Genomic adaptation of giant viruses in polar oceans

Lingjie Meng[1], Tom O. Delmont[2,3], Morgan Gaïa[2,3], Eric Pelletier[2,3], Antonio Fernàndez-Guerra[4,5], Samuel Chaffron[3,6], Russell Y. Neches[1], Junyi Wu[1], Hiroto Kaneko[1], Hisashi Endo[1], Hiroyuki Ogata[1,*]

Affiliations:

1. Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, 611-0011, Japan
2. Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ Evry, Université Paris-Saclay, F-91057 Evry, France
3. Research Federation for the study of Global Ocean systems ecology and evolution, FR2022/Tara GOsee, F-75016 Paris, France
4. Microbial Genomics and Bioinformatics Research G, Max Planck Institute for Marine Microbiology, Bremen, Germany
5. Lundbeck Foundation GeoGenetics Centre, GLOBE Institute, University of Copenhagen, Copenhagen, Denmark
6. Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000 Nantes, France

Corresponding author:

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

Despite being perennially frigid, polar oceans form an ecosystem hosting high and unique biodiversity. Various organisms show different adaptative strategies in this habitat, but how viruses adapt to this environment is largely unknown. Viruses of phyla *Nucleocytoviricota* and *Mirusviricota* are groups of eukaryote-infecting large and giant DNA viruses with genomes encoding a variety of functions. Here, by leveraging the Global Ocean Eukaryotic Viral database, we investigate the biogeography and functional repertoire of these viruses at a global scale. We first confirm the existence of an ecological barrier that clearly separates polar and nonpolar viral communities, and demonstrate that temperature drives dramatic changes in the virus–host network at the polar/nonpolar boundary. Ancestral niche reconstruction suggests that adaptation of these viruses to polar conditions has occurred repeatedly over the course of evolution, with polar-adapted viruses in the modern ocean being scattered across their phylogeny. Numerous viral genes are specifically associated with polar adaptation, although most of their homologues are not identified as polar-adaptive genes in eukaryotes. These results suggest that giant viruses adapt to cold environments by changing their functional repertoire, and this viral evolutionary strategy is independent of the polar adaptation of their hosts.

Main

Polar regions are recognized as among the coldest environments on Earth, with strong seasonal variations in light cycles. Nevertheless, these regions could nourish a diverse range of creatures, from microscopic organisms to large animals, thanks to the primary production by phytoplankton. Organisms adapted to polar environments exhibit distinctive physiological or morphological characteristics, which augment their fitness in these extreme but lush environments. For example, polar bears show characteristic morphological traits whose underlying genetic variations occurred in their ancestral gene pools[1]. Both Arctic and Antarctic fishes encode antifreeze proteins that allow them to maintain physiological activity in cold

waters[2,3], while some psychrophilic bacteria produce oxygen-scavenging enzymes or modify their membrane fatty acid composition[4,5].

How do viruses adapt to polar environments? In the ocean, viruses are the most abundant biological entities[6] and play important roles in the regulation of microbial host communities, carbon and nutrient cycling, and horizontal gene transfer among organisms[7–10]. Recent metagenomic studies have revealed that both Arctic[11,12] and Antarctic[13,14] environments harbour diverse viruses, with an elevated diversity of prokaryotic dsDNA viruses in the Arctic Ocean[11]. A large proportion of Arctic-specific genes from these viruses were suggested under positive selection based on their mutation patterns. This implied a role for gene repertoire in viral adaptation, although most of those genes were of unknown function. It is also known that phylogenetically closely related viruses can display different responses in their infection dynamics to varying temperature[15,16], suggesting that virus–host systems adapt to thermal changes. Another study showed that a prokaryotic virus reduced its genome in response to decreased culture temperature[17]. These studies imply possible adaptive mechanisms of viruses to low temperature or polar ecosystems. However, our knowledge on such viral adaptations is still limited.

In our previous study, we revealed a remarkable shift in the community composition of eukaryotic dsDNA viruses from nonpolar to polar biomes[18]. These viruses, classified in phylum *Nucleocytoviricota* ("giant viruses"), are known for their large genomes encoding hundreds to thousands of genes[19,20]. These viruses are ancient[21], diverse[22,23], abundant[24,25], and active[26] in the ocean. Despite the existence of a clear polar/nonpolar barrier for these viruses, the underlying genomic and ecological mechanisms are unknown. How frequently these viruses have crossed this polar barrier over evolutionary time also remains unclear. As the genomes of *Nucleocytoviricota* dynamically evolve by losing and gaining different functions[19,27], we hypothesized that adaptation to polar environments impacts their gene repertoire.

In this study, we investigated genomes of eukaryotic large DNA viruses to characterize their genome-level adaptations to polar environments. We leveraged recently reconstructed viral and eukaryotic environmental genomes from the multidisciplinary *Tara* Oceans international research project[28,29]. The viral genomic data include environmental genomes of viruses from phylum *Nucleocytoviricota* and a recently discovered phylum, *Mirusviricota*[28].

We first assess the existence of a polar barrier for giant viruses by analysing viral community composition and by computing robust temperature optima for viruses and their predicted hosts. We then perform ancestral state reconstruction for polar and nonpolar niches along the phylogenomic tree of these viruses to quantitatively estimate the adaptive evolutionary events. Finally, we delineate the functions that are specific to "polar" viruses and present evidence that viral genomic adaptation to low temperature is independent from the adaptation of their hosts.

Results and Discussion

Polar barrier for giant viruses

We investigated the biogeography of giant virus genomes from the Global Ocean Eukaryotic Viral (GOEV) database[28]. Their abundance profiles across *Tara* Oceans samples from different size fractions (Supplementary Fig. 1a,b; Supplementary Table 1) revealed 1,380 viral genomes that showed signals (>25% of the genome length was mapped by reads, see methods in our previous paper[28]) in at least one sample out of 928 samples (The details of biogeography were in the supplementary text; Supplementary Fig. 1-3). The presence/absence distribution of viral genomes across biomes revealed a large number of genomes specific to the Polar biome. Out of 569 genomes detected in polar regions, 262 (46.05%) were exclusive to the Polar biome (Supplementary Fig. 4a). Accordingly, biome-based classification of viral communities (i.e., Polar, Coastal, Trades, and Westerlies) had significant explanatory power

for community variation (Supplementary Fig. 4b,c; ANOSIM, $P < 0.01$). The R value increased from 0.4021 to 0.6141 after merging three nonpolar biomes, demonstrating the existence of a clear polar barrier for giant virus communities. The viral communities of Arctic regions were also characterized by their relatively high abundances showing peaks in cumulative coverage plots for different size fractions (Supplementary Fig. 1b). The major groups of viruses in this area were *Algavirales*, followed by *Imitervirales* as in other areas of the ocean (Supplementary Fig. 2c).

We inferred a virus–plankton network through co-occurrence analysis to further characterize the polar barrier in the context of virus-host interaction. In this analysis, we combined our virus genome data with previously reconstructed marine eukaryotic genome data[29]. In total, 2,135 virus–eukaryote associations (edges) were inferred in the network, with the majority (91.94%) of them being positive associations (Fig. 1a; Supplementary Table 3). Virus–eukaryote pairs with strong associations (edge weight ≥ 0.4) showed significantly higher protein similarities between their genomes than those without strong associations (no edges or edges with weight <0.4) (Wilcoxon rank-sum test, $P = 1\times10^{-13}$) (Fig. 1b). Such an increase of sequence similarity can be due to horizontal gene transfers between viruses and hosts[30,31], supporting the prediction of true virus–host relationships in the reconstructed network. A previous study revealed that the structure of the network for marine eukaryotes and prokaryotes correlates with the temperature optima of species[32]. By estimating robust temperature optima for individual viruses and eukaryotes[33], we identified a strong correlation between the temperature optima and the structure of the virus–eukaryote network (Fig. 1a). A dramatic structural change in the network at the temperature-dependent polar/nonpolar boundary is the source of the uniqueness of polar viral communities.

Latitudinal diversity gradients are characterized by relatively low polar and high temperate biodiversity[34] and are widespread across all ranges of marine microorganisms[35].

Previous studies revealed a similar latitudinal diversity gradient for giant viruses[35], but not for prokaryotic dsDNA viruses[11,35] and RNA viruses[36]. In this study, various diversity gradient patterns were observed among viruses of different size fractions and main taxonomic groups (Supplementary Fig. 1d; Supplementary Fig. 5). The reasons underlying the Arctic diversity hotspots for some viruses (e.g., viruses in large size fractions and mirusviruses) may reflect their host ranges as previously suggested[35]. Notably, eukaryotic nodes (i.e., potential hosts) associated with viruses showed a pattern distinct from the general diversity gradient trend with increasing diversity towards high latitudinal regions (Supplementary Fig. 6).

Potential hosts for polar viruses

A phylogeny-informed filtering method, Taxon Interaction Mapper (TIM)[37,38], was applied to the edges of the network to assign predicted host taxa to viral clades. This method assigned predicted host taxa (five taxa in total) to 34 viral clades (Supplementary Fig. 7a). These predictions are summarized in Supplementary Fig. 7b and included known virus–host relationships: *Mesomimiviridae* (from *Imitervirales)* and Phaeocytales[39–41]; *Mesomimiviridae* and Pelagomonadales[42,43]; and *Prasinovirus* (from *Algavirales)* and Mamiellales[44,45].

Recent discoveries of giant endogenous viral elements (GEVEs) that are widespread across different eukaryotes demonstrated the impacts of giant viruses' infection on host genome evolution[46–49]. We systematically analysed insertions of genomes of giant viruses and their satellite viruses (i.e., virophages) in marine eukaryotic genomes[29]. Among the five eukaryotic taxa predicted to contain viral hosts, the diatom order Chaetocerotales showed the largest number of insertion signals of both giant viruses and virophages (Supplementary Fig. 7b), suggesting infection of dsDNA viruses in these diatoms. Because only ssDNA and ssRNA viruses have been reported to infect species of diatoms[50], we further analysed draft genomes of two isolated *Chaetoceros* species[51,52], revealing three putative GEVEs in *C. tenuissimus* and

149 one GEVE in *C. muelleri*. Two viral DNA polymerase genes detected in the *Chaetoceros*
150 genomes were phylogenetically placed close to *Asfuvirales* and *Imitervirales* clades
151 (Supplementary Fig. 7c), corroborating the virus–host relationships of
152 *Imitervirales*/Chaetocerotales and *Asfuvirales*/Chaetocerotales predicted by our phylogeny-
153 informed co-occurrence method. Because chaetocerotalid diatoms are abundant and diverse in
154 both the Arctic and Southern Oceans[53,54], this unidentified virus–host relationship may account
155 for the diversity of giant viruses in high-latitude regions.
156
157 **Recurrent polar adaptations throughout viral evolution**
158 To investigate viral adaptation across the polar barrier, we assigned ecological niche
159 categories, either "Polar" or "Nonpolar", to individual viral genomes. Of 1,380 viral genomes,
160 450 genomes were classified as Polar, while 818 genomes were classified as Nonpolar (Fig.
161 2a,b). 111 genomes were labelled "Unknown" because of their ambiguous distribution patterns.
162 This ecological niche assignment was consistent with the robust temperature and latitude
163 optima (Supplementary Fig. 8a). We then investigated the niche category assignment in the
164 phylogenomic tree of viruses and found numerous clades of Polar viruses scattered across the
165 tree (Fig. 3a). One Polar clade included an Arctic-original metagenome-assembled genome
166 (MAG) and organic lake phycodnaviruses derived from an Antarctic organic lake[14]
167 (Supplementary Fig. 8b). *Emiliania huxleyi* viruses, known to occur at high latitudes, were also
168 assigned to Polar clades. All six genomes of *Proculviricetes*[28], a recently discovered class-level
169 group recovered exclusively from the Arctic and Southern Oceans, were classified as Polar
170 viruses. These examples corroborate the reliability of the ecological niche assignment using
171 global-scale abundance profiles.
172 We then performed Polar/Nonpolar state reconstruction for ancestral nodes in the tree
173 using a maximum likelihood approach (see Methods). As a result, 118 Nonpolar-to-Polar and

174 95 Polar-to-Nonpolar niche adaptations were inferred along the branches of the tree (Fig. 2a).
175 These adaptations thus occurred recurrently throughout the evolution of these viruses starting
176 from the root of the tree, which was inferred as Nonpolar. Yet, our data could not exclude the
177 possibility of a polar-origin scenario due to the difficulty in determining the root of the tree of
178 giant viruses. The divergence of these viruses is estimated to predate the divergence of
179 eukaryotes[21,23]. Most of the reconstructed niche adaptations occurred relatively recently after
180 the formation of genera, but some adaptations were inferred to have occurred during the early
181 stage of evolution, corresponding to order-level divergence (Fig. 2c).
182
183 **Polar-specific viral functions and their phylogenetic distributions**
184 Genomic adaptation (i.e., adaptation by alteration of gene repertoire) to polar regions
185 was investigated based on functions encoded in the viral genomes. We first annotated genes in
186 the viral genomes with the KEGG Orthologs (KOs). For KOs (n = 1591) that were observed in
187 more than four genomes, we calculated robust temperature and latitude optima (Supplementary
188 Table 4). The temperature optima ranged from −1.54 °C to 27.31 °C, and the latitude optima
189 from 5.25 ° to 78.96 °. The distribution of these values revealed two major groups of KOs: one
190 distributed in high-latitude/low-temperature regions (n = 314, 19.74%) and another in lower-
191 latitude/higher-temperature regions (n = 1,277, 80.26%) (Fig. 3a). The 314 Polar-specific
192 genes had temperature optima below 10 °C and latitude optima above 50 °. The temperature
193 and latitude optima for conserved core genes of giant viruses were found in the second group,
194 being distributed at around 13–14 °C and 37–40 °, respectively.
195 We then calculated the phylogenetic diversity of individual KOs using the viral
196 phylogenomic tree as a reference to assess the breadth of their phylogenetic distribution
197 (Supplementary Fig. 9a). Overall, Polar-specific KOs showed a relatively low phylogenetic
198 diversity (median = 6.94) compared with other KOs (median = 9.67) (Wilcoxon rank-sum

test, $P < 0.01$), indicating relatively narrow phylogenetic distributions of the Polar-specific KOs. To further characterize the phylogenetic distributions of the 314 Polar-specific KOs, we examined the strength of phylogenetic signals in their distribution using a model comparison approach (see Methods). This analysis revealed that the reference phylogenomic tree has insufficient explanatory power for the phylogenetic distribution of 193 Polar-specific KOs (61%) out of the 314 KOs (chi-squared test, $P < 0.05$). It is thus inferred that additional factors rather than speciation history impacted the phylogenetic distribution of these KOs; environmental conditions or associated host distributions could be such factors.

Polar-specific viral functions and metabolic pathways

The proportion of polar-specific KOs (among all genes with KO annotations in a viral genome) was significantly higher in Polar genomes (15.84% on average) compared to Nonpolar (6.95%) and Unknown (7.93%) genomes (Supplementary Fig. 9b; Kruskal-Wallis test, $P < 0.01$). Among Polar-specific KOs, ceramide glucosyltransferase (K00720) and dihydrofolate reductase (K18589) were exclusively distributed in polar genomes. Ceramide glucosyltransferase catalyzes sphingolipid glycosylation, indicating the biosynthesis of viral sphingolipids may improve the fitness of polar viruses[55]. Dihydrofolate reductase could provide dTMP pools for low GC content viruses, and a possible role of this function is to facilitate the replication of viruses in the persistent infections[56]. Additionally, nitrate transporter (K02575) had a high ratio of polar to nonpolar phylogenetic diversity (ratio = 7.96), thus showing a comparatively wide phylogenetic distribution in Polar genomes. The nitrate transporter pathway has a role in assimilating extracellular nitrate/nitrite, implying a potential role for Polar viruses to reprogram host metabolism to fit the nitrate-deficient polar oceans[57]. Some metabolic functions, including CoA biosynthesis (4'-phosphopantetheinyl transferase)

and secondary metabolite biosynthesis (hydroxymandelonitrile lyase and 2-polyprenyl-6-hydroxyphenyl methylase), also showed a high phylogenetic diversity for Polar genomes.

At the pathway level, we found that six pathways were significantly enriched in Polar KOs (Fig. 3b; Fisher's exact test, $P < 0.05$). Biosynthesis of unsaturated fatty acids, was found to be the most significantly enriched with polar KOs. A high proportion of unsaturated fatty acids is known as an adaptive trait for bacteria inhabiting low temperature and high pressure environment[58]. Giant viruses isolated from high latitude areas are known to encode enzymes for the biosynthesis of unsaturated fatty acid[41] and may rewire the host physiology of fatty acid[55]. The N-glycan biosynthesis pathway also had a relatively high ratio of Polar-specific KOs. N-glycan influences the virus replication cycle, including virus recognition and virus release[59]. Neuroactive ligand-receptor interaction was the pathway most significantly enriched with polar specific KOs, implying the ability of polar viruses to regulate signal transduction. Collectively, these results underscore the importance of membrane-related pathways, including unsaturated fatty acid and specific membrane-related functions, in polar virus–host interactions.

Other potential polar adapted functions

In addition to the above statistical analyses based on the temperature and latitude optima, we performed an enrichment analysis of KOs by examining their presence in Polar and Nonpolar genomes at different evolutionary scales to capture a variety of situations in the phylogenetic distributions of KOs. Specifically, this analysis was performed at four different lineage levels (i.e., root, main group, family, and genus). The analysis revealed 265 functions that were significantly enriched in Polar genomes inside at least one lineage (Fisher's exact test, $P < 0.05$; Supplementary Table 4). These KOs enriched in Polar viral genomes showed lower temperature optima than other KOs (Supplementary Fig. 9c; Wilcoxon rank-sum test, $P < 0.01$). For a finer-grained observation, we focused on one *Mesomimiviridae* clade, containing a

similar number of Polar (n = 32) and Nonpolar (n = 40) genomes scattered in a subtree of the phylogenomic tree. In this example, four functions were found in more than five genomes from different Polar clades (Fig. 4a). Three of them (K01627, K00979, K06041) co-occurred in the same genomes and formed a near-complete CMP−KDO biosynthesis module in the lipopolysaccharide biosynthesis pathway (Fig. 4b). Lipopolysaccharides are the main component of the Gram-negative bacterial outer membrane, and enzymes of CMP−KDO biosynthesis were found in the genome of Cafeteria roenbergensis virus[60]. This result suggests that the genomes in the examined Polar clade have adapted to the polar environment by coating virions with bacteria-like glycoconjugate to enhance their interactions with Polar hosts.

The KO system can annotate only functionally known genes, and therefore we calculated robust temperature and latitude optima for gene cluster communities, *de novo* clusters of viral genes[28]. The result indicated a slightly higher proportion of Polar-specific gene clusters (26.43%) than obtained by KO annotations (19.74%) (Fig. 4c; Supplementary Fig. 10a), indicating the presence of genes of unknown function that show Polar-specific distributions. We also found that Polar genomes have a slightly but significantly higher proportion of Alanine-rich low-complexity regions than Nonpolar and Unknown genomes (Supplementary Fig. 9d; Dunn's test, $P < 0.05$, following a significant Kruskal-Wallis test, $P = 0.0002$). These low-complexity sequences potentially have an anti-freeze function, as alanine-rich helical structure is one of the significant characteristics of type I antifreeze proteins for ice growth inhibition[61]. Additionally, the proportion of Polar viral genomes that encoded antifreeze protein homologs (n = 7, 1.6%) was higher than the genomes of other groups (n = 6, 0.65%), although the difference was not statistically significant ($P > 0.05$).

Polar-specific functions in microbial-eukaryotes

Finally, to examine whether genomic adaptation occurs in eukaryotic plankton in polar regions and to test if the adaptation is related to the one in viruses, we calculated the temperature and latitude optima for KOs (n = 11,988) assigned to genes in eukaryotic genomes. A similar pattern of Polar and Nonpolar KO groups was identified, although the proportion of the Polar KO group (n = 523, 4.36%) was much smaller than that for viruses (19.74%) (Fig. 4c; Supplementary Fig. 10b). Interestingly, of the 523 KOs in the eukaryotic Polar group, only four were found in the viral Polar group. These were PPM family protein phosphatase, L-galactose dehydrogenase, transcription factor S, and ATP-dependent DNA helicase DinG. This result indicates that most Polar viral functions do not exhibit the same temperature/latitude optima seen in eukaryotic genomes. The result further suggests that virus–host horizontal gene transfer is not the primary driver of viral polar adaptation, and that genomic adaptations are uncoupled between viruses and eukaryotes.

Conclusions

Functional repertoire is considered an important trait for the adaptation of organisms. Previous discoveries of functionally related genes in viruses[55,62] indicated that functional repertoire could also be important for adaptive evolution of viruses. However, this has rarely been addressed for large and giant DNA viruses at a wide geographic scale as compared with cellular organisms. Thanks to the recent progress in metagenomics, we investigated the links between the biogeography, host types, and gene repertoire of viruses infecting marine eukaryotes. We confirmed the existence of a strong polar/nonpolar barrier for these viruses and revealed size fraction-dependent Arctic diversity hotspots for some virus groups, which may reflect a high diversity of their hosts in cold environments. Temperature was an important factor that shaped the virus–host interactions of polar environments. Consistent with these findings, our analyses suggested a presently unidentified virus–host relationship between polar

297 diatoms and giant viruses. Our phylogenomic tree and ancestral state reconstruction revealed
298 back-and-forth adaptations between lower- and higher-temperature niches that occurred
299 recurrently throughout the long evolutionary course of these viruses. Numerous functions,
300 especially ones related to host interactions, were found to be specific to viral polar adaptation,
301 but most of them were not identified as polar-specific functions in eukaryotes. Furthermore,
302 the gene repertoire of these large DNA viral genomes appears more evolutionarily flexible and
303 responsive to temperature change than that of eukaryotic genomes. The discovery of this
304 difference in gene repertoire between polar and nonpolar viruses infecting marine eukaryotes
305 prompts concern about the influence of climate warming on the marine ecosystem, given the
306 importance of these viruses in regulating their host communities and biogeochemical cycling.
307

Methods

Global Ocean Eukaryotic Viral (GOEV) database

Metagenomic datasets and environmental data are provided in Supplementary Table 1. The Global Ocean Eukaryotic Viral (GOEV) database contained 1,817 viral genomes[28,30,63]. Taxonomic inference, read mapping, gene call and gene annotation of the GOEV were performed in a previous work[28]. 1380 detected viruses were classified into six main taxonomic groups: five orders (i.e., *Algavirales*, *Asfuvirales*, *Imitervirales*, *Pandoravirales*, and *Pimascovirales*) and the newly discovered phylum, *Mirusviricota*. Six different size fractions were used in this study: 0.22–1.6 µm or 0.22–3.0 µm ("Pico"), 0.8–5 µm ("Piconano"), 5–20 µm ("Nano"), 20–200 µm ("Micro"), 200–2,000 µm ("Macro"), and 0.8–2,000 µm ("Broad"). The size fraction below 0.22 µm was excluded in this study because of the low relative abundance and high overlap with species from the Pico size fraction. Mean coverage of these viruses was transformed into RPKM (Reads Per Kilobase of exon per Million mapped reads) using the formula: numReads / (genomeLength/1000 * totalNumReads/1,000,000). RPKM profile was used for the ecological analyses in this study.

Phylogenetic tree construction

Phylogenetic trees used in this study were reconstructed using IQ-TREE v.1.6.2[64]. The viral species tree was reconstructed with the site-specific frequency PMSF model following a best-fitting model according to the BIC from the ModelFinder Plus option. The PolB tree was of *Nucleocytoviricota* reference genomes and *Chaetoceros* genomes was reconstructed with the LG+F+I+G4 model. Tree structure manipulation and analysis were done using ETE3 toolkit v.3.1.1[65]. iTOL v.6 was used to visualize the phylogenetic trees[66]. Phylogenetic diversity was calculated using the 'pd' function in the R package '`picante`'[67] for polar and nonpolar genome subsets.

Ecological analyses

Diversity analyses were performed using R v.4.0.1[68] in Rstudio v.1.3.959[69]. To evaluate the diversity of each sample, the richness (number of MAGs), Shannon's index and Pielou's evenness were calculated with the package '`vegan`'[70]. Compositional variation among samples was assessed with a non-metric multidimensional scaling (NMDS) ordination based on Bray-Curtis dissimilarity. Samples with low viral abundance and richness produce outliers that reduce the readability of the NMDS ordination plot. To avoid such a bias, samples for which the sum of cumulative coverage was less than 10 or richness was less than 5 (set as the cutoff threshold) were removed from the compositional variation analyses. Statistical significance of differences among the sample groups (size fractions and biomes) was tested using an ANOSIM (analysis of similarities) with 9,999 permutations. The significance threshold was set to a *p*-value of 0.01. The plots and maps of sampling stations were generated by packages '`ggplot2`'[71] and '`rgdal`'[72].

Gene annotation and clustering

Genes were predicted using Prodigal v.2.6.3[73] within anvi'o v6.1[74] with the default parameters. Gene cluster communities were classified through the AGNOSTOS[75] workflow. Those two steps were performed and described in a previous work[28]. For functional annotation, genes were assigned to KEGG Orthologs (KOs) using eggNOG-mapper v.2.1.5[76] ("Diamond" with an E-value cut-off of 1.0×10^{-5}). Viral marker genes were searched with in-house HMM profiles from NCVOG (nucleocytoplasmic virus orthologous genes)[77] and GVOG (giant virus orthologous groups)[22] databases using HMMER v.3.2.1 (http://hmmer.org) with an E-value of 1×10^{-3}. [78]Antifreeze proteins were detected using InterProScan v.5.44-79.0[78]. Low-complexity

357 regions of protein sequences were identified using the option '-qmask seg' in usearch
358 v.11.0.667[79].
359
360 **Virus–plankton interaction network**
361 We determined the relative abundance matrix for the virus MAGs from the Pico size
362 fractions and relative abundance matrices for eukaryotic MAGs from five cellular size fractions
363 (Piconano, Nano, Micro, Macro, and Broad). To create the input files for network inference,
364 we combined the viral matrix with each of the eukaryotic matrices (corresponding to different
365 size fractions), and only the samples represented by both viral and eukaryotic MAGs were
366 placed in new files. Relative abundances in the newly-generated matrices were normalized
367 using centred log-ratio (*clr*) transformation after adding a pseudo-count of one to all matrix
368 elements because zero cannot be transformed in *clr*. Normalization and filtering were
369 separately applied to viral and eukaryotic MAGs.We then removed the MAGs that had fewer
370 than three sample observations. Network inference was performed using FlashWeave
371 v.0.15.0[80] with Sensitive mode to set a threshold of $\alpha < 0.01$ as the statistical significance and
372 without the default normalization step. All detected pairwise associations were then assigned
373 a weight that ranged between −1 and +1. The network was visualized with Cytoscape v.3.7.1[81]
374 using the prefuse force-directed layout. Proteins between linked genome pairs were aligned
375 using BlastP in Diamond v.2.0.6[82] with an E-value cut-off of 1.0×10^{-50}.
376
377 **Host prediction**
378 First, we pooled network associations from five size fractions by keeping the best positive
379 or negative associations (i.e., the edges with the highest absolute weights). We used a
380 phylogeny-guided filtering approach, Taxon Interaction Mapper (TIM)[37], to predict the host
381 using the global nucleocytoplasmic large DNA virus (NCLDV)–eukaryote network. TIM

provides a list of nodes in the viral tree and associated NCBI taxonomies (order, class, and phylum) of eukaryotes that show significant enrichment in the leaves under the nodes. All the virus–eukaryote associations were mapped on the viral phylogenetic tree to calculate the significance of the enrichment of specific associations using TIM, and the result was visualized with iTOL v.6.

Endogenous viral signals

We searched the viral signals in 713 genomes from the eukaryotic environmental genomes database using VirSorter2 v.2.2.3[29,83]. Both NCLDV and *Lavidaviridae* (virophage) genomic insertions (or co-binning) were searched using --min-score 0.85 and 0.95 for NCLDV and virophage, respectively. We next obtained long-read assembled genomes of two *Chaetoceros* isolates, *C. muelleri* [52]and *C. tenuissimus* [51]. Giant Endogenous Viral Elements (GEVEs) were detected using ViralRecall v.2.1 (-s 5 -w 10)[84]. *Nucleocytoviricota* DNApolB sequences in *Chaetoceros* genomes were detected using HMMER v.3.2.1 search against an in-house DNApolB database. *Chaetoceros*-originating DNApolB sequences were manually concatenated if they were in the same contig and had continuous gene IDs. *Chaetoceros*-originating DNApolBs were aligned with other reference NCLDV DNApolBs using MAFFT-linsi v.7.453[85], and the phylogenetic tree was constructed using IQ-TREE as described above[64].

Size index

Each *Tara* Oceans metagenome corresponds to a specific filtering size fraction (Pico, Piconano, Nano, Micro, Macro, and Broad size fractions as defined above), which were sorted as a list by increasing size. An index constant was set for each size fraction from small to large: Pico = 1, Piconano & Broad = 2, Nano = 3, Micro = 4, Macro = 5 (the Broad and Piconano size fractions were merged because of their similar relative abundances and lack of Arctic

samples for the Piconano fraction). We calculated the size index for a given genome by first multiplying the RPKM of the genome in a sample by the corresponding index constant, then dividing the sum of the products by the overall sum of the RPKMs of the genomes from all samples.

Biome and size niche

Each sample was associated with one specific marine biome (Coastal, Trades, Westerlies, or Polar). To investigate the difference between polar and nonpolar regions, we pooled Coastal, Trades, and Westerlies samples as "Nonpolar". First, we assigned each genome to Polar or Nonpolar if a genome was exclusive to either nonpolar or polar biomes. Additionally, on the basis of RPKM profiles, we calculated the significance using the Wilcoxon rank-sum test. Adjustments for multiple testing were performed using the Benjamini-Hochberg (BH). The significance threshold was set to a corrected *P*-value of 0.05. Similar assignments were performed for two size fractions: intercellular (Pico-size) and intracellular (Piconano, Nano, Micro, Macro, and Broad).

Robust ecological optimum and tolerance

We calculated the robust ecological optimum for a genome (or a gene), which reflects the optimal living condition regarding a given environmental parameter and a tolerance range around this optimum defined by lower and upper bounds[32,33]. For each genome (or a gene), we computed the proportion of RPKM in a given sample relative to the sum of RPKM over all samples. We then used these proportions to populate a weighted vector of a fixed size (n = 10,000) with environmental values accordingly. The ecological optimum is then defined as the median value (Q2) of this vector, and the tolerance (niche) range is given by the interquartile range (Q3 to Q1; some environmental parameter values were missing [nonavailable (NA)] for

some samples). To avoid inferring spurious ecological optima and tolerance ranges for genomes (or genes) for which there were many missing values, we set a minimum threshold of 10 observations with non-NAs and a minimum fraction of 30% non-NA values.

Ancestral states estimation and Relative Evolution Divergency

Ancestral states of Nonpolar and Polar viruses were estimated using the function "ace" (Ancestral Character Estimation) in the R package 'ape'[86]. The input files were a rooted phylogenetic tree based on the four-hallmark gene set described above. In the tree, we retained only viruses with biome assignments of Polar or Nonpolar, and excluded viruses with "Unknown" biomes. We used type = "discrete", method = "ML", and model = "ER" (one-parameter equal rates model). The ancestral states were analysed based on a series of likelihood values for Polar and Nonpolar. Relative Evolutionary Divergence (RED) values were calculated using the "get_reds" function in the package "castor"[87].

KO enrichment in Polar viral genomes

"Polar", "Nonpolar", or "Unknown" biome niche was assigned to each viral genome as described previously. For individual lineages at four taxonomic levels (root, main group, family, and genus), the enrichment of a given KO in Polar genomes assessed using Fisher's exact test in SciPy v.1.7.1[88]. Adjustments for multiple testing were performed using the Benjamini-Hochberg (BH). The significance threshold was set to a corrected *P*-value of 0.05.

Phylogenetic signal of functions

We hypothesized that the phylogenetic distributions of some polar specific functions (i.e., "trait distribution") could be better explained in part by environment selection rather than only by speciation history. We therefore compared two models, (i) the Brownian motion

model (Pagel's lambda = 1, used as the null hypothesis in which the distribution of a trait is simply explained by species tree) and (ii) the Lambda model (0 ≤ Pagel's lambda ≤ 1; lambda = 0 corresponds to the lack of phylogenetic signal in the distribution of a trait), by the likelihood ratio test using the function "fitContinuous" in an R package "geiger"[89]. The *p*-values to reject the null hypothesis were calculated by assuming chi-squared distribution with 1 d.f. for the likelihood-ratio test statistic and adjusted using the BH procedure. The threshold was set to a corrected *p*-value of 0.05

Acknowledgements

This work was supported by JSPS/KAKENHI (18H02279 and 22H00384, to H. O.), and the Collaborative International Joint Research Program of the Institute for Chemical Research, Kyoto University (No. 2021-29, 2022-26 to T.O.D.; No. 2022-27, to S. C.), and the H2020 European Commission project AtlantECO (award number 862923, to S. C.). Computational time was provided by the SuperComputer System, Institute for Chemical Research, Kyoto University. We further thank the *Tara* Oceans consortium, and the people and sponsors who supported *Tara* Oceans. *Tara* Oceans (including 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 XXX of *Tara* Oceans. We thank Gabe Yedid, PhD, from Edanz (http://jp.edanz.com) for editing a draft of this manuscript.

Author contributions

L. M. and H. O. designed the study. L. M. performed the primary biogeographical analysis. T.O.D completed the genome-resolved metagenomic analysis. M. G. performed phylogenomic

482 analyses. E. P. generated the reads mapping data. A. F-G provided de novo clusters of viral
483 genes. R.Y.N, J. W, H. K. contributed to the bioinformatics analysis. All the authors
484 contributed to interpreting the data and writing the manuscript.

485

486 Competing interest statement

487 The authors declare no competing interests.

References

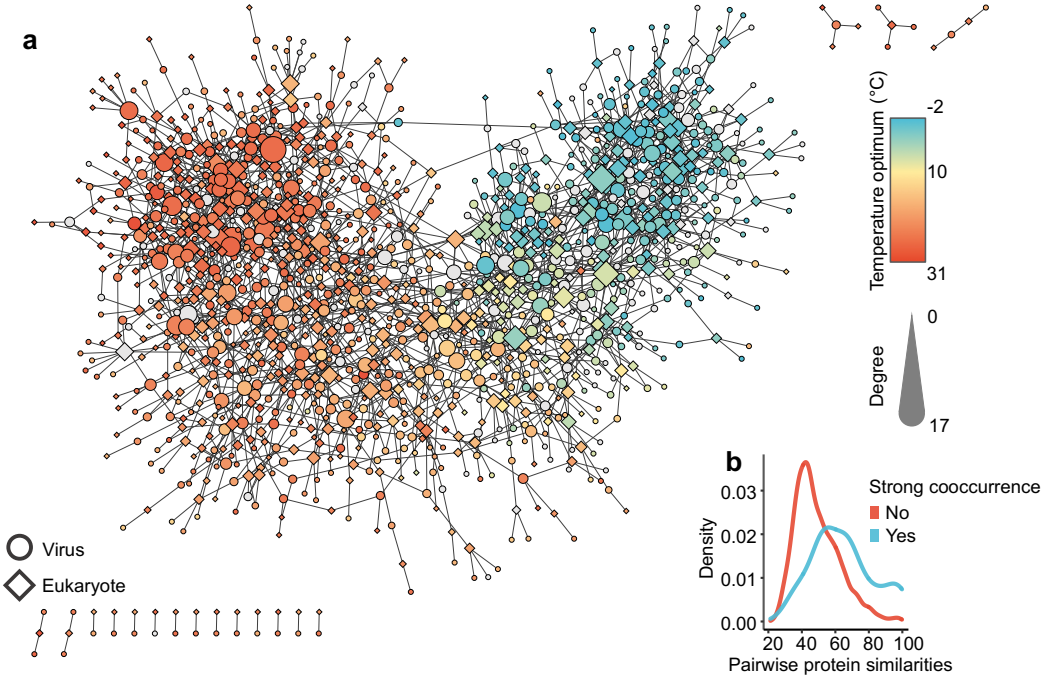
1. Alfredo, J., Castruita, S. & Westbury, M. V. Analyses of key genes involved in Arctic adaptation in polar bears suggest selection on both standing variation and de novo mutations played an important role. *BMC Genomics* **0**, 1–8 (2020).
2. DeVries, A. L. & Cheng, C. H. C. Antifreeze Proteins and Organismal Freezing Avoidance in Polar Fishes. *Fish Physiology* **22**, 155–201 (2005).
3. Kim, B. M. *et al.* Antarctic blackfin icefish genome reveals adaptations to extreme environments. *Nat Ecol Evol* **3**, 469–478 (2019).
4. Methé, B. A. *et al.* The psychrophilic lifestyle as revealed by the genome sequence of Colwellia psychrerythraea 34H through genomic and proteomic analyses. *Proc Natl Acad Sci U S A* **102**, 10913–10918 (2005).
5. Hassan, N. *et al.* Temperature Driven Membrane Lipid Adaptation in Glacial Psychrophilic Bacteria. *Front Microbiol* **11**, 1–10 (2020).
6. Suttle, C. A. Marine viruses - Major players in the global ecosystem. *Nat Rev Microbiol* **5**, 801–812 (2007).
7. Wilhelm, S. W. & Suttle, C. a. Viruses and Nutrient Cycles in the Sea aquatic food webs. *Bioscience* **49**, 781–788 (1999).
8. Guidi, L. *et al.* Plankton networks driving carbon export in the oligotrophic ocean. *Nature* **532**, 465–470 (2016).
9. Irwin, N. A. T., Pittis, A. A., Richards, T. A. & Keeling, P. J. Systematic evaluation of horizontal gene transfer between eukaryotes and viruses. *Nat Microbiol* **7**, 327–336 (2022).
10. Proctor, L. M. & Fuhrman, J. A. Viral mortality of marine bacteria and cyanobacteria. *Nature* **343**, 59–597 (1990).
11. Gregory, A. C. *et al.* Marine DNA Viral Macro- and Microdiversity from Pole to Pole. *Cell* **177**, 1109-1123.e14 (2019).
12. Xia, J. *et al.* Tight association between microbial eukaryote and giant virus communities in the Arctic Ocean. *Limnol Oceanogr* **67**, 1343–1356 (2022).
13. López-Bueno, A. *et al.* High diversity of the viral community from an Antarctic lake. *Science (1979)* **326**, 858–861 (2009).
14. Yau, S. *et al.* Virophage control of antarctic algal host-virus dynamics. *Proc Natl Acad Sci U S A* **108**, 6163–6168 (2011).
15. Demory, D. *et al.* A thermal trade-off between viral production and degradation drives virus-phytoplankton population dynamics. *Ecol Lett* **24**, 1133–1144 (2021).
16. Demory, D. *et al.* Temperature is a key factor in Micromonas-virus interactions. *ISME Journal* **11**, 601–612 (2017).
17. Ogunbunmi, E. T. *et al.* Low-Temperature Adaptation Targets Genome Packing Reactions in an Icosahedral Single-Stranded DNA Virus. *J Virol* **96**, (2022).
18. Endo, H. *et al.* Biogeography of marine giant viruses reveals their interplay with eukaryotes and ecological functions. *Nat Ecol Evol* **4**, 1639–1649 (2020).
19. Koonin, E. v. & Yutin, N. *Evolution of the Large Nucleocytoplasmic DNA Viruses of Eukaryotes and Convergent Origins of Viral Gigantism*. *Advances in Virus Research* vol. 103 (Elsevier Inc., 2019).
20. Koonin, E. v. & Yutin, N. Origin and evolution of eukaryotic large nucleo-cytoplasmic DNA viruses. *Intervirology* **53**, 284–292 (2010).
21. Guglielmini, J., Woo, A. C., Krupovic, M., Forterre, P. & Gaia, M. Diversification of giant and large eukaryotic dsDNA viruses predated the origin of modern eukaryotes. *Proc Natl Acad Sci U S A* **116**, 19585–19592 (2019).

22. Aylward, F. O., Moniruzzaman, M., Ha, A. D. & Koonin, E. v. A phylogenomic framework for charting the diversity and evolution of giant viruses. *PLoS Biol* **19**, 1–18 (2021).
23. Mihara, T. *et al.* Taxon richness of "Megaviridae" exceeds those of bacteria and archaea in the ocean. *Microbes Environ* **33**, 162–171 (2018).
24. Monier, A., Claverie, J. M. & Ogata, H. Taxonomic distribution of large DNA viruses in the sea. *Genome Biol* **9**, R106 (2008).
25. Hingamp, P. *et al.* Exploring nucleo-cytoplasmic large DNA viruses in Tara Oceans microbial metagenomes. *ISME Journal* **7**, 1678–1695 (2013).
26. Carradec, Q. *et al.* A global ocean atlas of eukaryotic genes. *Nat Commun* **9**, 373 (2018).
27. Schulz, F. *et al.* Giant viruses with an expanded complement of translation system components. **85**, 82–85 (2017).
28. Gaïa, M., Meng, L., Pelletier, E., Forterre, P. & Vanni, C. Plankton-infecting relatives of herpesviruses clarify the evolutionary trajectory of giant viruses. (2022).
29. Delmont, T. O. *et al.* Functional repertoire convergence of distantly related eukaryotic plankton lineages abundant in the sunlit ocean. *Cell Genomics* **2**, 100123 (2022).
30. Schulz, F. *et al.* Giant virus diversity and host interactions through global metagenomics. *Nature* **578**, 432–436 (2020).
31. Moniruzzaman, M., Weinheimer, A. R., Martinez-Gutierrez, C. A. & Aylward, F. O. Widespread endogenization of giant viruses shapes genomes of green algae. *Nature* **588**, 141–145 (2020).
32. Chaffron, S. *et al.* Environmental vulnerability of the global ocean epipelagic plankton community interactome. *Sci Adv* **7**, 1–15 (2021).
33. Cristóbal, E., Ayuso, S. V., Justel, A. & Toro, M. Robust optima and tolerance ranges of biological indicators: A new method to identify sentinels of global warming. *Ecol Res* **29**, 55–68 (2014).
34. Hillebrand, H. On the generality of the latitudinal diversity gradient. *American Naturalist* **163**, 192–211 (2004).
35. Ibarbalz, F. M. *et al.* Global Trends in Marine Plankton Diversity across Kingdoms of Life. *Cell* **179**, 1084-1097.e21 (2019).
36. Dominguez-Huerta, G. *et al.* Diversity and ecological footprint of Global Ocean RNA viruses. *Science (1979)* **376**, 1202–1208 (2022).
37. Kaneko, H. *et al.* Eukaryotic virus composition can predict the efficiency of carbon export in the global ocean. *iScience* **24**, 102002 (2021).
38. Meng, L. *et al.* Quantitative Assessment of Nucleocytoplasmic Large DNA Virus and Host Interactions Predicted by Co-occurrence Analyses. *mSphere* **6**, (2021).
39. Jacobsen, A., Bratbak, G. & Heldal, M. Isolation and characterization of a virus infecting Phaeocystis pouchetii (Prymnesiophyceae). *J Phycol* **32**, 923–927 (1996).
40. Baudoux, A. C. & Brussaard, C. P. D. Characterization of different viruses infecting the marine harmful algal bloom species Phaeocystis globosa. *Virology* **341**, 80–90 (2005).
41. Blanc-Mathieu, R. *et al.* A Persistent Giant Algal Virus, with a Unique Morphology, Encodes an Unprecedented Number of Genes Involved in Energy Metabolism. *J Virol* **95**, 1–23 (2021).
42. Moniruzzaman, M. *et al.* Genome of brown tide virus (AaV), the little giant of the Megaviridae, elucidates NCLDV genome expansion and host-virus coevolution. *Virology* **466–467**, 60–70 (2014).
43. Downes Gastrichl, M. *et al. Aureococcus Regional Brown Tide anophagefferens Blooms*. vol. 27 (2004).

44. 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**, 1–9 (1991).
45. Clerissi, C., Desdevises, Y. & Grimsley, N. Prasinoviruses of the Marine Green Alga Ostreococcus tauri Are Mainly Species Specific. *J Virol* **86**, 4611–4619 (2012).
46. Moniruzzaman, M., Weinheimer, A. R., Martinez-Gutierrez, C. A. & Aylward, F. O. Widespread endogenization of giant viruses shapes genomes of green algae. *Nature* **588**, 141–145 (2020).
47. Maumus, F. & Blanc, G. Study of gene trafficking between acanthamoeba and giant viruses suggests an undiscovered family of amoeba-infecting viruses. *Genome Biol Evol* **8**, 3351–3363 (2016).
48. Armaleo, D. *et al.* The lichen symbiosis re-viewed through the genomes of Cladonia grayi and its algal partner Asterochloris glomerata. *BMC Genomics* **20**, 1–33 (2019).
49. Irwin, N. A. T., Pittis, A. A., Richards, T. A. & Keeling, P. J. Systematic evaluation of horizontal gene transfer between eukaryotes and viruses. *Nat Microbiol* **7**, 327–336 (2022).
50. Coy, S. R., Gann, E. R., Pound, H. L., Short, S. M. & Wilhelm, S. W. Viruses of eukaryotic algae: Diversity, methods for detection, and future directions. *Viruses* **10**, (2018).
51. Nelson, D. R. *et al.* Large-scale genome sequencing reveals the driving forces of viruses in microalgal evolution. *Cell Host Microbe* **29**, 250-266.e8 (2021).
52. Hongo, Y. *et al.* The genome of the diatom Chaetoceros tenuissimus carries an ancient integrated fragment of an extant virus. *Sci Rep* **11**, 1–13 (2021).
53. Malviya, S. *et al.* Insights into global diatom distribution and diversity in the world's ocean. *Proc Natl Acad Sci U S A* **113**, E1516–E1525 (2016).
54. Nef, C., Madoui, M., Pelletier, É. & Bowler, C. Whole-genome scanning reveals selection mechanisms in epipelagic Chaetoceros diatom populations. 1–39 (2022).
55. Rosenwasser, S. *et al.* Rewiring host lipid metabolism by large viruses determines the fate of Emiliania huxleyi, a bloom-forming alga in the ocean. *Plant Cell* **26**, 2689–2707 (2014).
56. Trimble, J. J., Murthy, S. C. S., Bakker, A., Grassmann, R. & Desrosiers, R. C. A gene for dihydrofolate reductase in a herpesvirus. *Science (1979)* **239**, 1145–1147 (1988).
57. Yamamoto-Kawai, M., Carmack, E. & McLaughlin, F. Nitrogen balance and Arctic throughflow. *Nature* **443**, 43 (2006).
58. Wirsen, C. O., Jannasch, H. W., Wakeham, S. G. & Canuel, E. A. Membrane lipids of a psychrophilic and barophilic deep-sea bacterium. *Curr Microbiol* **14**, 319–322 (1986).
59. Li, Y. *et al.* The Importance of Glycans of Viral and Host Proteins in Enveloped Virus Infection. *Front Immunol* **12**, 1–12 (2021).
60. 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 (2010).
61. Davies, P. L., Roach, A. H. & Hew, C. L. DNA sequence coding for an antifreeze protein precursor from winter flounder. *Proc Natl Acad Sci U S A* **79**, 335—339 (1982).
62. 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 (2019).

63. 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**, 1–11 (2020).
64. Minh, B. Q. *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol Biol Evol* **37**, 1530–1534 (2020).
65. Huerta-Cepas, J., Serra, F. & Bork, P. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Mol Biol Evol* **33**, 1635–1638 (2016).
66. Letunic, I. & Bork, P. Interactive tree of life (iTOL) v5: An online tool for phylogenetic tree display and annotation. *Nucleic Acids Res* **49**, W293–W296 (2021).
67. Kembel, S. W. *et al.* Picante: R tools for integrating phylogenies and ecology. *Bioinformatics* **26**, 1463–1464 (2010).
68. R Core Team. A Language and Environment for Statistical Computing (R Foundation for Statistical Computing, 2021). (2021).
69. Team, Rs. RStudio: Integrated Development for R. RStudio, PBC, Boston, MA. *http://www.rstudio.com/* (2020).
70. Oksanen, J. *et al.* vegan: Community Ecology Package. R package. version 2.5-3. *https://CRAN.R-project.org/package=vegan* (2018).
71. Wickham, H. ggplot2: Elegant Graphics for Data Analysis. *Springer-Verlag New York* **ISBN 978-3**, (2016).
72. Bivand, R. *et al.* rgdal: bindings for the geospatial data abstraction library 2017. *URL https://CRAN. R-project. org/package= rgdal. R package version* 1 (2018).
73. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 1–11 (2010).
74. Eren, A. M. *et al.* Anvi'o: An advanced analysis and visualization platformfor 'omics data. *PeerJ* **2015**, 1–29 (2015).
75. Vanni, C. *et al.* Unifying the known and unknown microbial coding sequence space. *Elife* **11**, 1–60 (2022).
76. Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Mol Biol Evol* **38**, 5825–5829 (2021).
77. Yutin, N., Wolf, Y. I., Raoult, D. & Koonin, E. v. Eukaryotic large nucleo-cytoplasmic DNA viruses: Clusters of orthologous genes and reconstruction of viral genome evolution. *Virol J* **6**, 1–13 (2009).
78. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
79. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461 (2010).
80. Tackmann, J., Matias Rodrigues, J. F. & von Mering, C. Rapid Inference of Direct Interactions in Large-Scale Ecological Networks from Heterogeneous Microbial Sequencing Data. *Cell Syst* **9**, 286-296.e8 (2019).
81. Paul Shannon *et al.* Cytoscape: A Software Environment for Integrated Models. *Genome Res* **13**, 426 (2003).
82. Buchfink, B., Reuter, K. & Drost, H.-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat Methods* **18**, 366–368 (2021).
83. Guo, J. *et al.* VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. *Microbiome* **9**, 1–13 (2021).
84. Aylward, F. O. & Moniruzzaman, M. ViralRecall-A Flexible Command-Line Tool for the Detection of Giant Virus Signatures in 'Omic Data. *Viruses* **13**, 15–17 (2021).

682 85. Katoh, K., Misawa, K., Kuma, K. I. & Miyata, T. MAFFT: A novel method for rapid
683 multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* **30**,
684 3059–3066 (2002).
685 86. Paradis, E. & Schliep, K. ape 5.0: an environment for modern phylogenetics and
686 evolutionary analyses in R. *Bioinformatics* **35**, 526–528 (2019).
687 87. Louca, S. & Doebeli, M. Efficient comparative phylogenetics on large trees.
688 *Bioinformatics* **34**, 1053–1055 (2018).
689 88. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in
690 Python. *Nat Methods* **17**, 261–272 (2020).
691 89. Pennell, M. W. *et al.* geiger v2. 0: an expanded suite of methods for fitting
692 macroevolutionary models to phylogenetic trees. *Bioinformatics* **30**, 2216–2218
693 (2014).
694
695



696
697 **Fig. 1 | Virus–plankton interaction network. a,** Five individual networks inferred using input
698 matrices for the relative frequencies of eukaryotes (five size fractions) and giant viruses (Pico-
699 size fraction). The best positive or negative association (i.e., the edges with the highest absolute
700 weights between two genomes) were selected to build the integrated network. Node colour
701 represents the temperature optima of each genome for viruses and eukaryotes. A total of 1,347
702 nodes (567 eukaryotes and 780 viruses) are in the network. Of these nodes, 1,191 nodes (554
703 eukaryotes and 637 viruses) are coloured according to their temperature optima. **b,** The
704 distribution of pairwise sequence similarity of proteins (one protein from the eukaryotic
705 genome and one from the viral genome). Blue line indicates the distribution for pairs with a
706 strong virus–eukaryote association in the network (edge weight of ≥ 0.4), while the red line is
707 for pairs lacking a strong association. The two distributions are significantly different (P =
708 1×10^{-13}, Wilcoxon signed-rank test).
709

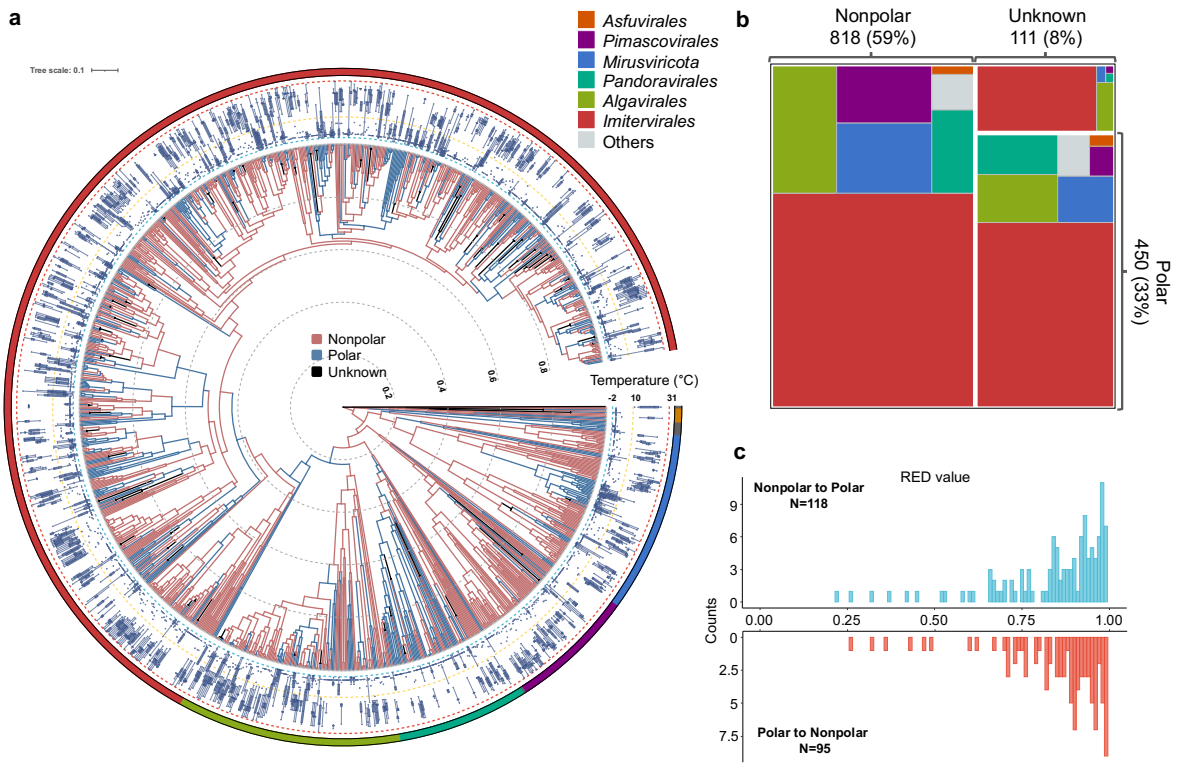


Fig. 2 | Inferred ancestral polar and non-polar niches for viruses. a, Ancestral "Polar" and "Nonpolar" states were estimated using the phylogenetic tree based on a one-parameter equal rates model. The outermost layer shows the taxonomy of six main groups. The boxplots in the second layer show the temperature optima of the viral genomes. Only polar and nonpolar genomes were included in the tree. **b,** The treemap diagram shows the number of viruses assigned to Polar, Nonpolar or "Unknown" biomes. Colours indicate the main taxonomic groups. **c,** Histograms of Relative Evolutionary Divergence (RED) values for the nodes at which "polar" or "nonpolar" adaptation events were inferred.

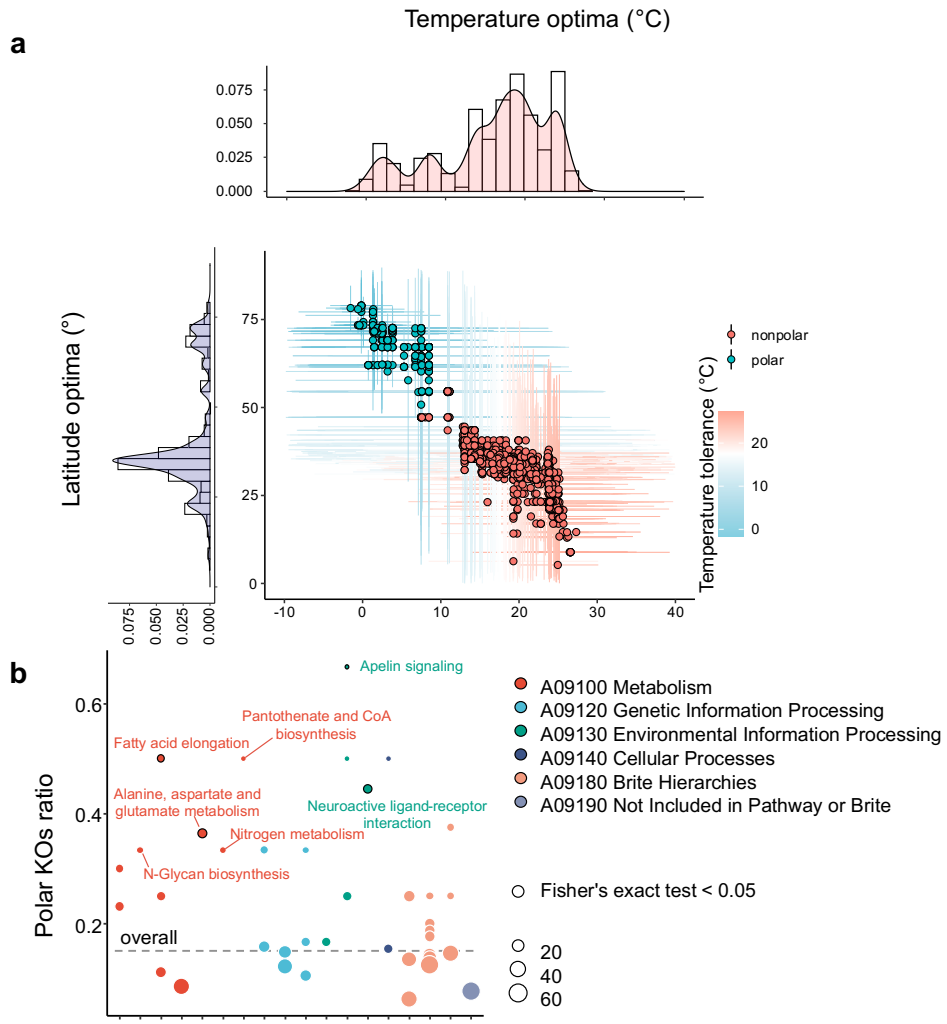


Fig. 3 | Ecological niche of KEGG Orthologs (KOs) and polar-enriched pathways. a, Distribution of the temperature optima and latitude optima for KEGG Orthologs (KOs) found in viral genomes. Colours of dots represent the Polar or Nonpolar niche for each KO. Bars indicate the tolerance ranges of temperature (horizontal) and latitude (vertical). Histograms show the distributions of temperature and latitude optima. **b,** Ratio of Polar KOs in each pathway. Black-framed circles correspond to pathways in which Polar KOs were significantly enriched ($P < 0.05$, Fisher's exact test). The overall ratio of Polar KOs to all KOs is indicated by a dotted line.

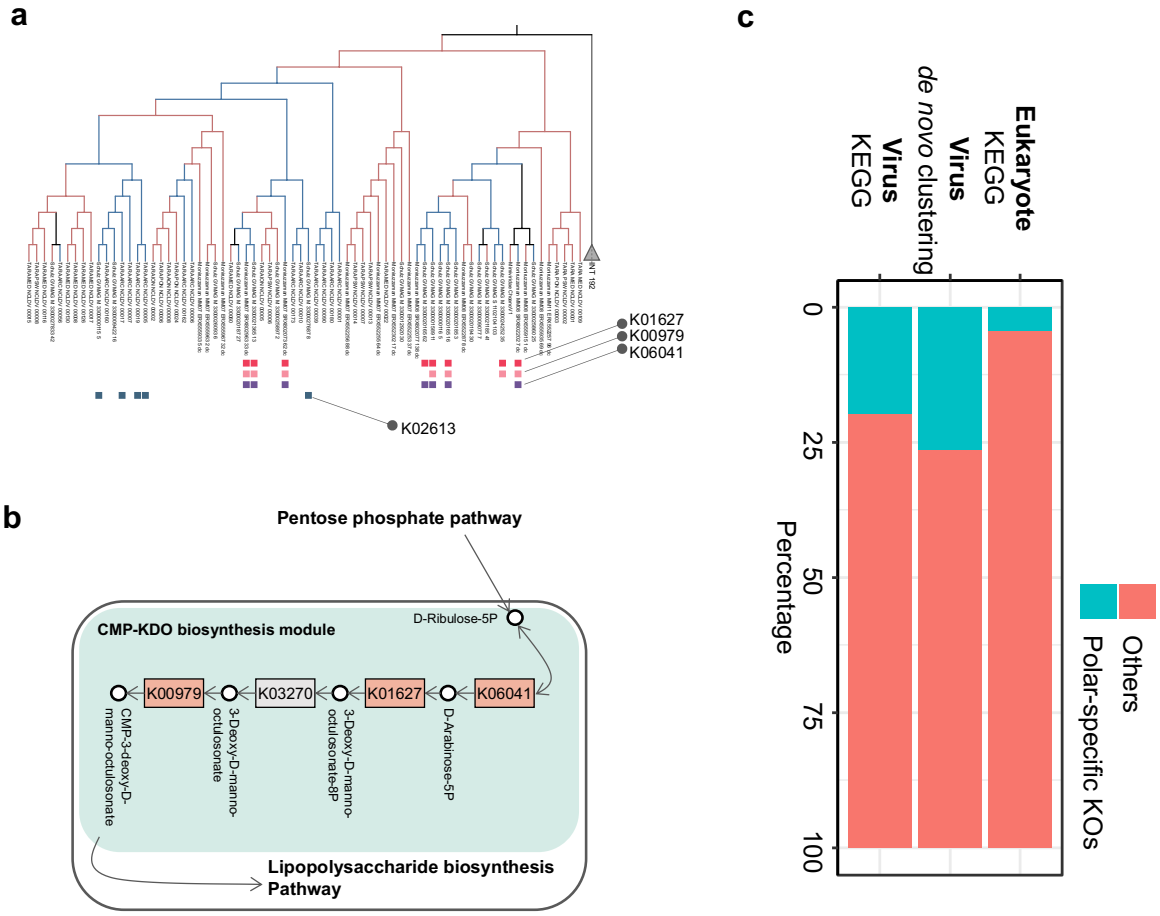


Fig. 4 | Independent genomic adaptation of giant viruses. 244 functions (KOs) were enriched at individual lineages. One example was given in **a,** Four KOs that were present exclusively in more than five Polar genomes in a selected *Mesomimiviridae* clade. Three of them (K01627, K00979, K06041) were encoded in the same genomes and formed a near-complete CMP−KDO biosynthesis module shown in **b,** Schematic of the three Polar enzymatic steps in the CMP–KDO biosynthesis module. **c,** Proportion of Polar and Nonpolar specific functions (KOs and GCCs) in viruses and eukaryotes.