
802 sequence motifs in regulatory regions. *Nucleic Acids Res* **29**, 2607-2618,
803 doi:10.1093/nar/29.12.2607 (2001).
- 804 68 Keeling, P. J. *et al.* The Marine Microbial Eukaryote Transcriptome Sequencing
805 Project (MMETSP): illuminating the functional diversity of eukaryotic life in the
806 oceans through transcriptome sequencing. *PLoS Biol* **12**, e1001889,
807 doi:10.1371/journal.pbio.1001889 (2014).
- 808 69 Jain, C., Rodriguez, R. L., Phillippy, A. M., Konstantinidis, K. T. & Aluru, S. High
809 throughput ANI analysis of 90K prokaryotic genomes reveals clear species
810 boundaries. *Nat Commun* **9**, 5114, doi:10.1038/s41467-018-07641-9 (2018).
- 811 70 Anderson, M. J. A new method for non-parametric multivariate analysis of

- 812 variance. *Austral Ecology* **26**, 32-46, doi:10.1111/j.1442-9993.2001.01070.pp.x
813 (2001).
- 814 71 de Vargas, C. *et al.* Ocean plankton. Eukaryotic plankton diversity in the sunlit
815 ocean. *Science* **348**, 1261605, doi:10.1126/science.1261605 (2015).
- 816 72 Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical
817 and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical*
818 *Society: Series B (Methodological)* **57**, 289-300, doi:10.1111/j.2517-
819 6161.1995.tb02031.x (1995).

820

821 **Acknowledgement**

822 This work was supported by JSPS/KAKENHI (Nos. 26430184, 18H02279, and
823 19H05667 to H.O. and Nos. 19K15895 and 19H04263 to H.E.), Scientific Research on
824 Innovative Areas from the Ministry of Education, Culture, Science, Sports and
825 Technology (MEXT) of Japan (Nos. 16H06429, 16K21723, and 16H06437 to H.O.),
826 Kyoto University Research Coordination Alliance (funding to H.E.), and the
827 Collaborative Research Program of the Institute for Chemical Research, Kyoto University
828 (Nos. 2019-30 and 2020-27). Computational time was provided by the SuperComputer
829 System, Institute for Chemical Research, Kyoto University. We further thank the *Tara*
830 Oceans consortium, the projects OCEANOMICS (ANR-11-BTBR-0008) and France
831 Genomique (ANR-10-INBS-09), and the people and sponsors who supported *Tara*
832 Oceans. *Tara Oceans* (that includes both the *Tara Oceans* and *Tara Oceans Polar Circle*
833 expeditions) would not exist without the leadership of the *Tara Expeditions Foundation*
834 and the continuous support of 23 institutes (<https://oceans.taraexpeditions.org>). This
835 article is contribution number 108 of *Tara Oceans*.

836

837 **Author contributions**

838 HE and HO designed the study. HE performed most of the bioinformatics analysis.
839 RB-M and YL contributed to the bioinformatics analysis. GS, NH, KL, CdV, MBS, CB,
840 PW, LK-B, and SS contributed to the generation of primary data. CdV, MBS, CB, PW,

841 LK-B, SS, and HO coordinated *Tara* Oceans. All authors contributed to the writing of the
842 manuscript.

843

844 **Materials & Correspondence**

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

847

848 **Competing financial interests**

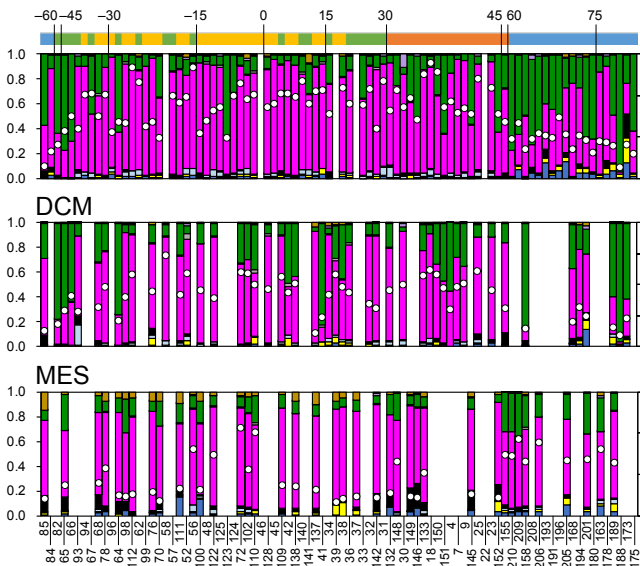
849 The authors declare no competing financial interests.

850

851

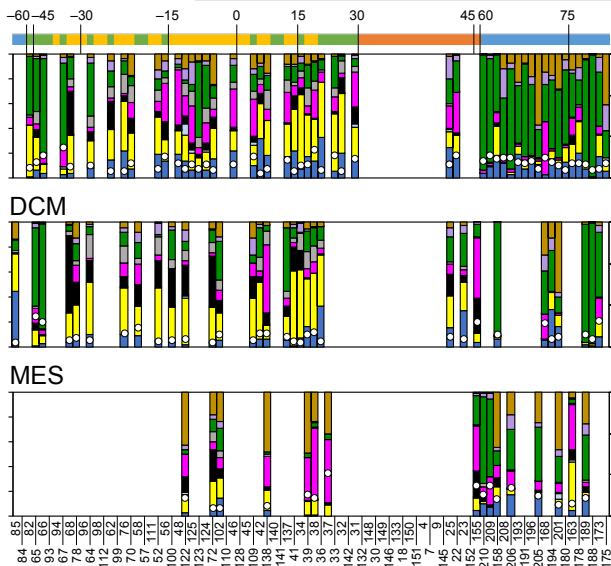
Relative contribution

Latitude (°N)

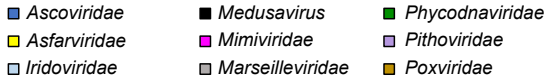


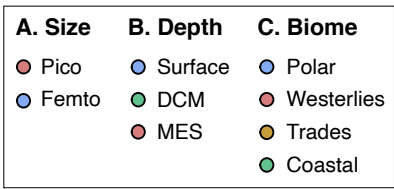
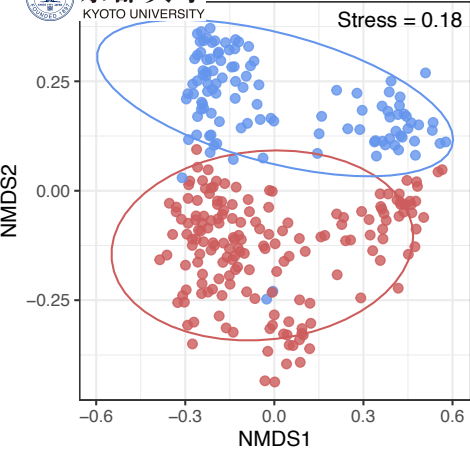
Station number

Surface

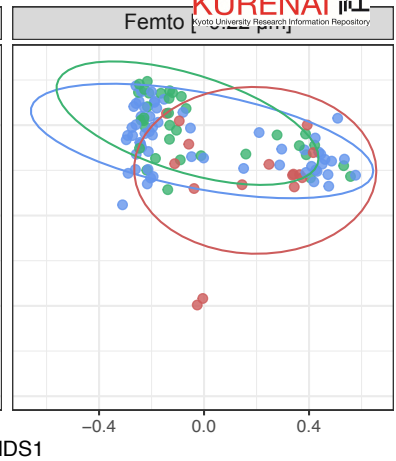
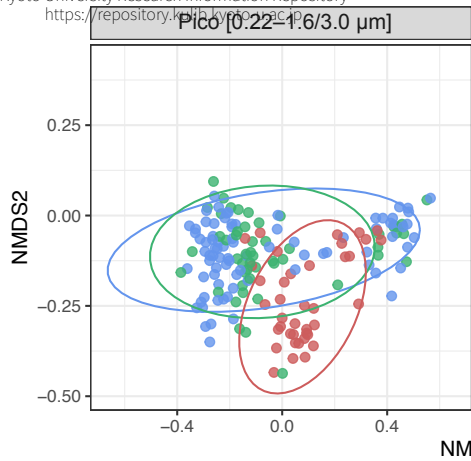


Number of phylotypes detected

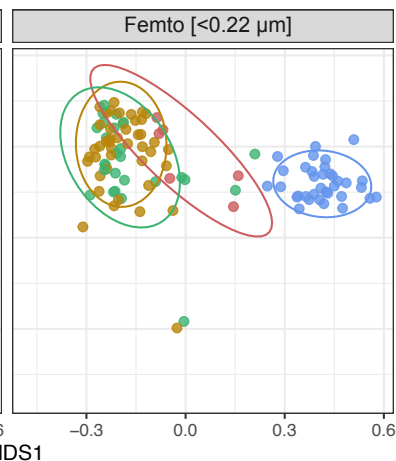
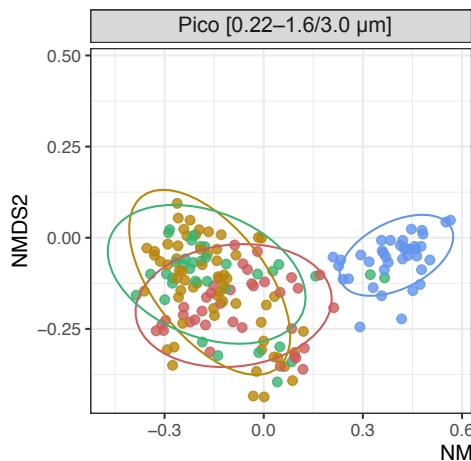




B



C



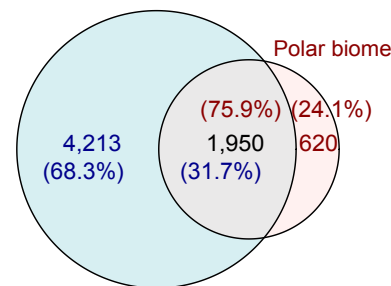
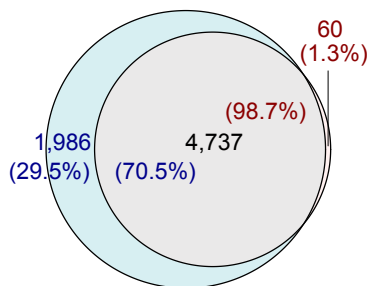
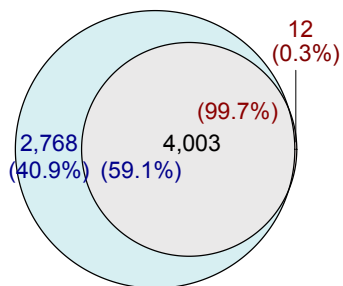
Pico-size
(0.22–1.6/3.0 μm)

Femto-size
(<0.22 μm)

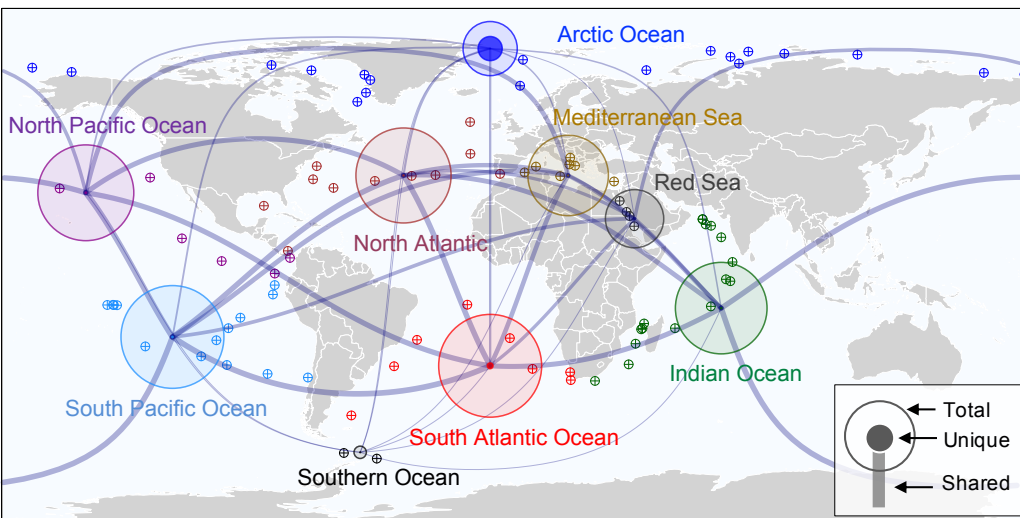
Euphotic zone
(Surface and DCM) **Mesopelagic zone**

Non-Polar biome
(Trades, Westerlies, and Coastal)

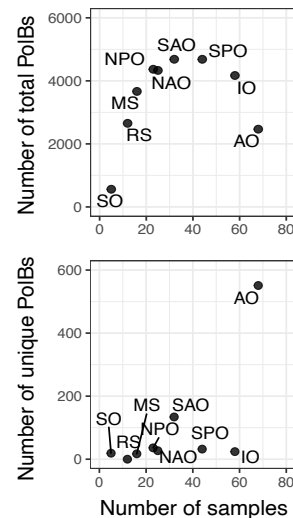
Polar biome

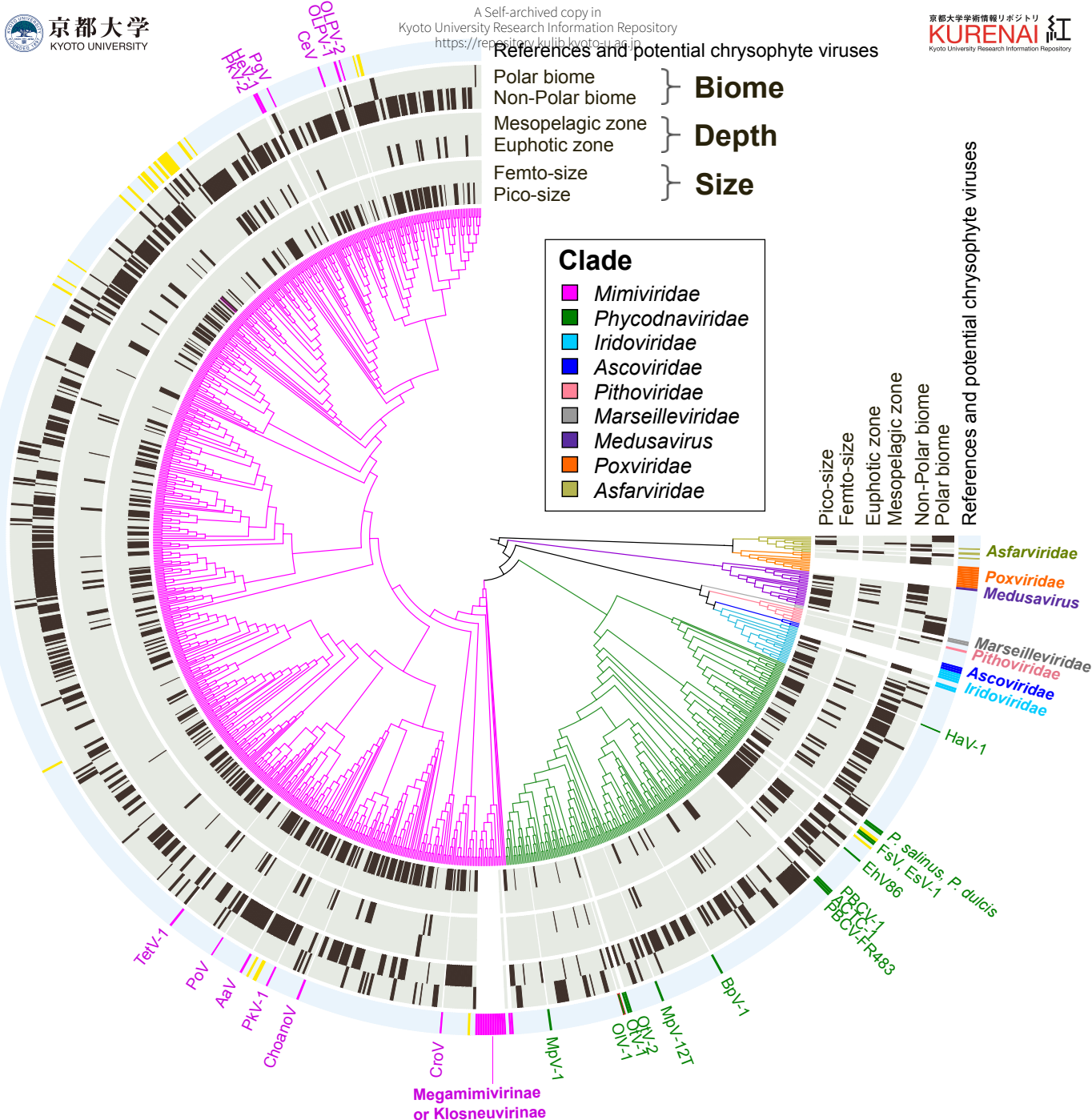


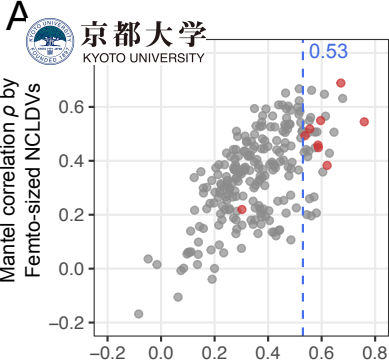
B



C

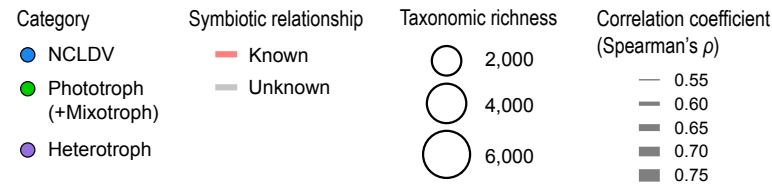
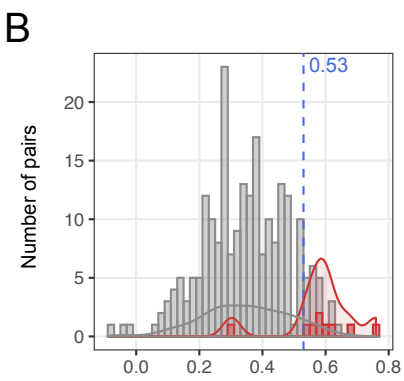
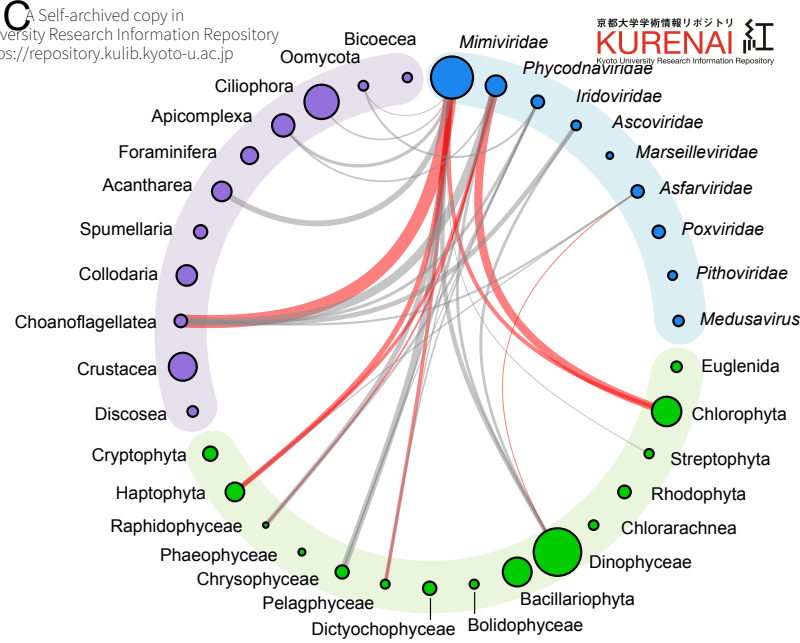






C Self-archived copy in
Kyoto University Research Information Repository
<https://repository.kulib.kyoto-u.ac.jp>

京都大学学術情報リポジトリ
KURENAI
Kyoto University Research Information Repository



Mantel correlation ρ by Pico-sized NCLDV

