Assessing the biogeography of marine giant viruses in four oceanic transects

Anh D. Ha[1], Mohammad Moniruzzaman[2], Frank O. Aylward[1,3*]
[1] Department of Biological Sciences, Virginia Tech, Blacksburg VA, 24061
[2] Rosenstiel School of Marine Atmospheric, and Earth Science, University of Miami, Coral Gables FL 33149
[3] Center for Emerging, Zoonotic, and Arthropod-Borne Infectious Disease, Virginia Tech, Blacksburg VA, 24061

***Correspondence:**
Frank O. Aylward
faylward@vt.edu

Running title: Biogeography of giant viruses in oceanic transects

Keywords: Giant Viruses, Nucleocytoviricota, Mesomimiviridae, Prasinoviridae, marine viruses.

Abstract

Viruses of the phylum *Nucleocytoviricota* are ubiquitous in ocean waters and play important roles in shaping the dynamics of marine ecosystems. In this study, we leveraged the bioGEOTRACES metagenomic dataset collected across the Atlantic and Pacific Oceans to investigate the biogeography of these viruses in marine environments. We identified 330 viral genomes, including 212 in the order *Imitervirales* and 54 in the order *Algavirales*. We found that most viruses appeared to be prevalent in shallow waters (<150 meters), and that viruses of the *Mesomimiviridae* (*Imitervirales)* and *Prasinoviridae* (*Algavirales)* are by far the most abundant and diverse groups in our survey. Five mesomimiviruses and one prasinovirus are particularly widespread in oligotrophic waters; annotation of these genomes revealed common stress response systems, photosynthesis-associated genes, and oxidative stress modulation that may be key to their broad distribution in the pelagic ocean. We identified a latitudinal pattern in viral diversity in one cruise that traversed the North and South Atlantic Ocean, with viral diversity

peaking at high latitudes of the northern hemisphere. Community analyses revealed three distinct *Nucleocytoviricota* communities across latitudes, categorized by latitudinal distance towards the equator. Our results contribute to the understanding of the biogeography of these viruses in marine systems.

Introduction

Large DNA viruses of the phylum *Nucleocytoviricota*, also known as “giant viruses”, are a diverse group of eukaryotic viruses with particle sizes typically larger than 0.2 μm in diameter and genome sizes reaching up to 2.5 Mbp [1–4]. Known members of the phylum are partitioned into six orders, namely *Algalvirales, Imitevirales, Pimascovirales, Pandoravirales, Asfuvirales,* and *Chitovirales*, and up to 32 potential families [5]. These viruses have an ancient origin and have underwent frequent gene exchange with their hosts [6, 7], and as a result their genomes often encode numerous genes involved in cellular processes such as glycolysis, the TCA cycle, amino acid metabolism, translation, light sensing, and cytoskeletal dynamics [8–15]. Giant viruses are known to infect a broad spectrum of eukaryotic hosts; while members of the *Imitervirales*, *Algavirales,* and *Pandoravirales* infect a wide range of algae and various heterotrophic protists, members of the *Asfuvirales, Chitovirales,* and *Pimascovirales* infect a wide range of protist and metazoan hosts [2, 16–19]. The diverse functional repertoires harbored in these viruses’ genomes are thought to render them able to manipulate the physiology and subvert the immune responses of their hosts during infection [9, 20, 21]. Giant viruses therefore play important roles in various ecological processes across the globe.

Although giant viruses are ubiquitous in the biosphere, they appear to be particularly abundant and diverse in marine environments. Early studies focusing on amplification and sequencing of the viral Family B DNA polymerase from seawater found that algal viruses within the

Nucleocytoviricota were widespread in a variety of marine environments [22, 23], an observation which was later confirmed through analysis of community metagenomic data [24, 25]. A recent comparative metagenomic study found that giant viruses are ubiquitous in the ocean, vary markedly across depth, and are prevalent in >0.22 um size fractions [26]. Field studies have estimated that the abundance of giant viruses can reach up to 10^4 -10^6 viruses per milliliter of seawater, with higher abundances typically recovered during algal blooms [27–32]. Giant viruses have been reported to infect many prevalent marine eukaryotic lineages, including chlorophytes, haptophytes, and choanoflagellates [15, 18, 33, 34], and they are therefore an important factor shaping marine ecological dynamics. Moreover, several studies have shown that giant viruses associated with algal blooms play key roles in carbon export to deeper waters [35–37], indicating they are critical components of global carbon cycles. Despite the ecological importance of giant viruses, our understanding of their diversity lags behind that of smaller viruses owing to the widespread use of filtration steps in viral diversity surveys, which often exclude larger viruses [38]. There is therefore a strong need for further studies to examine the biogeography and ecological dynamics of these large viruses in the ocean.

Here, we aimed to undertake a genome-based global survey of giant virus assemblages across the oceans and compare the diversity of these viruses in different geographic regions by leveraging the large number of metagenome-assembled genomes (MAGs) of these viruses that have been generated [9, 39–41]. We focused on the metagenomic data generated from samples collected in four transects in the Atlantic and Pacific oceans as part of the the GEOTRACES project [42], which provide clear, well-defined geographic and depth profiles. These metagenomes targeted the >0.2 um size fraction and are therefore suitable for examining the diversity of large viruses. Our work broadens our understanding of the biogeography of giant viruses in the ocean and reveals novel diversity patterns that will be important for understanding their role in marine environments.

Results and Discussion

Imitervirales and Algavirales orders are the most abundant and diverse among viruses recovered in bioGEOTRACES metagenomic data set

We examined 480 metagenomes collected and sequenced as part of the bioGEOTRACES component of the GEOTRACES project [42]. These samples were derived from four major cruises from different regions of the South Pacific and Atlantic Oceans (Fig. 1) in 2010-2011. This dataset targeted the >0.2-µm size fraction of microbial communities and was sampled along well-defined transects at various depths, making it suitable for assessing the geographic and depth distribution of marine giant viruses. In total, we identified 330 giant virus genomes with metagenomic reads mapping. To investigate the taxonomic distribution of the detected *Nucleocytoviricota* viruses, we constructed a phylogenetic tree of the viruses and 1,188 *Nucleocytoviricota* reference genomes (Fig. S1; see Methods). Out of the 330 genomes recovered, we were able to place the genomes within the orders *Imitervirales* [n=214], *Algavirales* [n=54], *Pimascovirales* [n=16], *Pandoravirales* [n=4], and *Asfuvirales* [n=1]. On the family level, the most well-represented groups were the *Mesomimiviridae* [n=146] and *Prasinoviridae* [n=42] (Fig. 2). We also identified 41 *Mirusviricota* viruses; although technically not members of the *Nucleocytoviricota*, these large DNA viruses are prevalent in the ocean and share many genomic features with giant viruses [41]. Of the 330 genomes identified, only 8 were derived from cultivated viruses, and 322 genomes are metagenome-assembled genomes (MAGs). Approximately half (159) of the 330 genomes are larger than 300 kbp (Table S1), underscoring the prevalence of viruses with large genomes throughout the ocean.

Of the 330 genomes with reads mapping, 14 viruses, including 8 *Algavirales* and 5 *Imitervirales* were recovered in all four bioGEOTRACES transects. Meanwhile, 182 viruses were found in

only one transect, among which 116 genomes were found exclusively in transect GA02, most of which were *Imitervirales* (Fig. 3A, Table S1). In term of total giant virus richness, the number of different genomes found in transect GA02, which traces along the Americas-Atlantic Ocean coastline (248 genomes total) by far exceeded that in the other three transects, especially compared to the pelagic transect GP13 where less than one-third of that number (71 genomes) were detected. This is likely a consequence of the much broader range of latitudes and biogeochemical regimes sampled by the GA02 transect compared to the others. Across the transects, 65 viruses were present in all three depth layers of the water columns (<80m, 80-150m, and >150m), most of which were *Imitervirales* and *Algavirales* (Fig. 3B). We found 103 viruses that solely appeared in the surface water of 80 meters up, while there were only 5 genomes unique to the deep water of below 150 meters.

Giant virus communities were mostly dominated by members of the *Imitervirales* and *Algavirales* orders, regardless of the transect location or depth of sampling (Fig. 4). On average, *Imitervirales* and *Algavirales* accounted for 56.4% and 32.6% of the total number of giant virus occurrences across all sampling locations, respectively. This result is consistent with previous observations that viruses of these two orders were the most abundant and widespread giant viruses in the Pacific and Atlantic Ocean [25, 26, 28]. Viruses within the *Imitervirales* were particularly widespread in communities across all depths sampled in the pelagic GP13 transect, with a mean contribution of 88.8%, and the majority (9) of the 11 viruses found exclusively in GP13 were *Imitervirales* (Fig. 2A; Fig. 4C). This pattern of *Imitervirales* dominance in pelagic waters was also observed in transect GA03 (Fig. 4B, inner samples), underscoring the prevalence of this group in oligotrophic gyres. In general, the spatial distribution of viruses in the ocean is shaped largely by the geographic distribution of their hosts [43], and the broad distribution of viruses within the *Imitervirales* is therefore likely a signature of their collective broad host range. Indeed, members of the *Imitervirales* order are known to infect an

exceptionally broad phylogenetic range of hosts [19], including marine haptophytes in the genera *Phaeocystis* and *Chrysochromulina,* which were found in high abundance in the open waters of the central Pacific Ocean [44, 45], as well as other widespread hosts such as the green algae*,* Choanoflagellates, and amoeboid protists [15, 46–49]. Aside from these established hosts, recent work using co-occurrence analyses have also identified a wide range of other potential eukaryotic hosts for viruses within the *Imitervirales* [12, 50, 51], suggesting that the hosts of viruses in this order is far broader than currently known. Lastly, given that many members of the *Imitervirales* gain entry to host cells through phagocytosis, it is likely that individual viral populations may infect a range of different host lineages. If this is the case, the broad representation of the *Imitervirales* in pelagic surface waters may represent the ability of viruses in this order to infect a range of mixotrophic and heterotrophic lineages.

The majority of viruses that are most widespread in the Atlantic Ocean in our survey (i.e., found in all three Atlantic transects GA02, GA03, and GA10, but not in the Pacific transect GP13) were *Algavirales* viruses (Fig. 3A). Viruses within the *Algavirales* order were especially abundant in the North Atlantic Ocean (northern samples in transect GA02) and in the coastal waters (eastern samples in transect GA10) where we observed a high abundance throughout the water column (Fig. 4A, 4D; Fig. 5C). Viruses of this order were also present in high abundance in surface waters (<50m) and euphotic waters at 50-150m deep on both two coastal sides of the transect GA03, while showing a sharp decreasing trend towards the pelagic waters of the North Atlantic and the Pacific Ocean (inners of transect GA03 and transect GP13, respectively). Once again, these changes in the abundance of viruses within the *Algavirales* likely reflect the distribution of their hosts. Members of the family *Prasinoviridae* are the most prevalent family we identified within the *Algavirales*, and members of this group are known to infect members of the prasinophyte genera *Ostreococcus, Bathycoccus*, and *Micromonas* [18], which have been found to be highly abundant in coastal systems. It was estimated that prasinophytes may account for

50-90% of total picoeukaryotic cells in coastal waters, while they only made up a much lower fraction (<20%) of those in pelagic waters [52]. Although not as abundant as their counterparts in coastal populations, there are several prasinophytes found widespread in oligotrophic waters of the open ocean, for instance *O. lucimarinus* and *Micromonas spp.*, allowing the broad presence of viruses infecting these hosts, which have been detected and documented [53, 54].

Mirusviruses, *Pimascovirales, Pandoravirales,* and *Asfuvirales* were present to a lesser degree in the four transects, with average contributions of 9.1%, 1.5%, 0.3%, and 0.03% of giant virus occurrence, respectively. Mirusviruses were prevalent across all four transects (Fig. 4), consistent with findings of a previous investigation using metagenomic read recruitments from Tara Oceans datasets [41]. These viruses appeared to be more abundant in coastal waters, which was maintained throughout the water column (transect GA02 and east of transect GA10), while in pelagic waters their abundance was more limited to sunlit waters at <100m (inners of transect GA03 and transect GP13, respectively) (Fig. S2A). This pattern is in agreement with the prediction that Mirusviruses infect a broad planktonic host range that includes many phototrophs [41]. All of the *Pimascovirales* genomes recovered in our survey were MAGs derived from marine metagenomes. Although currently little is known about the natural hosts of this viral group in the oceans, their prevalence across the four transects suggests that they infect widespread host taxa and play important roles in marine systems. Interestingly, we found that the recently-delineated family-level clade PM_01 was the most prevalent lineage of the *Pimascovirales* in the ocean, but no members of this group have been cultivated and their host range remains unknown. A recent study found a member of this lineage was prevalent in surface waters of Station ALOHA, consistent with the view that they are present in oligotrophic surface waters [55]. The only *Asfuvirales* virus found in our survey, GVMAG-M-3300027833-19 was recovered in the GA10 transect, which is located off the Atlantic coast of South Africa. This viral genome was also recently recovered in a TARA metagenomic sample in the same region

[17] and its transcriptomic activities were detected in the waters of the central California Current upwelling system in the North Pacific Ocean [12], suggesting that the virus may be widely distributed beyond the sampling scope of the bioGEOTRACES cruises. *Pandoravirales* viruses were present in all four transects in both coastal and pelagic waters, although at relatively low number of occurrences and abundance. Across all locations included in our survey, their distribution was strictly limited to shallower waters (<150 meters) (Fig. S2C). The *Pandoravirales* includes well-studied coccolithoviruses that are prevalent in *Emiliania huxleyi* blooms, suggesting that at least some members of this order will have highly variable abundance depending on host availability. An important caveat of our study is that we surveyed only metagenomes derived from >0.2 um size fractions; it is likely that some members of the *Pimascovirales* and *Asfuvirales*, which have on average smaller genome and virion sizes than members of the *Imitervirales*, may be more widespread in smaller size fractions, and are therefore more prevalent than indicated by our results here.

Giant viruses were apparently more diverse and abundant in surface waters (Fig. 3B, Fig. 4). Indeed, of all occurrences of giant viruses throughout the water column at all locations, 92.1% were located in waters at 150 meters or shallower. On average across all four transects, the viral richness at two depth ranges (<80m and 80-150m) were approximately 2.7 and 2.3 times higher than that at deep waters (>= 150m), respectively (Kruskal–Wallis and Dunn's test, P<0.001) (Fig. S3). In transects GA03 and GP13, the decline in total giant virus abundance with depth was particularly sharp (Fig. 5A). In deeper pelagic water (>200m), the giant communities appeared to be limited to just a few members of the *Imitervirales* (inners of transect GA03 and transect GP13), while in coastal waters at deeper than 200m, members of the *Algavirales* are also present with relatively high abundance, together with *Imitervirales* viruses (Fig. 4, Fig. 5). Apart from viruses of the *Algavirales* and *Imitervirales* orders, our read mapping approach did not recover any giant virus MAGs of any other orders in metagenomes sequenced from water

sampled from depths >200m. This may be partially due to biases in the reference database, which includes more genomes from surface water samples, but it seems likely that it is at least partially driven by the large diversity of giant viruses in surface waters.

Latitudinal pattern of giant virus diversity

We analyzed the giant virus communities along the transect GA02 in more detail to assess possible latitudinal gradients in giant virus diversity. The GA02 transect sampling sites follows the Americas-Atlantic Ocean coastline, spanning across a long range of latitudes from the parallel 50° North to 50° South and tracing a clear latitudinal gradient from the North Atlantic in the summer of 2010 to the south Atlantic in the austral summer of 2011. This sampling scheme may facilitate the detection of subtle latitudinal gradients that may be more difficult to resolve through comparison of samples collected across different ocean basins. In terms of community composition at each sampling location, the *Algavirales* assemblages seemingly dominated the giant virus communities in northern samples and decreased in abundance towards the south, replaced by the dominance of viruses of the order *Imitervirales* (Fig. S4).

We calculated taxonomic richness and Shannon's H diversity index in each depth-integrated sampling location to investigate latitudinal variation (Fig. 6). To avoid biases in diversity measurements due to unequal sequencing depth, we rarefied all metagenomic samples to 10M reads prior to calculation. We detected a latitudinal pattern of diversity along the GA02 transect with average diversity increasing with higher latitudes in the Northern Hemisphere and plateaued towards the South. Total giant virus communities peaked, both in terms of richness and alpha diversity in the further north of the North Atlantic Ocean (i.e. above 40° North) and steeply declined around the middle latitudes (20-40° North). The trend of increasing diversity from the equatorial zone towards higher latitudes was mirrored in the *Imitervirales* communities,

while varying marginally in the *Algavirales* communities for both viral genome richness and alpha diversity (Fig. 6). The clear peak in latitudinal diversity in the Northern Hemisphere is consistent with the trend of species richness observed for a large portion of the total of 65,000 marine species examined previously [56]. It is possible that stronger environmental instability, particularly the wide temperature variation in the northern hemisphere (excluding polar zones) [57] may explain the higher diversity compared to the south. A relatively similar northern spike of giant virus diversity has been reported from analyses of the Family B DNA Polymerase (PolB) genes in Tara Oceans datasets [26, 58], although the studies observed another increase in diversity near the southern middle latitudes. The discrepancy did not seem to result from the disparity in methodological approaches between our mapping strategy and the above two studies; we also performed calculation of diversity indices on the TARA Ocean datasets using our mapping method described herein, and observed a similar trend in giant virus diversity agreeing with in the two PolB studies in latitudinal locations of elevated diversity (Fig. S5). It is possible that the slightly differing results reflects the fact that the bioGEOTRACES and Tara Oceans samples were collected from different times with different cruise tracks.

The lower diversity, in terms of alpha diversity and richness of the giant virus communities near the equator in comparison with northern high latitudes did now follow conventional latitude diversity pattern, which posits that marine eukaryotic diversity generally increases toward the tropics [58, 59]. The difference could possibly be attributed to several potential causes. High temperature in the equatorial zone is potentially one underlying cause; decline in species diversity at higher temperatures has been observed in several marine taxa [60]. Previous work has widely reported that marine species tend to shift away from the tropics to higher latitudes due to climate warming [61–63]. A species distribution model has projected that marine species may acquire latitudinal peaks in total richness at approximately 40/30° absolute latitude and a loss of richness near the equator, highlighting the important role of temperature on these

distribution shifts [64]. Indeed, the metagenomic samples located in the GA02 transect included in our survey were collected during the months from late March to June, which were anticipated to be the warmest months of the year in the equatorial Atlantic Ocean [65]. Given the strong influence of seasonality on marine microbial communities, it is likely that latitudinal gradients in diversity are ephemeral and will vary throughout the year.

A non-metric multidimensional scaling (NMDS) analysis indicated that *Nucleocytoviricota* communities were clustered according to their latitudinal distance to the equator (Fig. 7). The giant virus community composition significantly differed between three latitudinal sectors in both the GA02 transect exclusively and all transects collectively (Permanova $p < 0.001$). This clustering is fairly consistent with traditional Longhurst oceanographic biogeographical biomes of plankton ecology, which were designated based on the distribution of chlorophyll, angle of sunlight, and cloudiness [66]. This may further support the view that the geographic distribution of viruses in ocean waters is mostly affected by the distribution of their hosts. In terms of viral community richness, we found only 5 giant virus genomes (2% of the total number of giant viruses found in the GA02 transect) shared across all three latitudinal zones (Fig. S6), all of which belonged to the order *Algavirales*. The high latitude zone (>40° latitude) harbored 93 unique genomes, while the mid-latitude zone (20° to 40° latitude) had 68 and the low latitude zone (between 20° equatorial) had 24.

Five Mesomimiviruses *and one* prasinovirus *are particularly widespread in oligotrophic waters*

The vast majority of the most abundant and widespread viruses in our survey belong to the orders *Imitervirales* and *Algavirales*. We observed six MAGs within the *Imitervirales* (5 genomes) and the *Algavirales* (1 genome) that were particularly widespread in oligotrophic waters (Fig. 8A), all detected across different water depths in at least 19 distinct sampling

locations. All the five *Imitervirales* could be classified into the recently-proposed *Mesomimiviridae* family (*Imitervirales* family 1) (Table S1). The *Mesomimiviridae* family is particularly widespread in marine systems and contains well-documented cultivated representatives that infect oceanic haptophytes, such as *Phaeocystis globosa* virus (PgV), *Chrysochromulina ericina* virus (CeV), and *Chrysochromulina parva* virus (CpV) [67–69]. Other members of the family have been found co-occurring and correlating with diatoms [50], suggesting that diatoms are potential hosts of this viral lineage. The only *Algavirales* virus belonged to the *Prasinoviridae* family (*Algavirales* family 1). The most broadly distributed genome, ERX556088.18.dc was recovered in 122 samples (more than 25% of the total number of samples analyzed overall) at 34 sampling locations. All the five Mesomimiviruses were extensively distributed in the Pacific transect GP13, while the prasinovirus, TARA_IOS_NCLDV_00011, was more widespread in the Atlantic Ocean (Fig. S7). All of these six genomes derived from marine environments, with genome sizes ranging from 108,412 bp to 483,524 bp and GC content varied from 26.2% to 34.6%.

Annotation of these genomes showed complex genomic repertoires, which is a common characteristic of viruses of the *Mesomimiviridae* and the *Prasinoviridae*. Complete or near-complete set of 9 giant virus core genes, including major capsid protein (MCP), A32-like packaging ATPase (A32), superfamily II helicase (SFII), family B DNA Polymerase (PolB), virus late transcription factor 3 (VLTF3), large and small RNA polymerase subunits (RNAPL and RNAPS, respectively), TFIIB transcriptional factor (TFIIB), and Topoisomerase family II (TopoII) were found in all of the *Mesomimiviridae* genomes, indicating that these are high quality genome assemblies (Fig. 8B). These core genes are broadly represented in genomes of *Nucleocytoviricota* and have previously been used as phylogenetic markers for these viruses [5, 7, 70]. Both RNAP subunits were absent in the *Prasinoviridae* genome, consistent with the lack of DNA-dependent RNA polymerase that has been previously reported for prasinoviruses [11].

Other genes encoding essential viral functions were also consistently found in these genomes, including ribonucleotide reductase, thymidylate synthase, dUTPase (for nucleotide metabolism), Nudix-like hydrolase, mRNA capping enzyme (transcription and RNA processing), and glycosyltransferase (virion morphogenesis) (Table S2).

Genes involved in translation have been widely reported in the genomes of viruses within the order *Imitervirales* [71–73]. We found several translation-related genes, including aminoacyl-tRNA synthetases, or aaRS (asparaginyl-tRNA synthetase), translation initiation factors (IF4E, eIF3, IF1A), translation elongation factors (eF-TU) in all of the *Mesomimiviridae* genomes. The aaRS genes catalyzes the linkage between tRNAs and amino acids during translation and may act as a mechanism for circumventing nutrient starvation in the host cell, allowing the virus to maintain viral replication in different nutritional conditions [74].

Throughout all six genomes, we also identified numerous genes involved in diverse metabolic processes (Fig. 8B, Table S2), which may be involved in rewiring host metabolism and cellular physiology during infection to support their own viral production. We found genes involved in central carbon metabolism, including enzymes for glycolysis, the TCA cycle, and beta oxidation in all of the genomes. Numerous genes involved in nutrient acquisition and processing, light-driven energy generation, and diverse transporters were also present, consistent with previous findings [9, 39]. Rhodopsins could potentially alter the host's sunlight-dependent energy transfer system [15], while chlorophyll a/b binding proteins might help maintain a stable light-harvesting capacity of host cells during infection [9]. The presence of genes involved in photosynthetic processes might be important for these viruses to infect a wide array of phototrophic or mixotrophic hosts in well-lit waters across the ocean. Genes encoding storage proteins and transporters, including ferritin-like proteins, amino acid permeases, transporters predicted to target sulfur, phosphorus, and iron are common in these genomes and may have a role in rewiring host's nutrient acquisition strategies to enhance viral propagation. Such set of viral-

encoded nutrient storage and transporters might be especially advantageous in marine environments, particularly in the oligotrophic waters of the South Pacific Ocean, where micronutrients such as iron are scarce [75] and the viruses need to employ their own transporters to boost nutrient acquisition. We also found homologs of genes involved in the regulation of cellular apoptosis, including caspase and tumor necrosis factor receptor. Manipulation of cell death is a common strategy employed by giant viruses to avert the impending cellular response to viral infection [76–78].

A broad array of stress response and repair genes found in all of the six genomes potentially equips the viruses with the ability to endure various external stresses common in oligotrophic waters, such as high temperatures, ultraviolet (UV) damage, and oxidative stress. We found genes involved in oxidative stress regulation, including thioredoxin, glutaredoxin, and superoxide dismutase (SOD) to be common among all genomes. Thioredoxin and SOD have been found expressed in several members of the *Imitervirales* [9, 79, 80] and were suggested to mitigate cellular oxidative stress by detoxifying harmful reactive oxygen species released by hosts during viral infection. SOD may also play an active role in reducing superoxide accumulation induced by UV exposures in direct sunlight, which may aid survival of viruses in the sunlit open waters [81]. It has been postulated that such viral-encoded redox genes allow the virus to infect a broad range of hosts [80]. Previous work has noted that giant viruses may carry genes that serve their own DNA repair to maintain high fidelity in genome replication [70, 82, 83]. We identified various DNA repair genes, including MutS mismatch repair and ultraviolet (UV) damage repair, such as ERCC4 nuclease [84] to be present in all of the mesomimivirus genomes. MutS homologs are widely present in genomes of mimivirus relatives [67, 85, 86] and are thought to associate with correcting mismatches to ensure the fidelity of viral genome replication. ERCC4-type repair nuclease might provide the viruses with crucial protection against DNA damages caused by UV irradiation. Although the prasinovirus' genome lack

homologs of MutS and ERCC4-type repair nuclease, it encodes numerous other putative DNA repair genes such as phosphatidylinositol 3-kinase and PD-(D/E)XK nuclease superfamily, which could also potentially aid in maintaining DNA integrity.

We also identified various genes predicted to encode enzymes for synthesizing glycans, which may be involved in the decoration of capsids with sugar moieties. These viral-encoded fibril structures are potentially useful for viruses to extend their host range and persist in the open waters. First, the oligosaccharides may enable the modification of virion surface to mimic the host's normal food source, e.g. organic debris and bacteria [87, 88], promoting phagocytosis of virion particles. This strategy of infection, which takes advantage of the 'generalized' feeding habit that many marine protists rely on, may obviate the requirement of building receptors to a specific host and thus allow for a broader array of hosts. In addition, glycosylated fibrils could possibly act as a protective layer to shield the viruses from unfavorable environmental conditions, therefore increasing viral persistence. Furthermore, a study of Acanthamoeba polyphaga mimivirus has found that viral particles covered with self-produced sugars are able to adhere to different organisms through glycoside interactions, including bacteria, fungi, and arthropods [89], without infecting them. These organisms thus may help disperse the viruses over a wide area of waters, increasing their chance of contact with drifted host cells and expanding spatial distribution across the ocean. We also observed that all of these viruses carry lectin-domain containing proteins, which may act as key mediators of host-virus recognitions and interactions [90, 91]. Although the exact role of the protein in viruses is still unclear, it is possible that they might leverage lectin domains to modulate interactions with hosts and achieve a broader host range.

Conclusion

In this study we conducted a metagenomic survey of giant viruses in the Atlantic and Pacific Oceans using the bioGEOTRACES datasets. We show that giant viruses of the orders *Imitervirales* and *Algavirales* are particularly widespread and abundant in epipelagic waters. Giant virus communities vary markedly by latitude, and in the GA02 transect in the Atlantic Ocean we detected a latitudinal pattern of diversity that peaks at high northern latitudes and plateaus towards the south. Lastly, we identified five genomes of the *Mesomimiviridae* family of the *Imitervirales* and one genome of the *Prasinoviridae* of the *Algavirales* that are particularly widespread in oligotrophic waters. Our comparative genomic analysis revealed that these genomes encoded diverse genes involved in central carbon metabolism, stress responses, and lectin-domain proteins potentially involved in host-virus interactions. We hypothesize that these genes may collectively expand the host range of these viruses, possibly explaining their particularly broad distribution. Overall, our study provides genomic insights into the distribution of giant viruses in the ocean and sheds light on the biogeography of these ecologically-important community members.

Materials and Methods

Nucleocytoviricota genome database compilation

We downloaded 1,382 *Nucleocytoviricota* genomes from the Giant Virus Database [5] and 696 viral MAGs assembled from 937 Tara Oceans metagenomes within the Global Ocean Eukaryotic Viral database [41]. All of these genomes were classified to the phylum *Nucleocytoviricota*, except those of the recently-discovered *Mirusviricota* lineage, which has a herpesvirus-like capsid and likely belongs to the realm Duplodnaviria. Although Mirusviruses represent a lineage distinct from the *Nucleocytoviricota*, we included them here because they represent a widespread lineage of marine large DNA viruses, and their genomes appear to be a

chimera of different viral lineages, including the *Nucleocytoviricota*. To remove possible contamination from cellular sources, we screened all viral genomes using ViralRecall [92] and removed all contigs that had a score < 0 (indicating stronger signals from cellular sources). We also excluded genomes of less than 100 kbp total sequence, not encoding PolB gene, and/or containing less than 2 out of 4 of the marker genes SFII, TFIIB, VLTF3, and A32. To avoid the presence of identical or highly similar genomes, we dereplicated the genome set with dRep v3.2.2 [93] using an average nucleotide identity threshold of 95%. We arrived at a database containing 1,629 viral genomes (1,518 *Nucleoviricota* and 111 *Mirusviricota*) for metagenomic read mapping.

Metagenome data set

We examined the metagenomic data from the >0.2-μm size fraction microbial communities of 480 samples collected by the international GEOTRACES program from May 2010 to December 2011. Accession numbers of the data are listed in Table S3. The samples were collected in four major cruise transects (GA02, GA03, GA10, and GP13) across the Atlantic and Pacific Oceans at 2-10 depths in each sampling location, ranging from 6m to 5601m. Sample processing was previously described in detail [42]. We calculated the geographical distance between sample locations in each transect based on recorded latitudes and longitudes using the function distHaversine from the R package geosphere.

Reads processing and mapping

We downloaded and trimmed reads from each of the metagenome samples with Trim Galore v. 0.6.4 using parameters "–length 50 -e 0.1 -q 5 –stringency 1 --phred33". We then mapped the trimmed reads onto the *Nucleocytoviricota* nucleotide sequences using coverM v0.6.1 (https://github.com/wwood/CoverM) in mode 'genome', with the parameter --min-read-percent-identity 0.95. We calculated relative abundance in reads mapped per kilobase of genome, per

million mapped reads (RPKM). To avoid the false detection of viral genomes due to spurious read mapping, we only retained genomes with breadth coverage >20% (i.e. more than 20% of the genome length were covered by any read) in subsequent analyses. This cutoff is based on recent work which suggested that a genome coverage of at least 20% is appropriate to indicate the presence of that genome in a sample [94]. After this filtering, we obtained a set of 330 *Nucleocytoviricota* genomes for subsequent analysis.

Phylogeny and clade delineation.

To provide phylogenetic context for the giant virus genomes that we identified, we constructed a multilocus phylogenetic tree of the *Imitervirales* order using a set of 7 marker genes: family B DNA Polymerase (PolB), A32-like packaging ATPase (A32), Poxvirus late transcription factor 3 (VLTF3), superfamily II helicase (SFII), alpha RNA polymerase subunits (RNAPL), TFIIB transcriptional factor (TFIIB), and Topoisomerase family II (TopoII). The concatenated alignment of these 7 markers was generated with the ncldv_markersearch.py script (github.com/faylward/ncldv_markersearch) and then trimmed with TrimAl v. 1.4.rev22 [95] (parameter -gt 0.1). The tree was inferred from the alignment using IQ-TREE version 2.2.0.3 [96] with the best fitting model determined by the ModelFinder Plus option in IQ-TREE, according to the Bayesian Information Criterion (BIC). We used the same order-, family-, and genus-level nomenclature for the *Nucleocytoviricota* as previously described [5].

Subsampling reads and calculating diversity

Comparison of diversity among samples, especially alpha diversity, may be erroneous due to differing library sizes [97]. To ensure equal library sizes across samples for diversity measurements, all samples were randomly subsampled without replacement to 10M reads using the reformat program provided in bbtools suite (Bushnell B. – sourceforge.net/projects/bbmap/). The subsampled reads were mapped against the viral

genome set using coverM as described above. We then calculated community richness and Shannon's diversity indices using the package 'vegan' (https://cran.r-project.org/web/packages/vegan/). Variation among community composition was analyzed with NMDS ordination based on Bray–Curtis dissimilarity using the function 'metaMDS', parameters k = 2, trymax = 100. Statistical analyses of difference in community composition were performed using a PERMANOVA test with the 'adonis' function, 9,999 permutations.

Depth distribution mapping and interpolation

We performed interpolation of viral depth distribution using the program Ocean Data View v5.6.0 [98] in DIVA gridding mode, with parameters signal-to-noise = 25, automatic scale lengths for the X- and Y-axis, quality limit = 3 to exclude bad estimates.

Protein annotations

We annotated proteins in six widespread giant virus genomes by comparing them to the EggNOG database 5.0.0 [99] and Pfam-A release 34 [100] hidden Markov models (HMMs) profile using HMMER v3.3.2 (parameter "-E 1e-3" for the EggNOG search and "–cut_nc" for the Pfam search) and retained only the best hits.

Data availability

The data sets analyzed in this study are already publicly available and were accessed as described in the Materials and Methods section.

Acknowledgements

We would like to thank the authors of Biller et al for providing access to the GEOTRACES metagenomes used in this study. We acknowledge the use of the Virginia Tech Advanced Research Computing Center for bioinformatic analyses performed in this study.

Competing interests

The authors declare no competing interests.

Funding

National Science Foundation (CAREER-2141862 to F.O.A.); Simons Early Career Award in Marine Microbial Ecology and Evolution (to F.O.A.), National Institutes of Health (1R35GM147290-01 to F.O.A).

References

1. Koonin EV, Dolja VV, Krupovic M, Varsani A, Wolf YI, Yutin N, et al. Global Organization and Proposed Megataxonomy of the Virus World. *Microbiol Mol Biol Rev* 2020; **84**.
2. Wilhelm S, Bird J, Bonifer K, Calfee B, Chen T, Coy S, et al. A Student's Guide to Giant Viruses Infecting Small Eukaryotes: From Acanthamoeba to Zooxanthellae. *Viruses* . 2017. , **9**: 46
3. Fischer MG. Giant viruses come of age. *Curr Opin Microbiol* 2016; **31**: 50–57.
4. Aylward FO, Moniruzzaman M. Viral Complexity. *Biomolecules* 2022; **12**.
5. Aylward FO, Moniruzzaman M, Ha AD, Koonin EV. A phylogenomic framework for charting the diversity and evolution of giant viruses. *PLoS Biol* 2021; **19**: e3001430.
6. Moniruzzaman M, Weinheimer AR, Martinez-Gutierrez CA, Aylward FO. Widespread endogenization of giant viruses shapes genomes of green algae. *Nature* 2020; **588**: 141–

145.

7. Guglielmini J, Woo AC, 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* 2019; **116**: 19585–19592.

8. Blanc-Mathieu R, Dahle H, Hofgaard A, Brandt D, Ban H, Kalinowski J, et al. A persistent giant algal virus, with a unique morphology, encodes an unprecedented number of genes involved in energy metabolism. *J Virol* 2021.

9. Moniruzzaman M, Martinez-Gutierrez CA, Weinheimer AR, Aylward FO. Dynamic genome evolution and complex virocell metabolism of globally-distributed giant viruses. *Nat Commun* 2020; **11**: 1710.

10. Rodrigues RAL, Arantes TS, Oliveira GP, Dos Santos Silva LK, Abrahão JS. The Complex Nature of Tupanviruses. *Adv Virus Res* 2019; **103**: 135–166.

11. Moreau H, Piganeau G, Desdevises Y, Cooke R, Derelle E, Grimsley N. Marine prasinovirus genomes show low evolutionary divergence and acquisition of protein metabolism genes by horizontal gene transfer. *J Virol* 2010; **84**: 12555–12563.

12. Ha AD, Moniruzzaman M, Aylward FO. High Transcriptional Activity and Diverse Functional Repertoires of Hundreds of Giant Viruses in a Coastal Marine System. *mSystems* 2021; **6**: e0029321.

13. Yutin N, Koonin EV. Proteorhodopsin genes in giant viruses. *Biol Direct* 2012; **7**: 34.

14. Rozenberg A, Oppermann J, Wietek J, Fernandez Lahore RG, Sandaa R-A, Bratbak G, et al. Lateral Gene Transfer of Anion-Conducting Channelrhodopsins between Green Algae and Giant Viruses. *Curr Biol* 2020; **30**: 4910–4920.e5.

15. Needham DM, Yoshizawa S, Hosaka T, Poirier C, Choi CJ, Hehenberger E, et al. A distinct lineage of giant viruses brings a rhodopsin photosystem to unicellular marine predators. *Proceedings of the National Academy of Sciences* . 2019. , **116**: 20574–20583

16. Koonin EV, Yutin N. Evolution of the Large Nucleocytoplasmic DNA Viruses of Eukaryotes

and Convergent Origins of Viral Gigantism. *Adv Virus Res* 2019; **103**: 167–202.

17. Karki S, Moniruzzaman M, Aylward FO. Comparative Genomics and Environmental Distribution of Large dsDNA Viruses in the Family Asfarviridae. *Frontiers in Microbiology* . 2021. , **12**
18. Weynberg KD, Allen MJ, Wilson WH. Marine Prasinoviruses and Their Tiny Plankton Hosts: A Review. *Viruses* 2017; **9**.
19. Claverie J-M, Abergel C. Mimiviridae: An Expanding Family of Highly Diverse Large dsDNA Viruses Infecting a Wide Phylogenetic Range of Aquatic Eukaryotes. *Viruses* . 2018. , **10**: 506
20. Rosenwasser S, Ziv C, van Creveld SG, Vardi A. Virocell Metabolism: Metabolic Innovations During Host–Virus Interactions in the Ocean. *Trends in Microbiology* . 2016. , **24**: 821–832
21. Forterre P. Manipulation of cellular syntheses and the nature of viruses: The virocell concept. *C R Chim* 2011; **14**: 392–399.
22. Chen F, Suttle CA, Short SM. Genetic diversity in marine algal virus communities as revealed by sequence analysis of DNA polymerase genes. *Appl Environ Microbiol* 1996; **62**: 2869–2874.
23. Short SM, Suttle CA. Sequence analysis of marine virus communities reveals that groups of related algal viruses are widely distributed in nature. *Appl Environ Microbiol* 2002; **68**: 1290–1296.
24. Monier A, Larsen JB, Sandaa R-A, Bratbak G, Claverie J-M, Ogata H. Marine mimivirus relatives are probably large algal viruses. *Virol J* 2008; **5**: 12.
25. Monier A, Claverie J-M, Ogata H. Taxonomic distribution of large DNA viruses in the sea. *Genome Biol* 2008; **9**: R106.
26. Endo H, Blanc-Mathieu R, Li Y, Salazar G, Henry N, Labadie K, et al. Biogeography of marine giant viruses reveals their interplay with eukaryotes and ecological functions. *Nat*

Ecol Evol 2020; **4**: 1639–1649.

27. Bellec L, Grimsley N, Derelle E, Moreau H, Desdevises Y. Abundance, spatial distribution and genetic diversity of Ostreococcus tauri viruses in two different environments. *Environ Microbiol Rep* 2010; **2**: 313–321.
28. Hingamp P, Grimsley N, Acinas SG, Clerissi C, Subirana L, Poulain J, et al. Exploring nucleo-cytoplasmic large DNA viruses in Tara Oceans microbial metagenomes. *ISME J* 2013; **7**: 1678–1695.
29. Sorensen G, Baker AC, Hall MJ, Munn CB, Schroeder DC. Novel virus dynamics in an Emiliania huxleyi bloom. *Journal of Plankton Research* . 2009. , **31**: 787–791
30. Wilson WH, Tarran GA, Schroeder D, Cox M, Oke J, Malin G. Isolation of viruses responsible for the demise of an *Emiliania huxleyi* bloom in the English Channel. *Journal of the Marine Biological Association of the United Kingdom* . 2002. , **82**: 369–377
31. Lehahn Y, Koren I, Schatz D, Frada M, Sheyn U, Boss E, et al. Decoupling physical from biological processes to assess the impact of viruses on a mesoscale algal bloom. *Curr Biol* 2014; **24**: 2041–2046.
32. Tarutani K, Nagasaki K, Yamaguchi M. Viral impacts on total abundance and clonal composition of the harmful bloom-forming phytoplankton Heterosigma akashiwo. *Appl Environ Microbiol* 2000; **66**: 4916–4920.
33. Sandaa RA, Heldal M, Castberg T, Thyrhaug R, Bratbak G. Isolation and characterization of two viruses with large genome size infecting Chrysochromulina ericina (Prymnesiophyceae) and Pyramimonas orientalis (Prasinophyceae). *Virology* 2001; **290**.
34. Brussaard CP, Short SM, Frederickson CM, Suttle CA. Isolation and phylogenetic analysis of novel viruses infecting the phytoplankton Phaeocystis globosa (Prymnesiophyceae). *Appl Environ Microbiol* 2004; **70**.
35. Kaneko H, Blanc-Mathieu R, Endo H, Chaffron S, Delmont TO, Gaia M, et al. Eukaryotic virus composition can predict the efficiency of carbon export in the global ocean. *iScience*

2020; **24**.

36. Laber CP, Hunter JE, Carvalho F, Collins JR, Hunter EJ, Schieler BM, et al. Coccolithovirus facilitation of carbon export in the North Atlantic. *Nature microbiology* 2018; **3**.
37. Kavagutti VS, Bulzu P-A, Chiriac CM, Salcher MM, Mukherjee I, Shabarova T, et al. High-resolution metagenomic reconstruction of the freshwater spring bloom. *Microbiome* 2023; **11**: 1–24.
38. Thurber RV, Haynes M, Breitbart M, Wegley L, Rohwer F. Laboratory procedures to generate viral metagenomes. *Nat Protoc* 2009; **4**: 470–483.
39. Schulz F, Roux S, Paez-Espino D, Jungbluth S, Walsh DA, Denef VJ, et al. Giant virus diversity and host interactions through global metagenomics. *Nature* 2020; **578**: 432–436.
40. Bäckström D, Yutin N, Jørgensen SL, Dharamshi J, Homa F, Zaremba-Niedwiedzka K, et al. Virus Genomes from Deep Sea Sediments Expand the Ocean Megavirome and Support Independent Origins of Viral Gigantism. *MBio* 2019; **10**.
41. Gaïa M, Meng L, Pelletier E, Forterre P, Vanni C, Fernandez-Guerra A, et al. Plankton-infecting relatives of herpesviruses clarify the evolutionary trajectory of giant viruses. *bioRxiv* . 2022. , 2021.12.27.474232
42. Biller SJ, Berube PM, Dooley K, Williams M, Satinsky BM, Hackl T, et al. Marine microbial metagenomes sampled across space and time. *Scientific Data* 2018; **5**: 1–7.
43. Chow C-ET, Suttle CA. Biogeography of Viruses in the Sea. *Annu Rev Virol* 2015; **2**: 41–66.
44. Endo H, Ogata H, Suzuki K. Contrasting biogeography and diversity patterns between diatoms and haptophytes in the central Pacific Ocean. *Sci Rep* 2018; **8**: 1–13.
45. Sow SLS, Trull TW, Bodrossy L. Oceanographic Fronts Shape Phaeocystis Assemblages: A High-Resolution 18S rRNA Gene Survey From the Ice-Edge to the Equator of the South Pacific. *Front Microbiol* 2020; **0**.
46. Schvarcz CR, Steward GF. A giant virus infecting green algae encodes key fermentation

genes. *Virology* 2018; **518**: 423–433.

47. Chelkha N, Levasseur A, Pontarotti P, Raoult D, Scola BL, Colson P. A Phylogenomic Study of Acanthamoeba polyphaga Draft Genome Sequences Suggests Genetic Exchanges With Giant Viruses. *Front Microbiol* 2018; **0**.
48. Fischer MG, Allen MJ, Wilson WH, Suttle CA. Giant virus with a remarkable complement of genes infects marine zooplankton. *Proc Natl Acad Sci U S A* 2010; **107**: 19508–19513.
49. Abrahão J, Silva L, Silva LS, Khalil JYB, Rodrigues R, Arantes T, et al. Tailed giant Tupanvirus possesses the most complete translational apparatus of the known virosphere. *Nat Commun* 2018; **9**: 1–12.
50. Meng L, Endo H, Blanc-Mathieu R, Chaffron S, Hernández-Velázquez R, Kaneko H, et al. Quantitative Assessment of Nucleocytoplasmic Large DNA Virus and Host Interactions Predicted by Co-occurrence Analyses. *mSphere* 2021; **6**.
51. Moniruzzaman M, Wurch LL, Alexander H, Dyhrman ST, Gobler CJ, Wilhelm SW. Virus-host relationships of marine single-celled eukaryotes resolved from metatranscriptomics. *Nat Commun* 2017; **8**: 16054.
52. Massana R. Picoeukaryotes. *Encyclopedia of Microbiology (Third Edition)*. 2009. Academic Press, pp 674–688.
53. Derelle E, Monier A, Cooke R, Worden AZ, Grimsley NH, Moreau H. Diversity of Viruses Infecting the Green Microalga Ostreococcus lucimarinus. *J Virol* 2015; **89**: 5812–5821.
54. Bellec L, Grimsley N, Desdevises Y. Isolation of prasinoviruses of the green unicellular algae Ostreococcus spp. on a worldwide geographical scale. *Appl Environ Microbiol* 2010; **76**: 96–101.
55. Farzad R, Ha AD, Aylward FO. Diversity and genomics of giant viruses in the North Pacific Subtropical Gyre. *Front Microbiol* 2022; **13**: 1021923.
56. Chaudhary C, Saeedi H, Costello MJ. Bimodality of Latitudinal Gradients in Marine Species Richness. *Trends Ecol Evol* 2016; **31**: 670–676.

57. Chown SL, Sinclair BJ, Leinaas HP, Gaston KJ. Hemispheric Asymmetries in Biodiversity—A Serious Matter for Ecology. *PLoS Biol* 2004; **2**: e406.

58. Ibarbalz FM, Henry N, Brandão MC, Martini S, Busseni G, Byrne H, Coelho LP, Endo H, Gasol JM, Gregory AC, Mahé F, Rigonato J, Royo-Llonch M, Salazar G, Sanz-Sáez I, Scalco E, Soviadan D, Zayed AA, Zingone A, Labadie K, Ferland J, Marec C, Kandels S, Picheral M, Dimier C, Poulain J, Pisarev S, Carmichael M, Pesant S}, de Vargas C Karp-Boss L Wincker P Lombard F Bowler C Zinger L Tara Oceans Coordinators Babin M Boss E Iudicone D Jaillon O Acinas Sg Ogata H Pelletier E Stemmann L Sullivan Mb Sunagawa S Bopp L. Global Trends in Marine Plankton Diversity across Kingdoms of Life. *Cell* 2019; **179**: 1084–1097.e21.

59. Hillebrand H. Strength, slope and variability of marine latitudinal gradients. *Marine Ecology Progress Series* . 2004. , **273**: 251–267

60. Tittensor DP, Mora C, Jetz W, Lotze HK, Ricard D, Berghe EV, et al. Global patterns and predictors of marine biodiversity across taxa. *Nature* 2010; **466**: 1098–1101.

61. Chen I-C, Hill JK, Ohlemüller R, Roy DB, Thomas CD. Rapid range shifts of species associated with high levels of climate warming. *Science* 2011; **333**: 1024–1026.

62. Poloczanska ES, Brown CJ, Sydeman WJ, Kiessling W, Schoeman DS, Moore PJ, et al. Global imprint of climate change on marine life. *Nat Clim Chang* 2013; **3**: 919–925.

63. Pinsky ML, Worm B, Fogarty MJ, Sarmiento JL, Levin SA. Marine taxa track local climate velocities. *Science* 2013; **341**: 1239–1242.

64. García Molinos J, Halpern BS, Schoeman DS, Brown CJ, Kiessling W, Moore PJ, et al. Climate velocity and the future global redistribution of marine biodiversity. *Nat Clim Chang* 2015; **6**: 83–88.

65. Li T, Philander SGH. On the Seasonal Cycle of the Equatorial Atlantic Ocean. *J Clim* 1997; **10**: 813–817.

66. Longhurst AR. Ecological Geography of the Sea. 2010. Elsevier.

67. Santini S, Jeudy S, Bartoli J, Poirot O, Lescot M, Abergel C, et al. Genome of Phaeocystis globosa virus PgV-16T highlights the common ancestry of the largest known DNA viruses infecting eukaryotes. *Proc Natl Acad Sci U S A* 2013; **110**.

68. Gallot-Lavallée L, Blanc G, Claverie JM. Comparative Genomics of Chrysochromulina Ericina Virus and Other Microalga-Infecting Large DNA Viruses Highlights Their Intricate Evolutionary Relationship with the Established Mimiviridae Family. *J Virol* 2017; **91**.

69. Stough JMA, Yutin N, Chaban YV, Moniruzzaman M, Gann ER, Pound HL, et al. Genome and Environmental Activity of a Chrysochromulina parva Virus and Its Virophages. *Front Microbiol* 2019; **10**.

70. Yutin N, Koonin EV. Hidden evolutionary complexity of Nucleo-Cytoplasmic Large DNA viruses of eukaryotes. *Virol J* 2012; **9**: 1–18.

71. Raoult D, Audic S, Robert C, Abergel C, Renesto P, Ogata H, et al. The 1.2-megabase genome sequence of Mimivirus. *Science* 2004; **306**.

72. Saini HK, Fischer D. Structural and functional insights into Mimivirus ORFans. *BMC Genomics* 2007; **8**.

73. Abrahão JS, Araújo R, Colson P, La Scola B. The analysis of translation-related gene set boosts debates around origin and evolution of mimiviruses. *PLoS Genet* 2017; **13**: e1006532.

74. Silva LCF, Almeida GMF, Assis FL, Albarnaz JD, Boratto PVM, Dornas FP, et al. Modulation of the expression of mimivirus-encoded translation-related genes in response to nutrient availability during Acanthamoeba castellanii infection. *Front Microbiol* 2015; **0**.

75. Behrenfeld MJ, Kolber ZS. Widespread iron limitation of phytoplankton in the south pacific ocean. *Science* 1999; **283**: 840–843.

76. Revilla Y, Cebrián A, Baixerás E, Martínez C, Viñuela E, Salas ML. Inhibition of apoptosis by the African swine fever virus Bcl-2 homologue: role of the BH1 domain. *Virology* 1997; **228**.

77. Bidle KD, Kwityn CJ. Assessing The Role Of Caspase Activity And Metacaspase Expression On Viral Susceptibility Of The Coccolithophore, Emiliania Huxleyi (Haptophyta). *J Phycol* 2012; **48**: 1079–1089.

78. Du J, Wang L, Wang Y, Shen J, Pan C, Meng Y, et al. Autophagy and apoptosis induced by Chinese giant salamander (Andrias davidianus) iridovirus (CGSIV). *Vet Microbiol* 2016; **195**: 87–95.

79. Schrad JR, Abrahão JS, Cortines JR, Parent KN. Structural and Proteomic Characterization of the Initiation of Giant Virus Infection. *Cell* 2020; **181**: 1046–1061.e6.

80. Lartigue A, Burlat B, Coutard B, Chaspoul F, Claverie J-M, Abergel C. The megavirus chilensis Cu,Zn-superoxide dismutase: the first viral structure of a typical cellular copper chaperone-independent hyperstable dimeric enzyme. *J Virol* 2015; **89**: 824–832.

81. Van Etten JL, Meints RH. Giant viruses infecting algae. *Annu Rev Microbiol* 1999; **53**: 447–494.

82. Furuta M, Schrader JO, Schrader HS, Kokjohn TA, Nyaga S, McCullough AK, et al. Chlorella virus PBCV-1 encodes a homolog of the bacteriophage T4 UV damage repair gene denV. *Appl Environ Microbiol* 1997; **63**: 1551–1556.

83. Redrejo-Rodríguez M, Ishchenko A, Saparbaev MK, Salas ML, Salas J. African swine fever virus AP endonuclease is a redox-sensitive enzyme that repairs alkylating and oxidative damage to DNA. *Virology* 2009; **390**: 102–109.

84. Manandhar M, Boulware KS, Wood RD. The ERCC1 and ERCC4 (XPF) genes and gene products. *Gene* 2015; **569**: 153.

85. Wilson WH, Gilg IC, Duarte A, Ogata H. Development of DNA mismatch repair gene, MutS, as a diagnostic marker for detection and phylogenetic analysis of algal Megaviruses. *Virology* 2014; **466-467**: 123–128.

86. Ogata H, Ray J, Toyoda K, Sandaa R-A, Nagasaki K, Bratbak G, et al. Two new subfamilies of DNA mismatch repair proteins (MutS) specifically abundant in the marine

environment. *ISME J* 2011; **5**: 1143–1151.

87. Piacente F, De Castro C, Jeudy S, Molinaro A, Salis A, Damonte G, et al. Giant Virus Megavirus chilensis Encodes the Biosynthetic Pathway for Uncommon Acetamido Sugars. *J Biol Chem* 2014; **289**: 24428–24439.
88. Parakkottil Chothi M, Duncan GA, Armirotti A, Abergel C, Gurnon JR, Van Etten JL, et al. Identification of an L-rhamnose synthetic pathway in two nucleocytoplasmic large DNA viruses. *J Virol* 2010; **84**: 8829–8838.
89. Rodrigues RAL, dos Santos Silva LK, Dornas FP, de Oliveira DB, Magalhães TFF, Santos DA, et al. Mimivirus Fibrils Are Important for Viral Attachment to the Microbial World by a Diverse Glycoside Interaction Repertoire. *J Virol* 2015; **89**: 11812.
90. Geslin C, Gaillard M, Flament D, Rouault K, Le Romancer M, Prieur D, et al. Analysis of the first genome of a hyperthermophilic marine virus-like particle, PAV1, isolated from Pyrococcus abyssi. *J Bacteriol* 2007; **189**: 4510–4519.
91. Rose SL, Fulton JM, Brown CM, Natale F, Van Mooy BAS, Bidle KD. Isolation and characterization of lipid rafts in Emiliania huxleyi: a role for membrane microdomains in host-virus interactions. *Environ Microbiol* 2014; **16**: 1150–1166.
92. Aylward FO, Moniruzzaman M. ViralRecall-A Flexible Command-Line Tool for the Detection of Giant Virus Signatures in ’Omic Data. *Viruses* 2021; **13**.
93. Olm MR, Brown CT, Brooks B, Banfield JF. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J* 2017; **11**: 2864–2868.
94. Weinheimer AR, Aylward FO. Infection strategy and biogeography distinguish cosmopolitan groups of marine jumbo bacteriophages. *ISME J* 2022; **16**: 1657–1667.
95. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 2009; **25**: 1972–1973.
96. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, et al.

IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol Biol Evol* 2020; **37**: 1530–1534.

97. Lemos LN, Fulthorpe RR, Triplett EW, Roesch LFW. Rethinking microbial diversity analysis in the high throughput sequencing era. *J Microbiol Methods* 2011; **86**: 42–51.

98. Schlitzer R. Data Analysis and Visualization with Ocean Data View. http://dx.doi.org/10.1016/s0098-3004(02)00040-7. .

99. Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, et al. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res* 2018; **47**: D309–D314.

100. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, et al. Pfam: The protein families database in 2021. *Nucleic Acids Res* 2020; **49**: D412–D419.

Figure legends

Figure 1. Global map of sampling locations, colored by transect. Blue dots indicate the start of cruise tracks.

Figure 2. Summary of the taxonomy of detected giant viruses. The area of each rectangle is proportional to the number of identified viral genomes in the respective taxon.

Figure 3. Unique genomes and genomes shared between the transects (A) and water depth layers (B). Horizontal bars (right) indicate the total number of genomes found in each transect; black dots indicate the presence in one or multiple transects; the corresponding vertical bars indicate the number of genomes with the presence described by the dots.

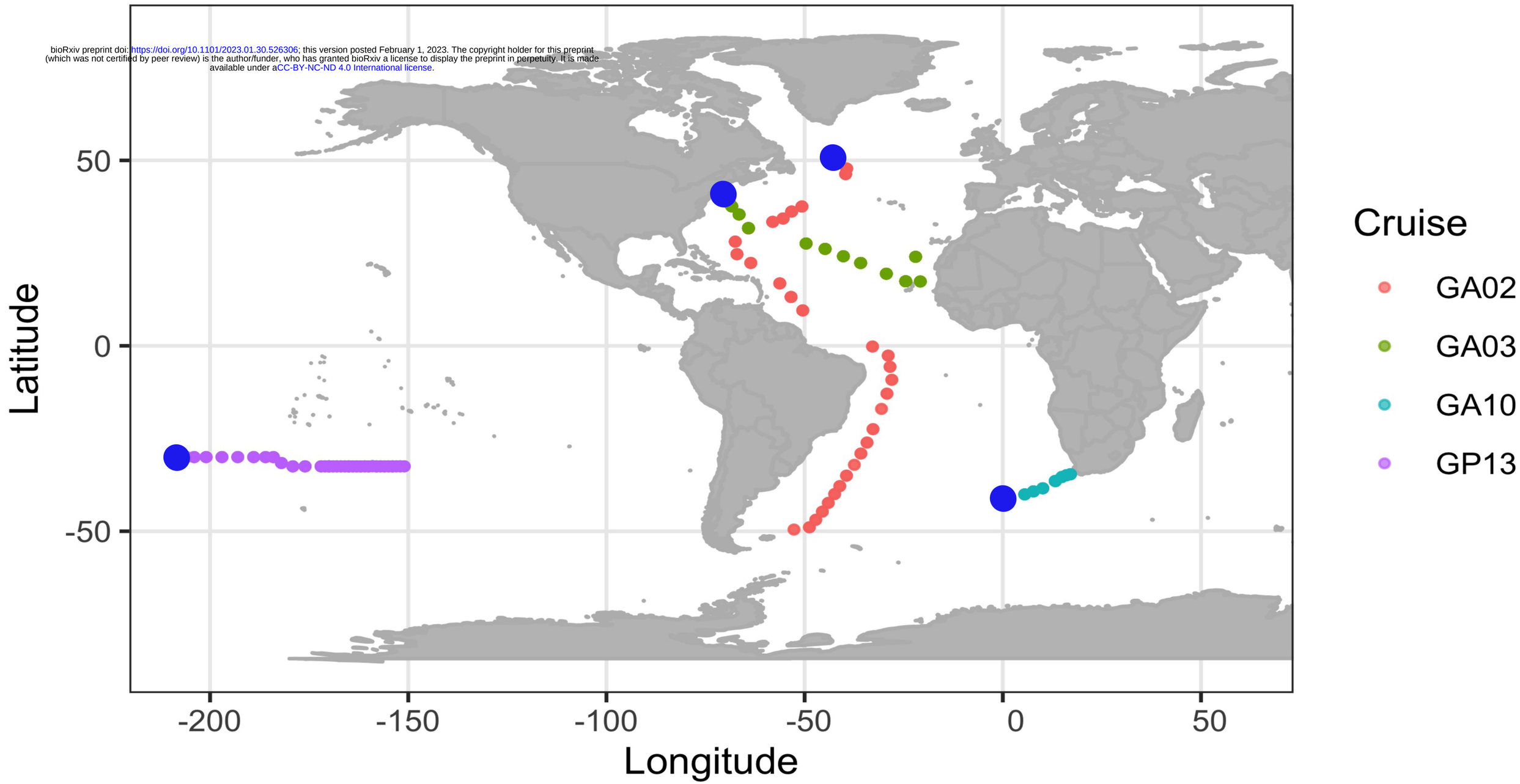
Figure 4. Distribution of giant viruses in each transect. Each column represents a sampling location. The y-axis shows the number of different viral genomes that were recovered at a given location, separated into three depth ranges (2-80m, 80-150m, and 150-5,500m). Locations are arranged in increasing distance from left to right on the x-axis, based on their distance from the starting location and follow the indicated orientation (N to S for GA02 and GA03, W to E for GA10 and GP13).

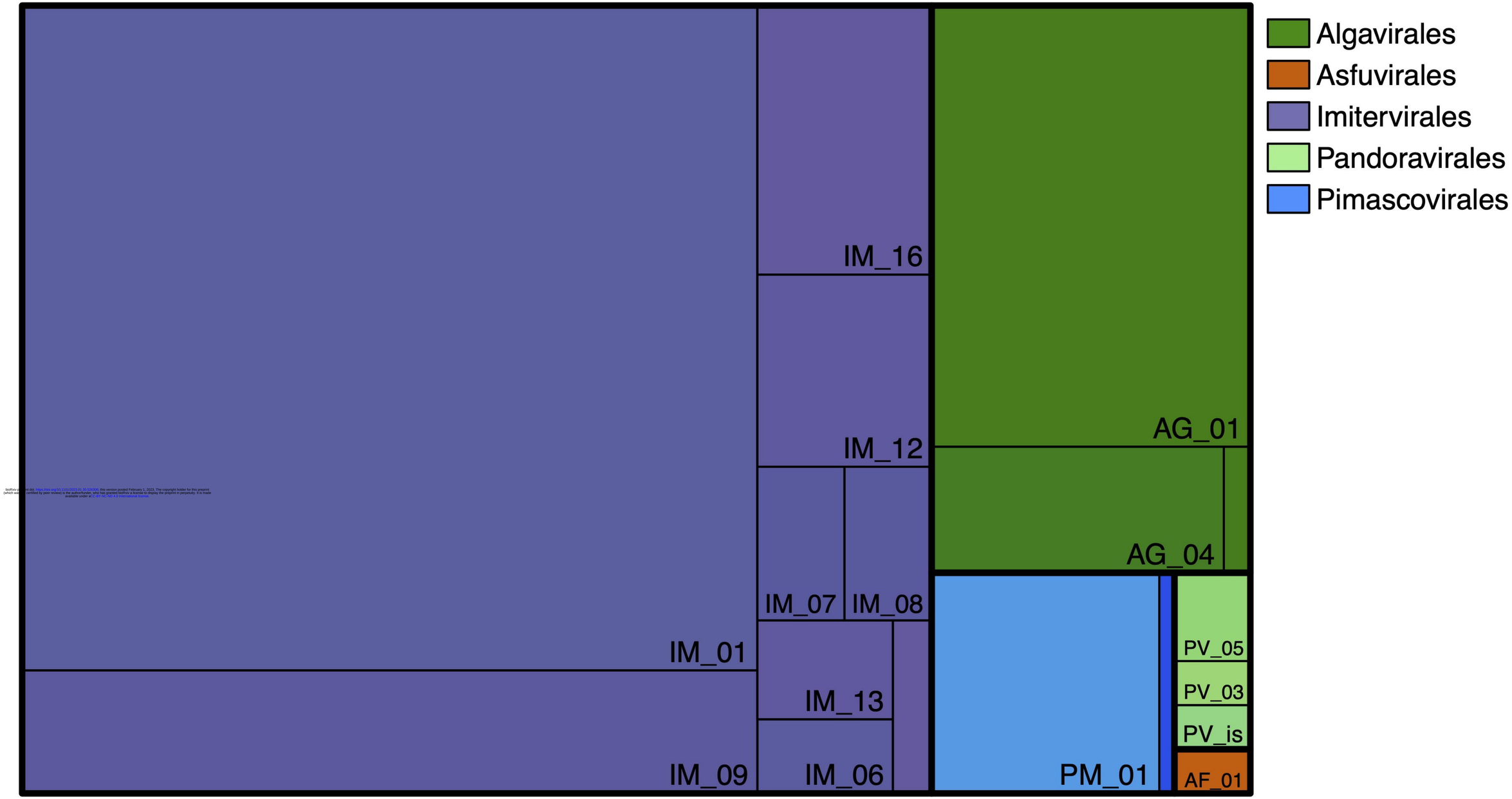
Figure 5. Distribution of viruses throughout the water column along the transects. The viral abundance (calculated in log RPKM) of (A) total giant viruses present in the transect (B) viruses of the *Imitervirales* order (C) viruses of the *Algavirales* order only. Samples were ordered based on the distance along transects, beginning from the first sampling location of cruise tracks (0 km). Black dots denote the sampling location along the transect of each sample.

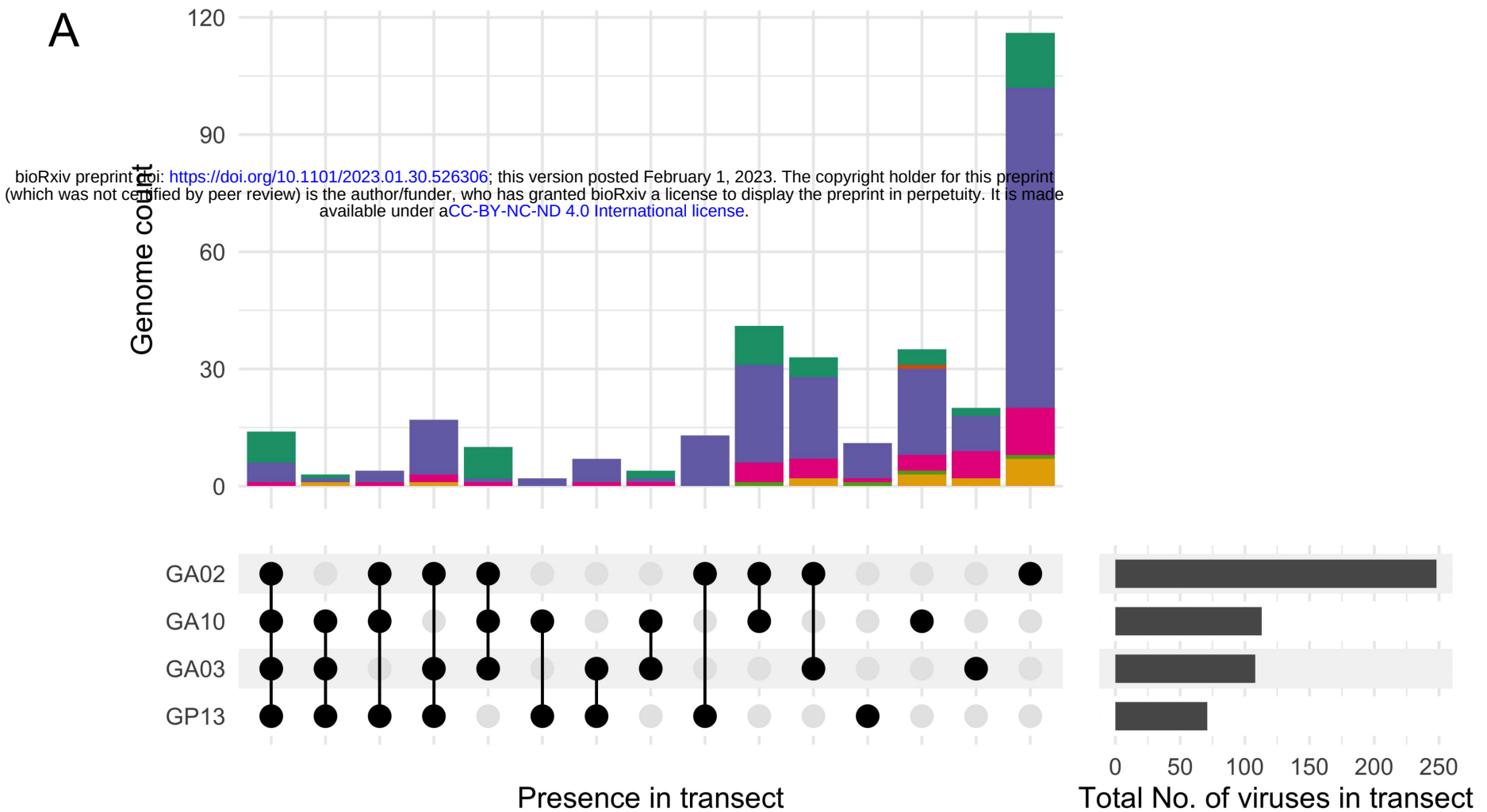
Figure 6. Latitudinal pattern of giant virus diversity across the transect GA02 showed in (A) Shannon's H index (B) Genome richness. Stars showing significant difference between two latitudinal groups (Wilcox test, p-values < 0.05) (* < 0.05, ** < 0.01, *** < 0.001, **** < 0.0001) Panels left: Total virus community; center: Imitervirales communities; right: Algavirales communities. EQ, Equator.

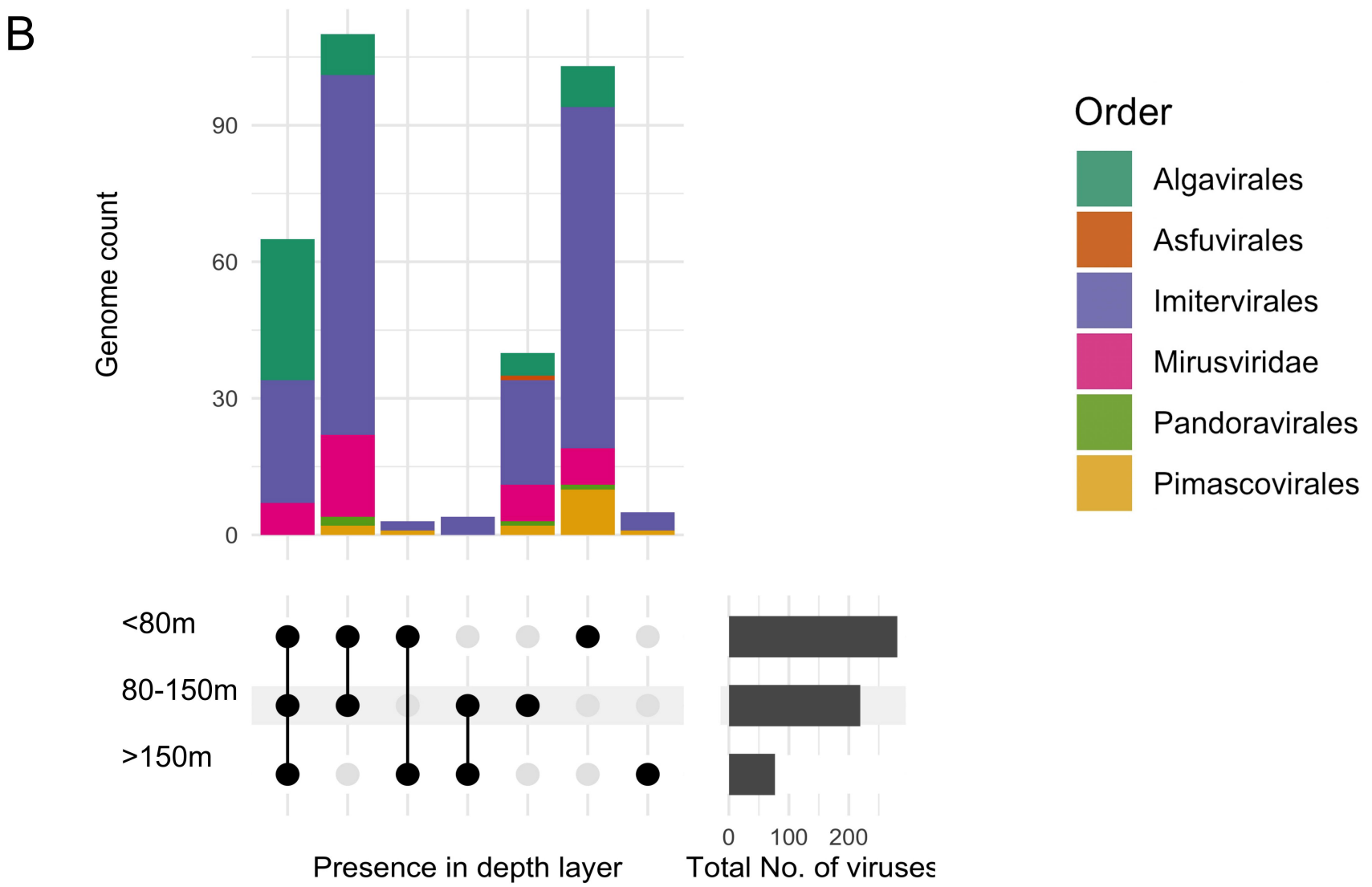
Figure 7. Community composition between latitudinal locations NMDS ordination based on Bray-Curtis distance matrices of viral communities collected in the (A) GA02 transect, stress = 0.3 (B) All four bioGEOTRACES, stress = 0.21. Latitudinal groups are color-coded by sample locations at higher than 40°N/S, from 20° to 40°N/S, and below 20°N/S (equatorial). Ellipses represent 95% confidence intervals. Viral communities are significantly different between groups (Permanova p <0.001).

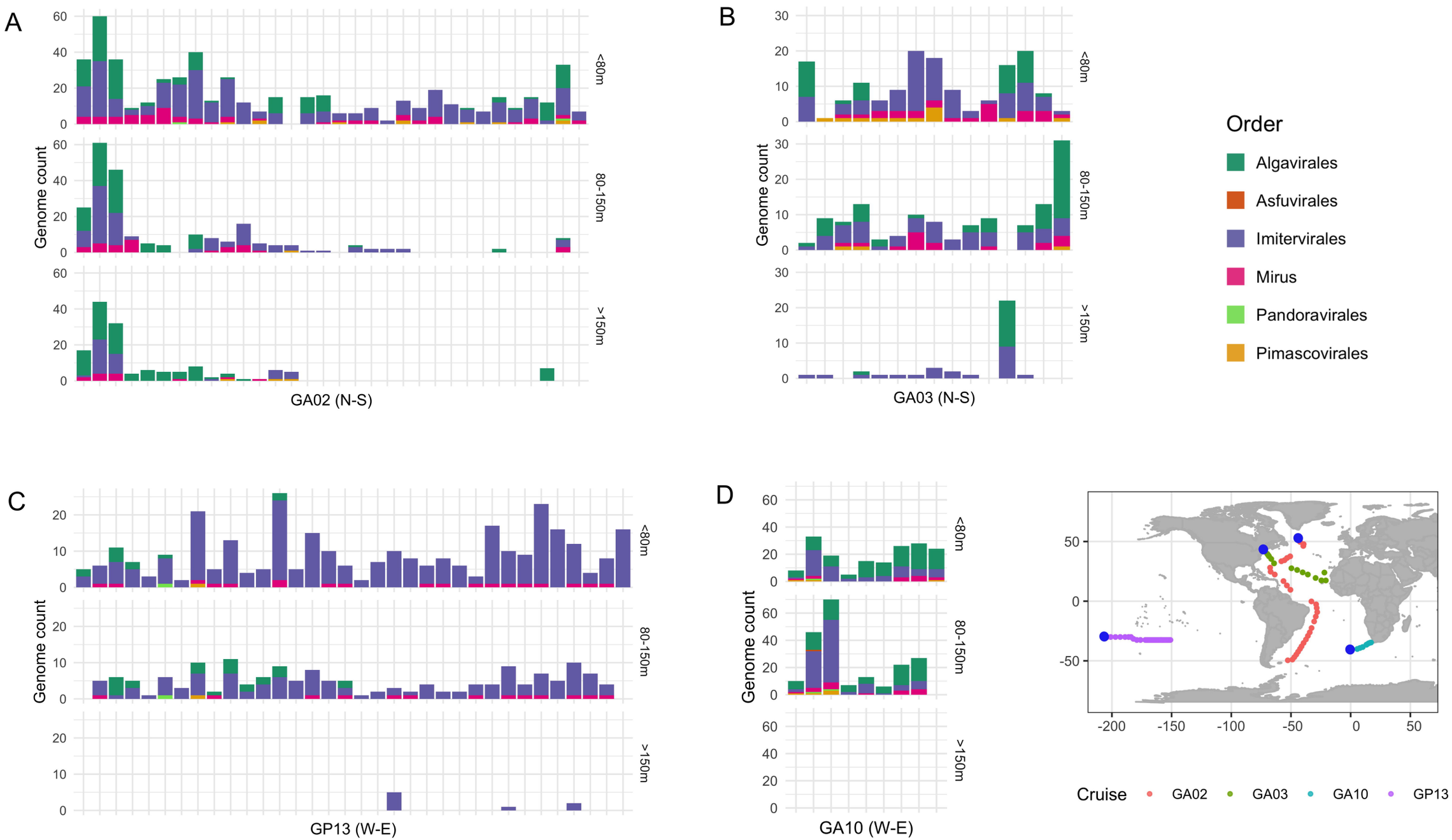
Figure 8. (A) General mapping statistics of viruses found in surface waters <150m. The y-axis shows the average abundance of a given individual genome (in RPKM), the x-axis shows the number of samples from which the virus was recovered. Dots are colored by the viral order and dot sizes represent the length of the genomes. (B) Genomic functional features of the six genomes that are widespread in oligotrophic waters. On the x axis, genomes are ranked from left to right in order of decreasing number of samples in which the viruses were detected; the horizontal colored bar shows the taxonomic order of the genome (purple: *Imitervirales*, green: *Algavirales*). The y axis denotes the functional annotation found in genomes; putative genes are color-coded by functional categories. Gene function abbreviations: PPDK, Pyruvate phosphate dikinase; GAPDH, Glyceraldehyde 3-P dehydrogenase; SDH, Succinate dehydrogenase; LHCB, Chlorophyll a/b binding protein; ACAD, Acyl-CoA dehydrogenase; ACBP, Acyl-CoA binding protein; GMD, GDP-mannose dehydrogenase; GMDH, GDP-mannose 4,6 dehydratase; GlcNAc epimerase, UDP-N-acetylglucosamine 2-epimerase; GNAT, Glucosamine-6-phosphate N-acetyltransferase; PCNA, Proliferating cell nuclear antigen; PI3K, Phosphatidylinositol 3-kinases; PDXK, PD-(D/E)XK nuclease superfamily; TNFR, Tumor necrosis factor receptor.

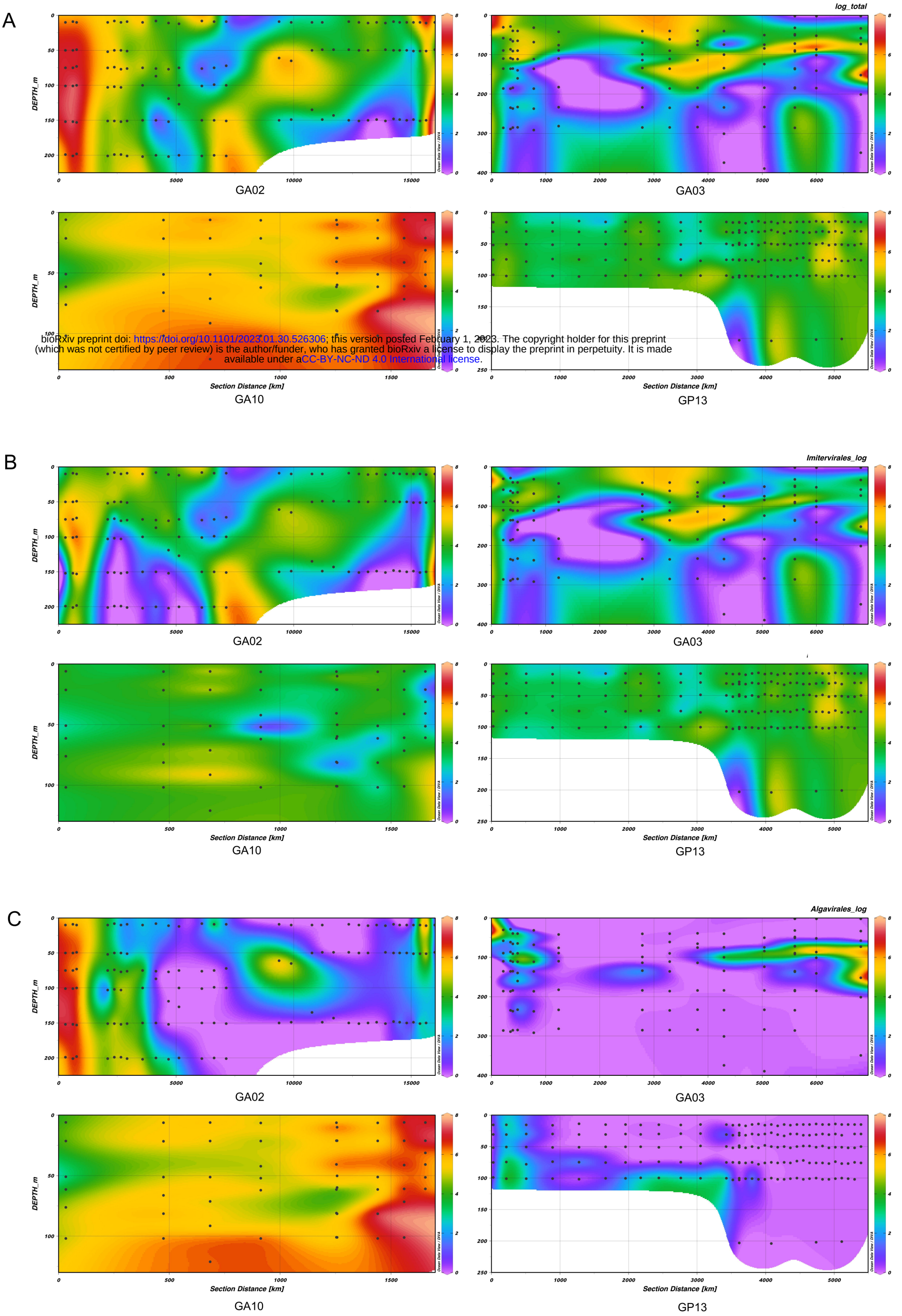


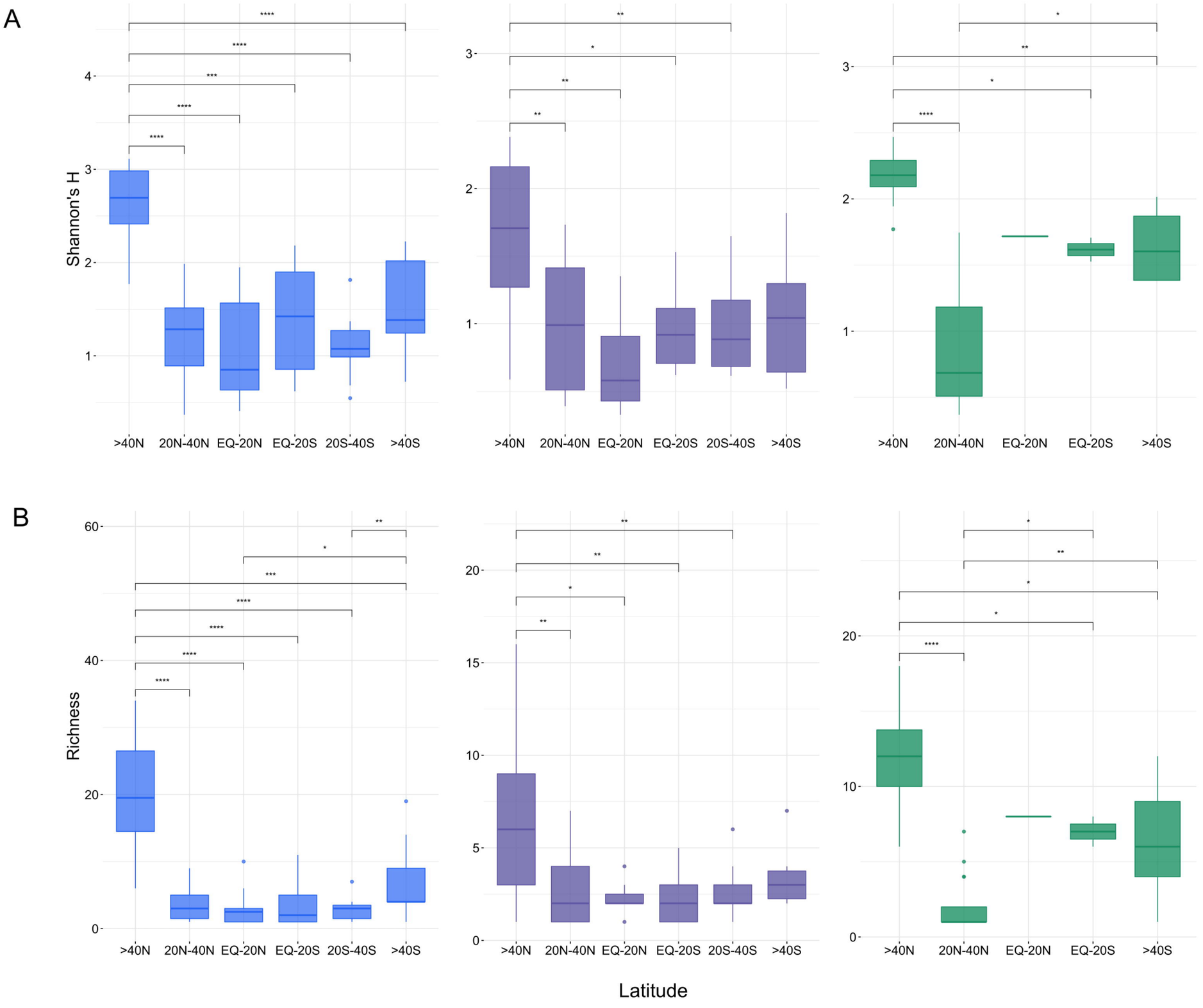


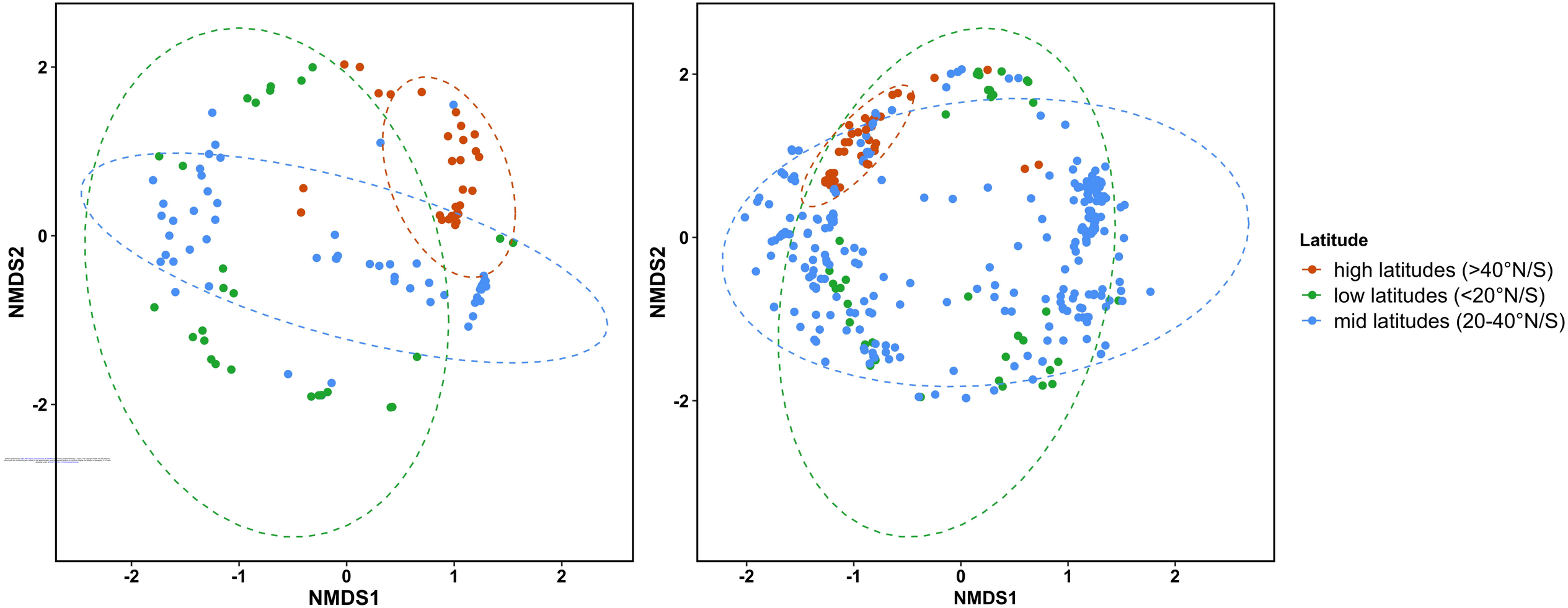




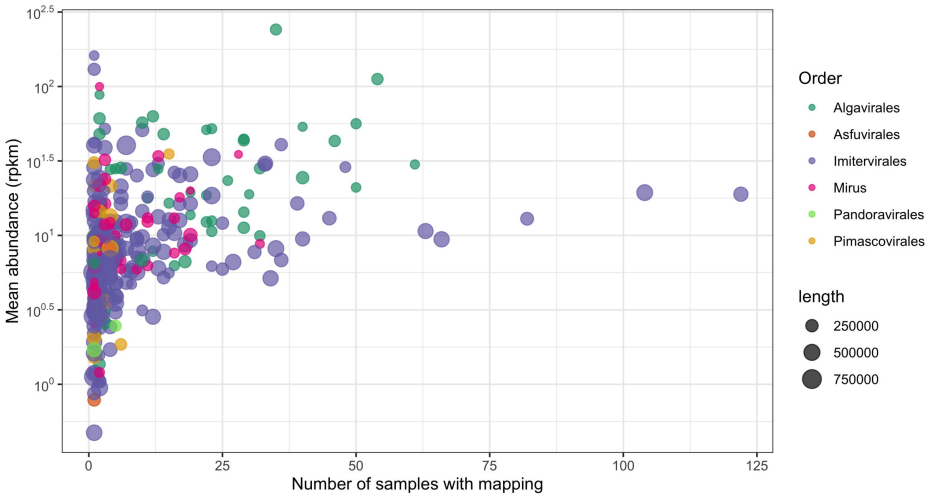

A
B
C
D
Genome count
<80m
80-150m
>150m
GA02 (N-S)
GA03 (N-S)
GP13 (W-E)
GA10 (W-E)
Order
Algavirales
Asfuvirales
Imitervirales
Mirus
Pandoravirales
Pimascovirales
Cruise
GA02
GA03
GA10
GP13




A
B
Shannon's H
Richness
>40N
20N-40N
EQ-20N
EQ-20S
20S-40S
>40S
Latitude


NMDS2
NMDS1
2
0
-2
-2
-1
0
1
2
Latitude
high latitudes (>40°N/S)
low latitudes (<20°N/S)
mid latitudes (20-40°N/S)

A


Mean abundance (rpkm)
Number of samples with mapping
Order
Algavirales
Asfuvirales
Imitervirales
Mirus
Pandoravirales
Pimascovirales
length
250000
500000
750000

B


Pathway
NCLDV core genes
Energy metabolism
Photosynthesis
Beta oxidation
Nutrient metabolism
Cytoskeleton
Sugar metabolism
Transporter
Signal transduction and regulation
Virus-host interaction
DNA repair
Programmed cell death
Stress regulation
A32
mcp
PolB
SFII
VLTF3
TFIIB
RNAPS
RNAPL
TopoII
PPDK
Aldolase
GAPDH
Phosphoglucomutase
Aconitase
SDH
Transketolase
LHCB
Rhodopsin
Chlorophyllase
ACAD
ACBP
Sulfotransferase
Exostosin
Ferritin
Pho regulon protein
Kinesin
GMD
GMDH
GlcNAc epimerase
GNAT
Glycosyltransferase
Mg2+ transporter
Amino acid permease
Ion channel
Sulfite exporter
VIT family transporter
Sugar efflux transporter
ABC transporter
Ser/Thr kinase
Ulp1 protease
Lectin
PCNA
MutS
Nucleotidyltransferase
YqaJ
PI3K
PDXK
ERCC4 nuclease
Caspase
TNFR
E3 ubiquitin ligase
Cold–shock protein
Lon protease
Superoxide dismutase
Heat–shock protein
Thioredoxin
Glutaredoxin
ERX556088.18.dc
GVMAG–S–3300012919–55
TARA_MED_NCLDV_00063
TARA_MED_NCLDV_00027
TARA_AOS_NCLDV_00003
TARA_IOS_NCLDV_00011