# Infection strategy and biogeography distinguish cosmopolitan groups of marine jumbo bacteriophages

Alaina R. Weinheimer[1], Frank O. Aylward[1,2]

[1]Department of Biological Sciences, Virginia Tech, Blacksburg, VA, USA. [2] Center for Emerging, Zoonotic, and Arthropod-borne Pathogens, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061-0913

Correspondence to email: faylward@vt.edu; address: 5092 Derring Hall, 926 W Campus Drive, Blacksburg, VA 24061; telephone: (540) 231-8657

Competing Interests: The authors declare no competing interests.
Funding Support: This work was supported by grants from the Institute for Critical Technology and Applied Science, the NSF (IIBR-1918271) and a Simons Early Career Award in Marine Microbial Ecology and Evolution to F.O.A. ARW was supported by an ICTAS Doctoral Scholars Fellowship.

# Infection strategy and biogeography distinguish cosmopolitan groups of marine jumbo bacteriophages

Alaina R. Weinheimer[1], Frank O. Aylward[1,2]

[1]Department of Biological Sciences, Virginia Tech, Blacksburg, VA, USA. [2] Center for Emerging, Zoonotic, and Arthropod-borne Pathogens, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061-0913

Correspondence to faylward@vt.edu

**Keywords**: Jumbo bacteriophage, viral diversity, viral gigantism, marine microbiology, microbial oceanography

## ABSTRACT

Recent research has underscored the immense diversity and key biogeochemical roles of large DNA viruses in the ocean. Although they are important constituents of marine ecosystems, it is often difficult to detect these viruses due to their large size and complex genomes. This is true for “jumbo” bacteriophages, which have genome sizes >200 kbp and large capsids reaching up to 0.45 µm in diameter. In this study, we sought to assess the genomic diversity and distribution of these bacteriophages in the ocean by developing a bioinformatic pipeline to generate and validate jumbo phage genomes from metagenomes. We recover 85 marine jumbo phages that ranged in size from 201-498 kilobases, and we examine their genetic similarities and biogeography together with a reference database of marine jumbo phage genomes. By analyzing Tara Oceans metagenomic data we show that jumbo phages are less common in <0.22 µm size fractions but are widespread in larger fractions, consistent with their large size. Our network-based analysis of gene sharing patterns reveals that jumbo bacteriophage belong to five genome clusters that are typified by diverse replication strategies, genomic repertoires, and potential host ranges. Our analysis of jumbo phage distributions in the ocean reveals that depth is a major factor shaping their biogeography, with some phage genome clusters exhibiting higher relative abundance in either surface or mesopelagic waters, respectively. Taken

together, our findings demonstrate that jumbo phages are widespread community members in the ocean with complex genomic repertoires and ecological impacts that warrant further targeted investigation.

## Introduction

Although historically noted for their small virion sizes and simple genomes [1], viruses with large particles and elaborate genomes have been discovered in recent decades throughout the biosphere [2–4]. These complex viruses not only invite intriguing evolutionary questions [5–7], but also expand the potential roles viruses have in shaping microbial community structure and biogeochemical cycling [4, 8–10]. One group of these larger viruses are jumbo bacteriophages (jumbo phages), which have traditionally been defined as *Caudovirales* with genomes over 200 kilobases in length [11]. While a recent survey of cultured jumbo phages by Iyer and colleagues [6] showed jumbo phages share some universal features and genes, such as encoding DNA polymerases and the terminase large subunit (TerL), these features do not distinguish them from smaller phages, and several lines of evidence suggest that jumbo phages emerged from smaller phages multiple times independently. For example, a recent phylogenetic study of cultured *Caudovirales* used concatenated protein alignments to generate phylogenies and found that the most supported clades within the *Caudovirales* family do not consistently correspond to genome length [12]. Furthermore, Iyer et al. found that jumbo phages cluster with smaller phages based on gene content and are best grouped by replication machinery, among other infection apparati [6]. Taken together, jumbo phages likely emerged from diverse smaller phages multiple times independently and form distinct clades within the *Caudovirales*.

Although the first jumbo phages were isolated as early as the 1970s [13], these viruses have remained relatively sparse in culture, representing less than 3% (n=93) of complete

*Caudovirales* genomes on NCBI's RefSeq Viral Genome Portal (downloaded July 5, 2020). All cultured jumbo phage capsids have morphologies belonging to the *Myoviridae* and *Siphoviridae* families and undergo infection cycles that reflect temporal patterns of lytic *Caudovirales* [14]. Some jumbo phages are known to stall infections resulting in "pseudolysogeny", however, which has been proposed as a competitive strategy against other phages to prevent superinfection [15]. Jumbo phages that have been studied extensively are primarily investigated for their unusually complex functional capabilities, such as encoding entire transcriptional apparatti [16] or sophisticated anti-CRISPR defense mechanisms [17, 18]. Although most cultivated jumbo phages have been isolated from a variety of Gram positive and Gram negative model and pathogenic bacteria [14], a recent metagenomic survey uncovered these viruses in diverse ecosystems around the world [4].

Despite this apparent broad environmental distribution, commonly-used methods for viral isolation and diversity surveys bias against the inclusion of jumbo phages. Because viruses have historically been considered to be smaller than cells, many viral diversity surveys specifically examine only small particle sizes. For example, in plaque assays, agar concentrations are often too high for larger phage particles to diffuse through compared to smaller particles [19]. Moreover, filters are often used to remove cells when preparing viral enrichments for metagenomic sequencing [20], which excludes larger viruses [9, 21]. Particularly in marine studies, the <0.22 μm fraction, sometimes even referred to as the "viral fraction" [22], is most commonly examined for viruses[23–25]. Jumbo phages can have particles over 0.45 μm in length (i.e. *Pseudomonas aeruginosa* phage PhiKZ)[13], however, and will therefore be excluded from <0.22 μm size fractions. Lastly, in bioinformatic pipelines, phage sequences are typically only assembled to the contig or scaffold level, which is sometimes sufficient for the assembly of most known smaller phage genomes [26], but often leaves larger phage genomes fragmented into multiple contigs and may require additional joining of contigs into bins [27]. Overall, considering these biases and the recently-discovered broad distribution

of these viruses [4], jumbo phages may represent underappreciated components of marine microbial communities and food webs that warrant further examination.

In this study, we examine the diversity and prevalence of jumbo phages in the global ocean. We developed a workflow for generating and validating high-quality jumbo bacteriophage bins from metagenomic data, with which we identified 85 bins of jumbo phages. We then compared the genetic content of these jumbo phages with other cultured phages of all sizes and metagenomic jumbo phages from other studies. We found that the jumbo phages of this study grouped into five distinct clusters that are distinguished by diverse replication machinery and infection strategies, implicating a broad range of potential jumbo-phage host interactions in the ocean. We then assessed the distribution of jumbo phages belonging to these genome clusters in the ocean by using metagenomic data from Tara Oceans [28,29]. Mapping the Tara Oceans metagenomic data onto the jumbo phage sequences revealed that these jumbos phages are collectively widely distributed in the ocean, but vary in biogeography based on cluster, with some more enriched in surface waters relative to deeper waters and vice versa. Upon examining the collective presence of jumbo phages in different filter fractions, we also found that commonly used size fractions in viral diversity surveys reduce the ability to recover and detect jumbo phage sequences. As these complications apply to many viral diversity studies in other environments, our results indicate that jumbo phages may play underappreciated roles in ecosystems around the globe.

## RESULTS AND DISCUSSION

### Detection and validation of high-quality jumbo phage bins

Due to the large size of jumbo bacteriophage genomes, it is likely that they are present in multiple distinct contigs in metagenomic datasets and therefore require binning to recover high-quality metagenome-assembled genomes (MAGs) [27]. This has been shown for large DNA viruses that infect eukaryotes, where several recent studies have successfully employed binning

127 approaches to recover viral MAGs [2, 3]. Here, we used the same 1,545 high-quality
128 metagenomic assemblies [30] used in a recent study to recover giant viruses of eukaryotes [3],
129 but we modified the bioinformatic pipeline to identify bins of jumbo bacteriophages. We first
130 binned the contigs from these assemblies with MetaBat2 [31], which groups contigs together
131 based on similar nucleotide composition and coverage profiles, and focused on bins of at least
132 200 kilobases in total length. We subsequently identified bins composed of bacteriophage
133 contigs through analysis with VirSorter2 [32], VIBRANT [33], and CheckV [34] (see Methods for
134 details).
135 The occurrence of multiple copies of highly-conserved marker genes is typically used to assess
136 the level of contamination present in metagenome-derived genomes of bacteria and archaea
137 [35]. Because bacteriophage lack these marker genes, we developed alternative strategies to
138 assess possible contamination in our jumbo phage bins. Firstly, we refined the set of bins by
139 retaining only those with no more than 5 contigs that were each at least 5 kilobases in length to
140 reduce the possibility that spurious contigs were put together. Secondly, we assessed the
141 possibility that two strains of smaller phages with similar nucleotide composition may be binned
142 together by aligning the contigs in a bin to each other. Bins that had contigs with high sequence
143 similarity across the majority of their lengths were discarded (Supplementary Figure 1). Thirdly,
144 we discarded bins if their contigs exhibited aberrant co-abundance profiles compared to other
145 contigs in the same bin. We identified these bins through read mapping and co-abundance
146 analysis of 225 marine metagenomes from Tara Oceans. Coverage variation was benchmarked
147 based on read-mapping results from artificially-fragmented reference genomes present in the
148 samples (See Methods for details). Only bins with coverage variation below our empirically-
149 derived threshold were retained. Using this stringent filtering, we identified 85 bins belonging to
150 jumbo bacteriophages that are present in Tara Oceans metagenomes. These bins ranged in
151 length from 202 bp to 498 kbp, and 31 consisted of a single contig, while 54 consisted of 2-5
152 contigs (Supplemental Figure 2).

To assess global diversity patterns of jumbo bacteriophage we combined these jumbo phage bins together with a compiled database of previously-identified jumbo phages that included all complete jumbo *Caudovirales* genomes on RefSeq (downloaded July 5th, 2020), the INPHARED database [36], the Iyer et al 2021 study [6], the Al-Shayeb et al. 2020 study [4], and marine jumbo phage contigs from metagenomic surveys of GOV 2.0 [25] (60 jumbo phages), ALOHA 2.0 [37] (8 jumbo phages), and one megaphage MAG recovered from datasets of the English Channel [38]. Ultimately, we arrived at a set of 223 jumbo phages present in at least one Tara Oceans sample that we analyzed further in this study and refer to as marine jumbo phages. Statistics on genomic features can be found in Supplemental Dataset 1.

**Marine jumbo phages belong to distinct groups with diverse infection strategies**

Because bacteriophages lack high-resolution, universal marker genes for classification, such as 16S rRNA in bacteria, phages are often grouped by gene content [39, 40]. Here, we generated a bipartite network that included the 85 bins of jumbo phages with a dataset of available *Caudovirales* complete genomes in RefSeq (3,012 genomes; downloaded July 5th, 2020) and the set of jumbo phages described above. To construct the bipartite network, we compared proteins encoded in all the phage genomes to the VOG database (vogdb.org), and each genome was linked to VOG hits that were present (Figure 1, Supplemental Dataset 2, see Methods for details). To identify groups of phage genomes with similar VOG profiles, we employed a spinglass community detection algorithm [41] to generate genome clusters. The marine jumbo phages of this study clustered into five groups that all included both jumbo and non-jumbo phage genomes (Figure 2a). We refer to these five clusters as Phage Genome Clusters (PGCs): PGC_A, PGC_B, PGC_C, PGC_D, and PGC_E. These PGCs included cultured phages and metagenome-derived jumbo phages found in various environments (i.e. aquatic, engineered) and isolated on a diversity of hosts (i.e. Firmicutes, Proteobacteria, Bacteroidetes) (Figure 2b,c). Of the marine jumbo phages, 134 belonged to PGC_A, 11 to PGC_B, 90 to PGC_C 27, 7 to PGC_D, and 1 to PGC_E (Figure 1b). In addition to this network-

based analysis, we also examined phylogenies of the major capsid protein (MCP) and the terminase large subunit (TerL) encoded by the marine jumbo phages and the same reference phage set examined in the network (Figure 1c, 1d). With the exception of PGC_A, the marine jumbo phages that belong to the same PGC appeared more closely related to each other than those belonging to different clusters. The evolutionary history of these marker genes therefore supports the results of the bipartite network genome clustering and is consistent with the view that genome gigantism evolved multiple times independently within the *Caudovirales* [6].

We then compared functional content encoded by the marine jumbo phages in the PGCs to identify functional differences that distinguish these groups. PGC_E was excluded from this analysis because this genome cluster contained only a single jumbo phage. Collectively, most genes of the marine jumbo phages could not be assigned a function (mean: 86.60%, std dev: 7.01%; Supplemental Dataset 3), which is common with environmental viruses [42, 43]. Genes with known functions primarily belonged to functional categories related to viral replication machinery, such as information processing and virion structure (Figure 3a), and these genes drove the variation between the genome clusters of marine jumbo phages (Figure 3b). Iyer et al provided a comparative genomic analysis that was able to identify three types of jumbo phages that are defined by different infection strategies and host interactions (referred to as Groups 1-3) [6]. We cross-referenced our PGCs and found that PGCs B, C, and D of this study corresponded to Iyer Groups 1, 2, and 3, respectively, suggesting that these genome clusters contain phages with distinct infection and replication strategies. PGC_A corresponded to multiple phage groups in the Iyer study, indicating that this genome cluster contains a particularly broad diversity of phages.

PGC_B consists of 238 phages (11 (4.6%) marine jumbo phages, including 10 bins generated here), and included cultured phage of Group 1 of Iyer et al, which is typified by *Pseudomonas aeruginosa* phage PhiKZ. Supporting this correspondence with Iyer Group 1, all marine jumbo phages of PGC_B encoded the same distinct replication and transcription machinery, including

205 a divergent family B DNA polymerase and a multi-subunit RNA polymerase (Figure 3b,
206 Supplemental Dataset 3). These marine jumbo phages also encoded a PhiKZ internal head
207 protein, and they uniquely encoded shell and tubulin homologs which has recently been found in
208 PhiKZ phages to assist in the formation of a nucleus-like compartment during infection that
209 protects the replicating phage from host defenses [17,18]. Although we could not confidently
210 predict hosts for the 11 metagenomic marine jumbo phages in this PGC_B (Supplemental
211 Dataset 1), the cultured phages of this genome cluster infect pathogenic bacteria belonging to
212 the phyla Proteobacteria (178 phages) and Firmicutes (6 phages) (Figure 2c), implicating a
213 potential host range for marine jumbo phages in PGC_B.
214 PGC_C is comprised of 156 phages total (90 marine jumbo phages (57.7%); 4 bins generated
215 from this study) and included reference jumbo phages in Iyer Group 2 (31, 19.9%) which are
216 typified by Alphaproteobacteria and Cyanobacteria phages. Likewise, the host range of other
217 cultured phages in PGC_C support the Group 2 correspondence, either infecting Cyanobacteria
218 (139 phages) or Proteobacteria (4 phages) (Figure 2c). Furthermore, all 3 marine metagenomic
219 phages in PGC_C for which hosts could be predicted were matched to Cyanobacteria hosts
220 (Supplemental Dataset 1). Functional annotations of PGC_C marine jumbo phages revealed
221 nearly all encode a family B DNA polymerase (97.8% phages) and the photosystem II D2
222 protein (PF00124, VOG04549) characteristic of cyanophages (90% phages) (Figure 3b). This
223 PGC included the reference *Prochlorococcus* phage P-TIM68 (NC_028955.1), which encodes
224 components of both photosystem I and II as a mechanism to enhance cyclic electron flow during
225 infection [44]. This suggests that an enhanced complement of genes used to manipulate host
226 physiology during infection may be a driver of large genome sizes in this group. Additionally,
227 most of the PGC_C marine jumbo phages encoded Lipopolysaccharide (LPS) biosynthesis
228 proteins (76%), which have been found in cyanophage genomes that may induce a
229 "pseudolysogeny" state, when infected host cells are dormant, by changing the surface of the
230 host cell and preventing additional phage infections [15] (Supplemental Dataset 3). Taken

231 together, most marine jumbo phages of PGC_C likely follow host interactions of jumbo
232 cyanophages, such as potentially manipulating host metabolism by encoding their own
233 photosynthetic genes and potentially inducing a pseudolysogenic state.
234 PGC_D comprised 47 phages, of which 7 were marine jumbo phages generated in this study
235 (14.9%). This group included Iyer Group 3 jumbo phages (15, 31.9%), which is primarily
236 distinguished by encoding a T7-type DNA polymerase but is not typified by a particular phage
237 type or host range. Supporting this grouping, all marine jumbo phages in this study encoded this
238 T7 DNA polymerase which belongs to family A DNA polymerases (Figure 3b, Supplemental
239 Dataset 3). Most of the other genes distinctively encoded by the marine jumbo phages in this
240 group included structural genes related to T7 (T7 baseplate, T7 capsid proteins), a eukaryotic
241 DNA topoisomerase I catalytic core (PF01028), and a DNA structural modification genes
242 (MmcB-like DNA repair protein, DNA gyrase B). Hosts of metagenomic marine jumbo phages in
243 PGC_D could not be predicted (Supplemental Dataset 1); however, cultured Group 3 jumbo
244 phages in PGC_D all infect Proteobacteria, primarily Enterobacteria and other pathogens.
245 PGC_A was the largest of the PGCs with 469 phages in total, including 134 marine jumbo
246 phages (63 bins from this study). This genome cluster contained the largest jumbo phages,
247 such as *Bacillus* phage G (498 kb) and the marine megaphage Mar_Mega_1 (656 kb) recently
248 recovered from the English Channel [38]. Unlike other PGCs, PGC_A contained mostly
249 metagenomic phages (401, 85%, Figure 2b,c). Considering PGC_A contains the largest jumbo
250 phages (Figure 1b, 2a), the vast genetic diversity in this PGC might explain why few genes were
251 found to distinguish this group. Of the genes unique to PGC_A, only one was present in at least
252 half of the phages (51.9%), which was a Bacterial DNA polymerase III alpha NTPase domain
253 (PF07733). The host ranges of cultured phages from this PGC further reflect the phage diversity
254 and included a variety of phyla and genera that can perform complex metabolisms or lifestyles,
255 such as the nitrogen-fixing Cyanobacteria of the *Nodularia* genus isolated from the Baltic Sea

(corresponding to phage sequences NC_048756.1 and NC_048757.1) and the Bacteroidetes bacteria *Rhodothermus* isolated from a hot spring in Iceland [45].

**Relative abundance of jumbo bacteriophages across size fractions**

To explore the distribution of the marine jumbo phages in the ocean, we first examined the size fractions in which the jumbo phages were most prevalent. To remove redundancy, we clustered the 223 jumbo phages into populations if they were highly similar at the nucleotide level ( >80% genes with >95% average nucleotide identity) [23], which resulted in 120 jumbo phage populations. Out of the 225 Tara Oceans metagenomes examined, 202 (89.8%) contained at least one jumbo phage population (median: 24, Supplemental Dataset 4). Jumbo phages were less frequently detected in samples below 0.22 µm (<-0.22 µm, 0.1-0.22 µm) than those above 0.22 µm (0.45-0.8 µm, 0.22-0.45 µm, 0.22-1.6 µm, 0.22-3 µm). Only 54.5% of samples in the <-0.22 µm fraction and 82.4% of samples in the 0.1-0.22 µm fraction had at least one jumbo phage present. Conversely, all samples in all larger size fractions above 0.22 µm had at least one jumbo phage present (Figure 4a). Furthermore, we detected 21 populations (14.9%) exclusively in samples above 0.22 um, compared to only one population (0.71%) exclusive to samples below 0.22 µm. A similar disparity in virus detection between size fractions has been reported for large eukaryotic viruses, where roughly 41% of phylotypes were present in samples of the picoplankton fraction of 0.22-3 µm but absent in size fractions below 0.22 µm [9]. In contrast to this study, where certain viral groups were more prevalent in larger size fractions than smaller, a jumbo phage's PGC membership or genome size did not affect its probability of detection at different size fractions (Supplemental Figure 3, Supplemental Figure 4).

We also compared jumbo phage diversity (defined as population richness), relative abundance, and community composition between the size fractions. Samples of the size fraction <-0.22 µm were significantly less diverse (P-value 0.0003, Wilcox test) and had significantly lower relative abundances of jumbo phages relative to the large size fractions of 0.1-0.22 and 0.22-1.6 (P-values 0.003, 0.031 Wilcox test, respectively). Samples of the 0.1-0.22 µm size fraction had the

highest diversity and relative abundances compared to the other fractions (Figure 4b,c). The enhanced signal of jumbo phages relative to other entities in the 0.1-0.22 μm size fraction may result from the reduction of cells below the 0.22 μm threshold and the reduction of most other viruses above the 0.1 μm limit. Despite these differences in diversity and relative abundances, jumbo phage community composition did not significantly differ between the >0.22 and <0.22 μm size fractions (P-value 0.8109, ANOSIM, presence/absence Bray-Curtis distance matrix, Figure 4d,e).

To directly test the effect of the 0.22 μm size fraction cut-off on jumbo phage recovery, we examined a subset of the samples that were co-collected at the same station and depth for the fractions below 0.22 μm (<-0.22 or 0.1-022) and above 0.22 μm (0.1-0.22 μm or 0.22-3 μm). The number of detected jumbo phage populations was significantly higher in samples above 0.22 μm than below 0.22 μm (P-value = 0.0004, Wilcox test, Supplemental Figure 5a). The relative abundance of jumbo phages was not significantly different between this size distinction (P-value = 0.1309, Wilcox test, Supplemental Figure 5b), however, nor was community composition (P-value = 0.1229, ANOSIM, presence/absence Bray-Curtis distance matrix, Supplemental Figure 5c,d). Taken together, these findings suggest that using size fractions below 0.22 μm to analyze phages limits the recovery of jumbo phage populations detected, likely due the large size of these viruses. The enhanced relative abundance of jumbo phages in the 0.1-0.22 μm relative to the other size fractions suggests that this may be the optimal size fraction to examine for studies that seek to target jumbo phages specifically.

**Biogeography of jumbo bacteriophages in the global ocean**

*Jumbo phage populations varied in depth distribution by PGC.* Jumbo populations in this study were found in all three Longhurst biomes (Trades, Westerlies, Coastal) and depths (Surface, Deep Chlorophyll Maximum, Mesopelagic) examined (Figure 5, Supplemental Figure 6; Supplemental Dataset 4). We focused these biogeographic analyses on the 0.22-1.6 or 0.22-3 μm size fractions because the most sites were available for these samples, and because we

found that jumbo phages were recovered at a higher frequency in larger size fractions. Jumbo phage communities differed significantly between depths (P-value = 0.0001, ANOSIM based on presence/absence Bray-Curtis distance matrix, Supplemental Figure 7c,d), consistent with the dramatic transition in community composition that occurs from surface waters to below the deep chlorophyll maximum [37, 46]. Specifically, the diversity of jumbo phages across depths varied by genome cluster, with PGC_A and PGC_C exhibiting higher abundance and diversity in the epipelagic (SRF and DCM), and PGC_B and PGC_D exhibiting higher abundance in the mesopelagic (Figure 5d, Supplemental Figure 7a,b). PGC_C is typified by cyanophages, providing a clear reason why this phage group is enriched in surface waters. Conversely, PGC_B is typified by *Pseudomonas aeruginosa* PhiKZ phages, suggesting these PGC_B marine jumbo phages may be infecting heterotrophic bacteria that potentially thrive in deeper waters. Overall relative abundance and diversity of jumbo phages in this study were significantly higher in the epipelagic zone, partly because most of these viruses are in PGC_A. This pattern held when examining only those samples that were co-collected at all three depths (Supplemental Figure 8,9). This general pattern therefore reflects what has been found in previous studies on the depth distribution and viral protein clusters, where more were unique in the euphotic (i.e. epipelagic) than aphotic depths [9, 25, 47, 48], although this contrasts what has recently been found in the Pacific Ocean, where overall viral diversity increased in the mesopelagic [37].

*Jumbo phage biogeography across biomes.* Jumbo phage communities in this study significantly differed in composition in surface samples between the Trades, Coastal, and Westerlies biomes, albeit weakly (P-value 0.0021, ANOSIM, presence/absence Bray-Curtis distance matrix, Supplemental Figure 10c,d). However, no biome appeared to be a particular hotspot for jumbo phages, as they were not significantly more diverse in any particular biome (Supplemental Figure 10a). Although the relative abundances of jumbo phages were higher in Trades samples relative to Coastal samples (Supplemental Figure 10b), this pattern did not

always hold when comparing biomes at different depths (Supplemental Figure 11). Similarly, jumbo phages from all PGCs could be detected in all biomes, but no PGC was consistently enriched in diversity and relative abundance in a particular biome (Supplemental Figure 12). A recent global study on marine viruses has found that viral diversity was better explained by ecological zones defined by physicochemical factors like temperature, rather than by Longhurst biomes defined by patterns of chlorophyll *a* concentrations [25], suggesting that Longhurst biomes may not be good predictors of viral diversity in general.

Jumbo phage populations ranged in endemicity with seven populations (5.8%) detected in only one Tara station and eleven (9.2%) present in 50 or more stations (Figure 5c). More endemic populations, present in only one Tara station, belonged to PGCs A, B, and C, while the most prevalent jumbo phage, those present in 50 or more stations, belonged to PGC_A and PGC_C. PGC_A and PGC_C contained the most populations, which likely explains the wide range of endemicity of phages in these clusters. Moreover, the cyanobacterial hosts that are known for many of the jumbo phages in PGC_C are widespread in the ocean, which may also explain the prevalence of this group of phages. In general, the heterogeneous distribution and abundance of these jumbo phages is consistent with the seed bank hypothesis, which postulates that viruses are passively dispersed throughout the ocean and viral community structure is shaped by local selective forces [23, 49]. This framework has previously been used to explain why phage distributions range from extremely cosmopolitan to extremely rare, which is a pattern that also appears to hold for jumbo bacteriophages.

## CONCLUSION

Large DNA viruses are becoming increasingly recognized as critical components of the virosphere, notable for their intriguing evolutionary histories [50, 51], vast functional capacities [3, 4], and global distribution [2, 4]. Here, we uncover the diversity and ecology of marine jumbo bacteriophages, which have historically been difficult to study due to biases in filtration and

isolation strategies commonly used in virology. We employed a binning strategy to generate and quality-check genomes of jumbo phages and used it to identify 85 high quality bins. We combined these bins together with reference jumbo bacteriophage genomes, and ultimately identified 120 populations that are present in Tara Oceans metagenomes. When compared with other metagenomic jumbo phages and cultured phages of all sizes, we found that marine jumbo phages primarily belong to five phage genome clusters (PGCs) that encode distinct replication machinery, biogeography, and potential hosts. For example, marine jumbo phages in PGC_C follow cyanophage infection strategies and ecology, as this cluster included cultured marine cyanophages and encoded classic family B DNA polymerases and photosynthesis enzymes characteristic of cyanophages. Furthermore, we found they are enriched in surface waters relative to the mesopelagic, consistent with the geographic range of their hosts. In contrast, marine jumbo phages of PGC_B included cultured PhiKZ phages of *Pseudomonas aeruginosa* and uniquely encoded multi-subunit RNA polymerases and tubulin, which are thought to play a role in the remarkable nucleus-like structures that these viruses employ as an anti-CRIPSR defense [17, 18]. PGC_B was particularly enriched in mesopelagic waters, suggesting that this complex infection strategy may be more common in the deep ocean. These results demonstrate the diverse biology and ecology of jumbo phages, consistent with the hypothesis that jumbo phages evolved from smaller phages on independent occasions [6].

We show that jumbo phages are widespread throughout the ocean and are typically more diverse and abundant in epipelagic waters, which reflect previous findings that surface waters usually harbor a higher per-sample alpha diversity of viral groups compared to deeper waters [9, 25]. Larger phages therefore appear to coexist in patterns broadly similar to smaller viruses despite the disadvantages of their size, such as smaller burst sizes and lower host contact rates [52]. In eukaryotic giant viruses it has been hypothesized that these disadvantages are offset by higher infection efficiency, broader host ranges, potentially decreased decay rates, and higher rates of successful attachments compared to smaller viruses [52]. Although some of these

advantages to viral gigantism may also apply to jumbo bacteriophages, it is unlikely that they are all applicable. For example, given that they are tailed *Caudovirales* [6], they likely possess higher host specificity in part due to their non-phagocytotic mode of infection. Nonetheless, the large genomes of jumbo bacteriophage often encode an expanded complement of genes used to manipulate host physiology during infection, and these may play critical roles in promoting infection efficiency or offsetting host defense mechanisms. The impressive complement of photosynthesis genes in PGC_C is at least partially responsible for the large genomes in this lineage, while the genes involved in anti-CRISPR defense found in PGC_B indicate that a host-virus arms race may be responsible for genome gigantism in this group. Interestingly, the largest number of jumbo phage genomes we identified belong to PGC_A, which is largely uncharacterized and composed of primarily metagenome-derived genomes, suggesting that these viruses have as-yet unidentified infection strategies. Overall, it is likely that the factors leading to genome gigantism in each of these genome clusters are different, given that they represent distinct lineages of *Caudovirales* that have independently evolved large genomes. Future work further characterizing the hosts of these jumbo phages and the details of their infection program, particularly in PGC_A, will therefore be critical to understanding mechanisms that underlie complexity in the virosphere and maintain diversity.

## Methods Summary

**Jumbo phage bin detection.** Metagenomic scaffolds were downloaded from 1,545 assembled by Parks et al. 2017 [30] and binned with MetaBAT2 [31] (-s 200000 --unbinned -t 32 -m 5000 --minS 75 --maxEdges 75). Bins were retained if they summed to at least 200,000 base-pairs and comprised <= 5 contigs (min. contig size 5kb). Proteins were predicted with prodigal [53] using default settings on each bin individually. Bins were retained if they lacked more than one ribosomal protein, lacked overlapping regions (via promer and gnuplot [54] with MUMmer 3.0 [55]), had fewer hits to NCLDV than phage (via LASTp [56] against RefSeq r99), lacked more

than one NCLDV marker gene (via HMM search (hmmer.org) against NCLDV marker gene HMM profiles [57]) and had even read coverage of Tara Ocean metagenomes via coverM [58] (https://github.com/wwood/CoverM). Jumbo phages were then detected by running the bins through VirSorter2 [32], VIBRANT [33], and CheckV [34]. Bins were considered putative phages if they had at least an average dsDNAphage score of >0.9 from VirSorter2 or VirSorter2 average score >0.5 and either i) CheckV quality of medium or higher (via pseudocontigs) or ii) VIBRANT consensus classification as viral. 85 bins were retained for downstream analyses. Bins were clustered into populations with a compiled set of jumbo bacteriophages (RefSeq phages over 200 kilobases, phage sequences over 200 kilobases from the INPHARED database [36], Iyer et al 2021 [6], Al-Shayeb et al. 2020 [4] jumbo phage genomes, GOV 2.0 [25], ALOHA 2.0 [37], and a megaphage assembled from the English Channel [38]) based on nucleotide sequences of genes (predicted with prodigal -d flag) aligned with BLASTn [59] (>95% average nucleotide identity, >80% genes) [23]. See Supplemental Methods for details.

**Evolutionary analyses.** Reference phage sequences were compiled from RefSeq's *Caudovirales* complete genomes (downloaded July 2020 from NCBI's Virus genome portal; 3,012 genomes) along with the curated jumbo phage set used in the population analysis. Proteins of jumbo bins and this reference set were predicted with prodigal and searched against the Virus Orthologous Groups (VOGs, vogdb.org) via HMM searches (E-value 0.001). Bipartite network was made based on shared VOGs using igraph (graph.incidence) (1.2.5) [60] in R (version 3.5.1) [61] with RStudio (version 1.1.456) [62]. Clusters were detected by the spinglass community detection algorithm [41] via igraph (50 spins for 100 iterations). Final clusters determined by iteration with highest modularity. Phages were further classified with vContact2 [39]; however, none of the jumbo phage bins belonged to clusters that could be assigned an ICTV taxa. Plots of the cluster composition from the bipartite network analysis were made with ggplot2 (3.1.1) [63] in R with Rstudio. TerL (terminase large subunit) and MCP (major capsid protein) trees were made with hits to TerL VOG families and MCP VOG families encoded by the

jumbo phages and reference hits (HMM searches, E-value < 0.001; Supplemental Dataset 2). Reference hits were de-replicated with CD-HIT (-c 0.9) [64] and filtered for size (See Supplementary Methods). Proteins were then aligned with Clustal Omega [65], trimmed with trimAl (-gt 0.1) [66] and constructed with IQ-TREE [67] (TEST model selection with ModelFinder [68]).

**Size fraction and ecological analyses.** Metagenomic reads from Tara Oceans were mapped onto the population representatives of the jumbo phage set (535 populations) with coverM (coverm genome; >10% genome covered, min ID 95%), and reads per kilobase per million (RPKM) was then calculated by dividing the number of reads aligned to a phage by the length of the phage in kilobases and then dividing that by the number of reads in the sample in millions, which accounts for differences in phage sequence length and differences in sample sequencing depth. Statistical analyses and plots were carried out in R with vegan (2.5-5) [69], ggplot2, maps (3.3.0) [70], and ggpubr (0.2.4) [71] packages. Community composition was compared between variables using ANOSIMs based on Bray-Curtis distances using both presence/absence and RPKM matrices with a significance p-value cutoff of 0.05. Statistical tests were carried out with the ggplot2 function stat_compare_means(label="p.signif").

**Annotation.** Amino acid sequences of genes were annotated with HMM searches (Evalue < 0.001) against the Pfam [72] (version 32), eggNOG [73], and VOG databases. Virion structural protein families were identified based on VOG hit descriptions (Supplemental Dataset 3). Consensus annotation was based on Pfam annotations and then the highest bit score between eggNOG and VOG hits. Functions were grouped into larger categories (Supplemental Dataset 3). Clusters from the network analyses were compared for functional composition by averaging the proportion of genes in a functional category using python (version 3) statistics and plotted with ggplot2 in R. Genes with the highest variance between clusters were identified based on the variance in proportion of genomes in a cluster with that gene. Those with over 0.2 variance and a known function were visualized with pheatmap [74] in R.

**Host prediction.** Hosts of the jumbo phage bins were predicted based on matching CRISPR spacers, tRNAs, and gene content. CRISPR spacers were predicted on the Genome Taxonomy Database (release 95) [75], metagenome assembled genomes (MAGs) of bacteria and archaea from the metagenomes that the jumbo phage bins derived (provided by Parks et al. 2017 [30]), and on the jumbo phage bins with minCED (derived from reference [76]). Spacers were aligned with BLASTn and matches were >24 bp with <= 1 mismatches [4]. tRNA sequences were predicted with tRNAscan-SE (-bacteria option) [77] on the MAGs and jumbo phage bins. Promiscuous tRNAs [78] were removed (BLASTn hits 100% ID, <= 1 mismatches). Jumbo phage tRNAs were aligned against the MAGs tRNAs and NCBI nr database (BLASTn 100% ID, <= 1 mismatches) [4]. Lastly, hosts were assigned based on the taxonomy of coding sequence matches to the MAGs (BLASTn). Hits to phyla were summed and a putative host phylum had three times the number of hits as the phylum with the next most hits as used in a previous study [4].

**Acknowledgements**

We thank members of the Aylward Lab for helpful feedback. This work was performed using compute nodes available at the Virginia Tech Advanced Research and Computing Center.

**Competing Interests**

The authors declare no competing financial interests.

**References**

1. Brüssow H, Hendrix RW. Phage Genomics: Small is Beautiful. *Cell* 2002; **108**: 13–16.
2. 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.
3. 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.

4. Al-Shayeb B, Sachdeva R, Chen L-X, Ward F, Munk P, Devoto A, et al. Clades of huge phages from across Earth's ecosystems. *Nature* 2020; **578**: 425–431.
5. Koonin EV, Yutin N. Origin and Evolution of Eukaryotic Large Nucleo-Cytoplasmic DNA Viruses. *Intervirology* . 2010; **53**: 284–292.
6. M Iyer L, Anantharaman V, Krishnan A, Burroughs AM, Aravind L. Jumbo Phages: A Comparative Genomic Overview of Core Functions and Adaptions for Biological Conflicts. *Viruses* 2021; **13**.
7. Raoult D, Forterre P. Redefining viruses: lessons from Mimivirus. *Nat Rev Microbiol* 2008; **6**: 315–319.
8. 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:** e00293-21.
9. 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.
10. 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* 2021; **24**: 102002.
11. Hendrix RW. Jumbo bacteriophages. *Curr Top Microbiol Immunol* 2009; **328**: 229–240.
12. Low SJ, Džunková M, Chaumeil P-A, Parks DH, Hugenholtz P. Evaluation of a concatenated protein phylogeny for classification of tailed double-stranded DNA viruses belonging to the order Caudovirales. *Nat Microbiol* 2019; **4**: 1306–1315.
13. Krylov VN, Zhazykov IZ. Pseudomonas bacteriophage phiKZ--possible model for studying the genetic control of morphogenesis. *Genetika* 1978; **14**: 678–685.
14. Yuan Y, Gao M. Jumbo Bacteriophages: An Overview. *Frontiers in Microbiology* 2017; **8**:

403.

15. Bertani B, Ruiz N. Function and Biogenesis of Lipopolysaccharides. *EcoSal Plus* 2018; **8**: https://doi.org/10.1128/ecosalplus.ESP-0001-2018.
16. Ceyssens P-J, Minakhin L, Van den Bossche A, Yakunina M, Klimuk E, Blasdel B, et al. Development of Giant Bacteriophage KZ Is Independent of the Host Transcription Apparatus. *Journal of Virology*. 2014; **88**: 10501–10510.
17. Mendoza SD, Nieweglowska ES, Govindarajan S, Leon LM, Berry JD, Tiwari A, et al. A bacteriophage nucleus-like compartment shields DNA from CRISPR nucleases. *Nature* 2020; **577**: 244–248.
18. Malone LM, Warring SL, Jackson SA, Warnecke C, Gardner PP, Gumy LF, et al. A jumbo phage that forms a nucleus-like structure evades CRISPR–Cas DNA targeting but is vulnerable to type III RNA-based immunity. *Nature Microbiology* . 2020; **5**: 48–55.
19. Serwer P, Hayes SJ, Thomas JA, Hardies SC. Propagating the missing bacteriophages: a large bacteriophage in a new class. *Virol J* 2007; **4**: 21.
20. Thurber RV, Haynes M, Breitbart M, Wegley L, Rohwer F. Laboratory procedures to generate viral metagenomes. *Nat Protoc* 2009; **4**: 470–483.
21. Palermo CN, Shea DW, Short SM. Analysis of Different Size Fractions Provides a More Complete Perspective of Viral Diversity in a Freshwater Embayment. *Appl Environ Microbiol* 2021; **87**: e00197-21.
22. Sieburth JM, Smetacek V, Lenz J. Pelagic ecosystem structure: Heterotrophic compartments of the plankton and their relationship to plankton size fractions 1. *Limnology and Oceanography* 1978; **23**: 1256–1263.
23. Brum JR, Ignacio-Espinoza JC, Roux S, Doulcier G, Acinas SG, Alberti A, et al. Ocean plankton. Patterns and ecological drivers of ocean viral communities. *Science* 2015; **348**: 1261498.
24. Roux S, Brum JR, Dutilh BE, Sunagawa S, Duhaime MB, Loy A, et al. Ecogenomics and

potential biogeochemical impacts of globally abundant ocean viruses. *Nature* 2016; **537**: 689–693.

25. Gregory AC, Zayed AA, Conceição-Neto N, Temperton B, Bolduc B, Alberti A, et al. Marine DNA Viral Macro- and Microdiversity from Pole to Pole. *Cell* 2019; **177**: 1109–1123.e14.
26. Roux S, Emerson JB, Eloe-Fadrosh EA, Sullivan MB. Benchmarking viromics: an evaluation of metagenome-enabled estimates of viral community composition and diversity. *PeerJ* 2017; **5**: e3817.
27. García-López R, Vázquez-Castellanos JF, Moya A. Fragmentation and Coverage Variation in Viral Metagenome Assemblies, and Their Effect in Diversity Calculations. *Front Bioeng Biotechnol* 2015; **3**: 141.
28. Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, et al. Ocean plankton. Structure and function of the global ocean microbiome. *Science* 2015; **348**: 1261359.
29. Sunagawa S, Coordinators TO, Acinas SG, Bork P, Bowler C, Eveillard D, et al. Tara Oceans: towards global ocean ecosystems biology. *Nature Reviews Microbiology* 2020; **18**: 428–445.
30. Parks DH, Rinke C, Chuvochina M, Chaumeil P-A, Woodcroft BJ, Evans PN, et al. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat Microbiol* 2017; **2**: 1533–1542.
31. Kang DD, Li F, Kirton E, Thomas A, Egan R, An H, et al. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* 2019; **7**: e7359.
32. Guo J, Bolduc B, Zayed AA, Varsani A, Dominguez-Huerta G, Delmont TO, et al. VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. *Microbiome* 2021; **9**: 37.
33. Kieft K, Zhou Z, Anantharaman K. VIBRANT: automated recovery, annotation and curation

of microbial viruses, and evaluation of viral community function from genomic sequences. *Microbiome* 2020; **8**: 90.

34. Nayfach S, Camargo AP, Schulz F, Eloe-Fadrosh E, Roux S, Kyrpides NC. CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nat Biotechnol* 2021; **39**: 578–585.
35. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome research* 2015; **7**: 1043-1055.
36. Cook R, Brown N, Redgwell T, Rihtman B, Barnes M, Clokie M, et al. INfrastructure for a PHAge REference Database: Identification of Large-Scale Biases in the Current Collection of Cultured Phage Genomes. *PHAGE: Therapy, Applications, and Research* 2021; e-pub ahead of print 5 October 2021; http://doi.org/10.1089/phage.2021.0007
37. Luo E, Eppley JM, Romano AE, Mende DR, DeLong EF. Double-stranded DNA virioplankton dynamics and reproductive strategies in the oligotrophic open ocean water column. *ISME J* 2020; **14**: 1304–1315.
38. Michniewski S, Rihtman B, Cook R, Jones MA, Wilson WH, Scanlan DJ, et al. Identification of a new family of ‘megaphages’ that are abundant in the marine environment.
39. Jang HB, Bolduc B, Zablocki O, Kuhn JH, Roux S, Adriaenssens EM, et al. Gene sharing networks to automate genome-based prokaryotic viral taxonomy. *Nat Biotechnology* 2019; **37**: doi:10.1101/533240.
40. Iranzo J, Krupovic M, Koonin EV. The Double-Stranded DNA Virosphere as a Modular Hierarchical Network of Gene Sharing. *MBio* 2016; **7**: e00978-16.
41. Reichardt J, Bornholdt S. Statistical mechanics of community detection. *Phys Rev E Stat Nonlin Soft Matter Phys* 2006; **74**: 016110.
42. Hurwitz BL, Sullivan MB. The Pacific Ocean virome (POV): a marine viral metagenomic dataset and associated protein clusters for quantitative viral ecology. *PLoS One* 2013; **8**:

e57355.

43. Pappas N, Dutilh BE. Finding functional associations between prokaryotic virus orthologous groups: a proof of concept. *BMC Bioinformatics* 2021; **22**: 1–11.

44. Fridman S, Flores-Uribe J, Larom S, Alalouf O, Liran O, Yacoby I, et al. A myovirus encoding both photosystem I and II proteins enhances cyclic electron flow in infected *Prochlorococcus* cells. *Nat Microbiol* 2017; **2**: 1350–1357.

45. Hjorleifsdottir S, Aevarsson A, Hreggvidsson GO, Fridjonsson OH, Kristjansson JK. Isolation, growth and genome of the *Rhodothermus* RM378 thermophilic bacteriophage. *Extremophiles* 2014; **18**: 261–270.

46. Mende DR, Bryant JA, Aylward FO, Eppley JM, Nielsen T, Karl DM, et al. Environmental drivers of a microbial genomic transition zone in the ocean's interior. *Nat Microbiol* 2017; **2**: 1367–1373.

47. Goldsmith DB, Brum JR, Hopkins M, Carlson CA, Breitbart M. Water column stratification structures viral community composition in the Sargasso Sea. *Aquatic Microbial Ecology* . 2015; **76**: 85–94.

48. Hurwitz BL, Brum JR, Sullivan MB. Depth-stratified functional and taxonomic niche specialization in the 'core' and 'flexible' Pacific Ocean Virome. *The ISME Journal* 2015; **9**: 472–484.

49. Breitbart M, Rohwer F. Here a virus, there a virus, everywhere the same virus? *Trends in Microbiology* 2005; **13**: 278–284.

50. Brüssow H. Huge bacteriophages: bridging the gap? *Environmental Microbiology* 2020; **22**: 1965–1970.

51. 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.

52. Edwards KF, Steward GF, Schvarcz CR. Making sense of virus size and the tradeoffs shaping viral fitness. *Ecol Lett* 2021; **24**: 363–373.

53. Hyatt D, Chen G-L, Locascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 2010; **11**: 119.
54. Williams T, Kelley C. Gnuplot 5.2 Manual: An Interactive Plotting Program. 2017.
55. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, et al. Versatile and open software for comparing large genomes. *Genome Biol* 2004; **5**: 1-9.
56. Kiełbasa SM, Wan R, Sato K, Horton P, Frith MC. Adaptive seeds tame genomic sequence comparison. *Genome Res* 2011; **21**: 487–493.
57. Aylward FO, Moniruzzaman M. ViralRecall-A Flexible Command-Line Tool for the Detection of Giant Virus Signatures in ’Omic Data. *Viruses* 2021; **13**.
58. wwood. GitHub - wwood/CoverM: Read coverage calculator for metagenomics. https://github.com/wwood/CoverM. Accessed 23 Jul 2021.
59. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol* 1990; **215**: 403–410.
60. igraph – Network analysis software. http://igraph.org. Accessed 23 Jul 2021.
61. R Core Team. R: A Language and Environment for Statistical Computing. 2019. R Foundation for Statistical Computing, Vienna, Austria.
62. RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/.
63. Wickham H. ggplot2. *Wiley Interdisciplinary Reviews: Computational Statistics* 2011; **3**: 180–185.
64. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 2012; **28**: 3150–3152.
65. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol* 2011; **7**: 539.

66. 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.
67. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* 2015; **32**: 268–274.
68. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods* 2017; **14**: 587–589.
69. Dixon P. VEGAN, a package of R functions for community ecology. *Journal of Vegetation Science* 2003; **14**: 927–930.
70. maps: Draw Geographical Maps. https://CRAN.R-project.org/package=maps. Accessed 22 Jul 2021.
71. Kassambara A. 'ggplot2' Based Publication Ready Plots [R package ggpubr version 0.4.0]. 2020.
72. 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.
73. 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 Research* 2019; **47**: D309–D314.
74. pheatmap: Pretty Heatmaps. https://CRAN.R-project.org/package=pheatmap. Accessed 22 Sep 2021.
75. Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil P-A, et al. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat Biotechnol* 2018; **36**: 996–1004.
76. Bland C, Ramsey TL, Sabree F, Lowe M, Brown K, Kyrpides NC, et al. CRISPR Recognition Tool (CRT): a tool for automatic detection of clustered regularly interspaced

palindromic repeats. *BMC Bioinformatics* 2007; **8**: 1-8.

77. Lowe TM, Chan PP. tRNAscan-SE On-line: integrating search and context for analysis of transfer RNA genes. *Nucleic Acids Res* 2016; **44**: W54–7.

78. Paez-Espino D, Eloe-Fadrosh EA, Pavlopoulos GA, Thomas AD, Huntemann M, Mikhailova N, et al. Uncovering earth's virome. *Nature* 2016; **536**: 425–430.

## Figure Legends

**Figure 1. Comparison of bins to reference gene content and evolutionary histories. (a)** Network with marine jumbos and references as nodes and edges based on shared VOGs. Marine jumbo phage nodes are colored by PGC as detected with spinglass community detection analysis, other nodes are in gray. (**b)** Barchart of the number of members in each PGC. PGCs with marine jumbo phages are denoted with a star and the number of marine jumbo phages in that PGC. Proportion of marine jumbo phages in that PGC is colored. (**c,d)** Phylogenies of TerL (**c**) and MCP (**d**) proteins with references and bins. Inner ring and branches are colored by the 5 PGCs that marine jumbo phages belong to. Navy blue circles in the outer ring denote marine jumbo phages.

**Figure 2. (a)** Boxplot of genome length in each network cluster (x-axis is PGC number). Star denotes PGC with a marine jumbo phage and the color matches the PGC letters of Figure 1. **(b)** Stacked barplot of the metagenome environment from which each phage derives from in each PGC (x-axis). Reference (yellow) are cultured phages, black is the bins of jumbo phages from this study **(c)** Stacked barplot of the host phylum of the RefSeq cultured phages in each cluster; metagenomic phages are in gray.

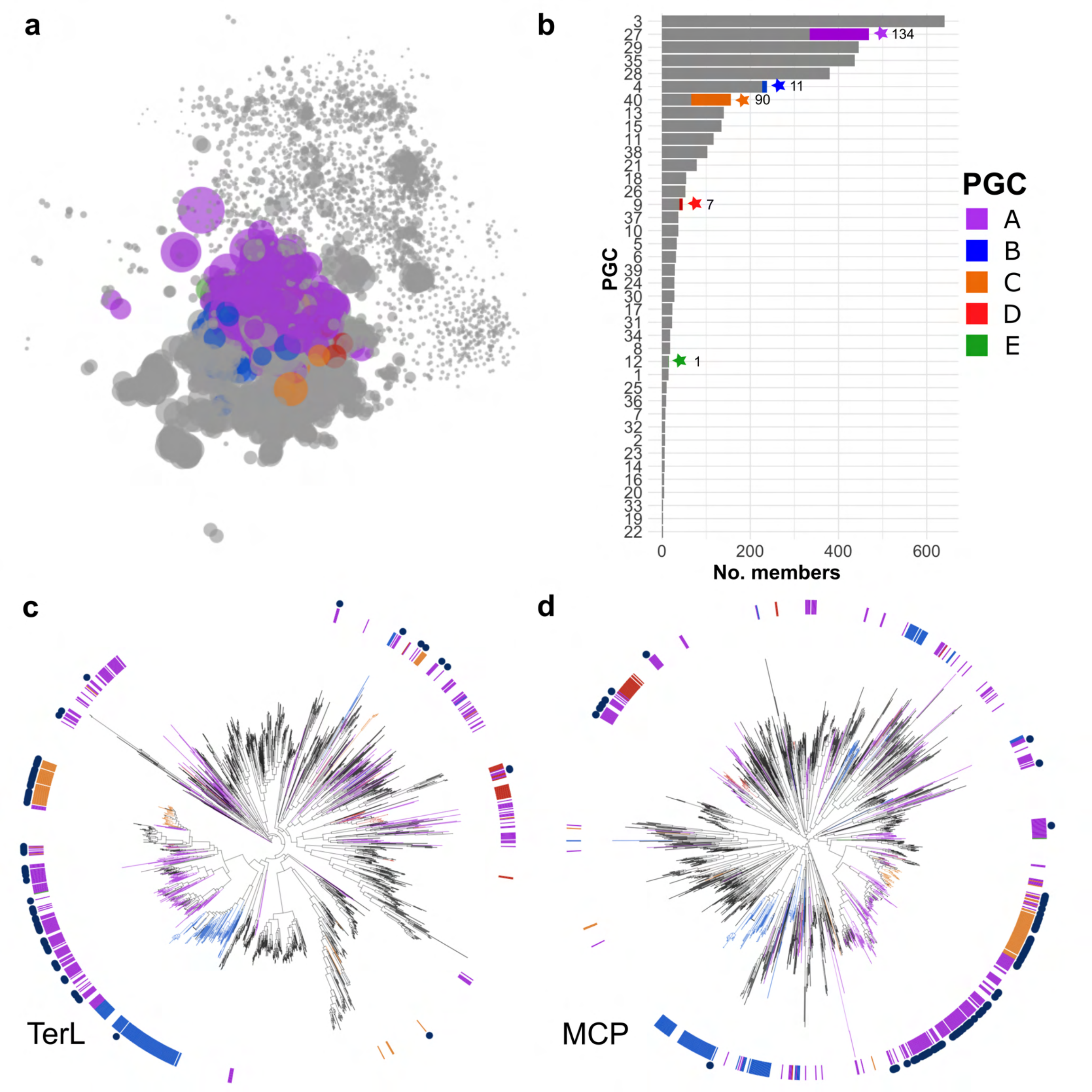
**Figure 3** (**a**) Functional categories for genes encoded by jumbo phages. (**b**) Heatmap of proportion of genomes in each cluster that contain the listed genes. Listed genes were selected

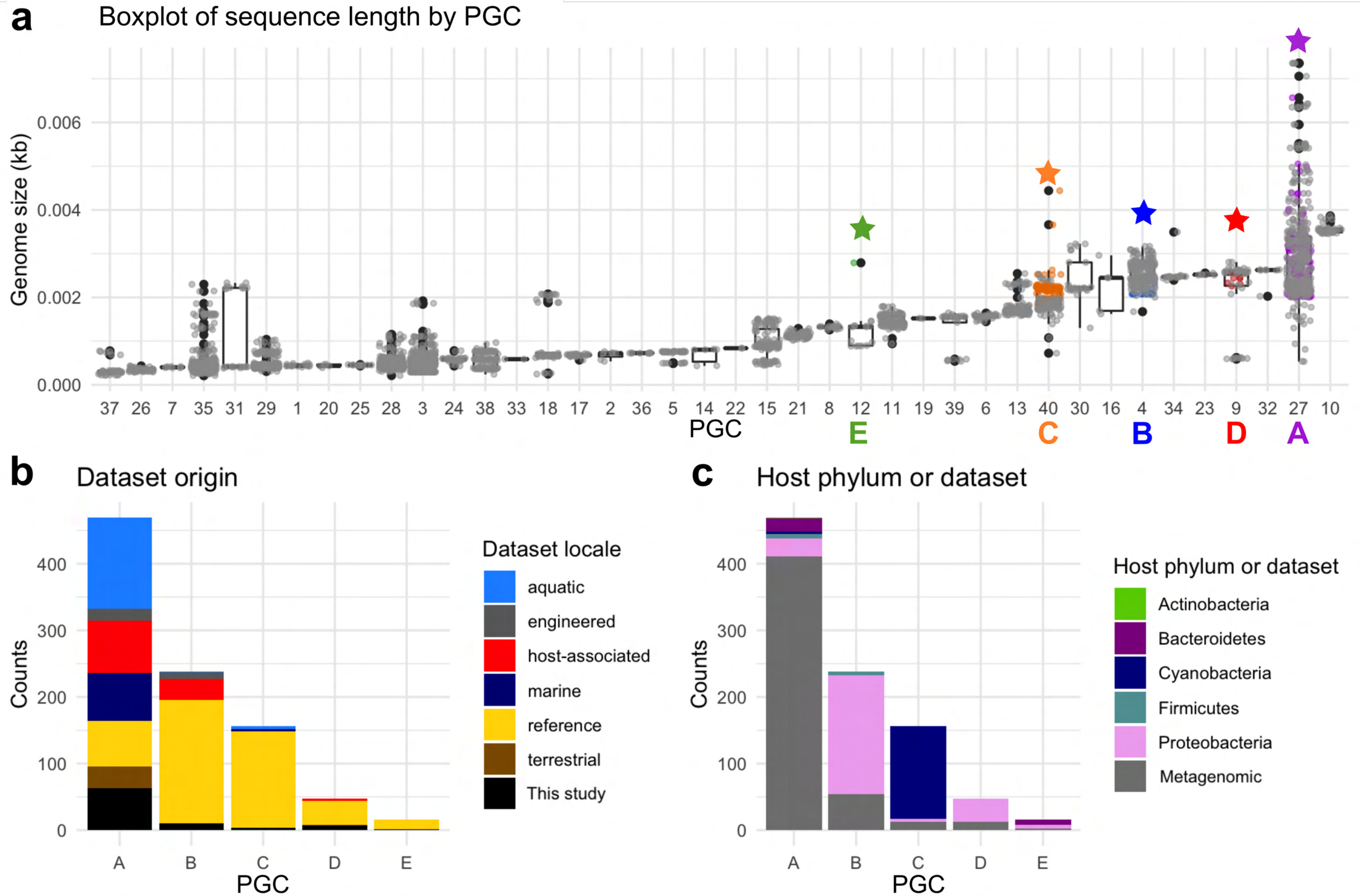
based on containing a known function and variance between the clusters above 0.2. Dendrogram was generated based on hierarchical clustering in pheatmap.

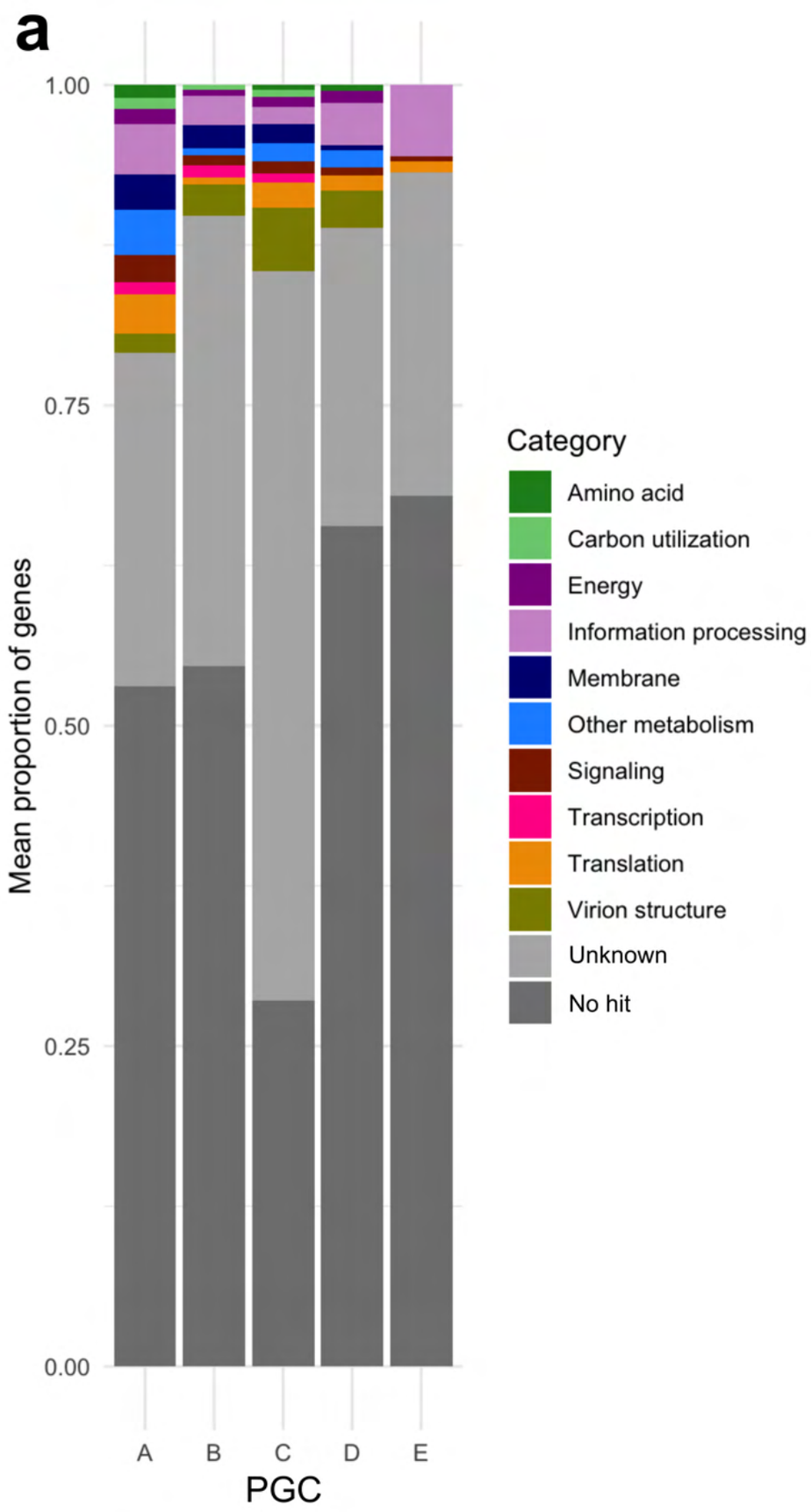
**Figure 4. Comparison of jumbo abundance and presence in samples of different filter size fractions**. Dark teal are fractions with minimum sizes of 0.22 µm or higher. Light teal are fractions with a maximum size of 0.22 µm or lower. **a)** Barchart of the proportion of samples with at least one marine jumbo phage (x-axis) by size fraction (y-axis) sorted from highest to lowest. **b)** Boxplot with x-axis as the number of marine jumbo phages found in a sample with size fraction on the y-axis sorted by median. **c)** Boxplot with x-axis as the relative abundance of marine jumbo phages found in a sample (RPKM) with size fraction on the y-axis sorted by median. Significance bars in c,d correspond to Wilcox test, with stars corresponding to p-values < 0.05 and those with p-values > 0.05 as not significant "ns" (stat_compare_means function). (**d,e**) NMDS plots (Bray Curtis dissimilarity distances) of jumbo phage composition in each sample using presence absence data (**d**) and relative abundance data (**e**). Samples are colored by size fraction distinction above 0.22 (dark teal) and below 0.22 (light teal). Ellipses calculated based on multivariate normal distribution. Samples above and below 0.22 were not significantly different from each other using ANOSIM (p-values > 0.05).

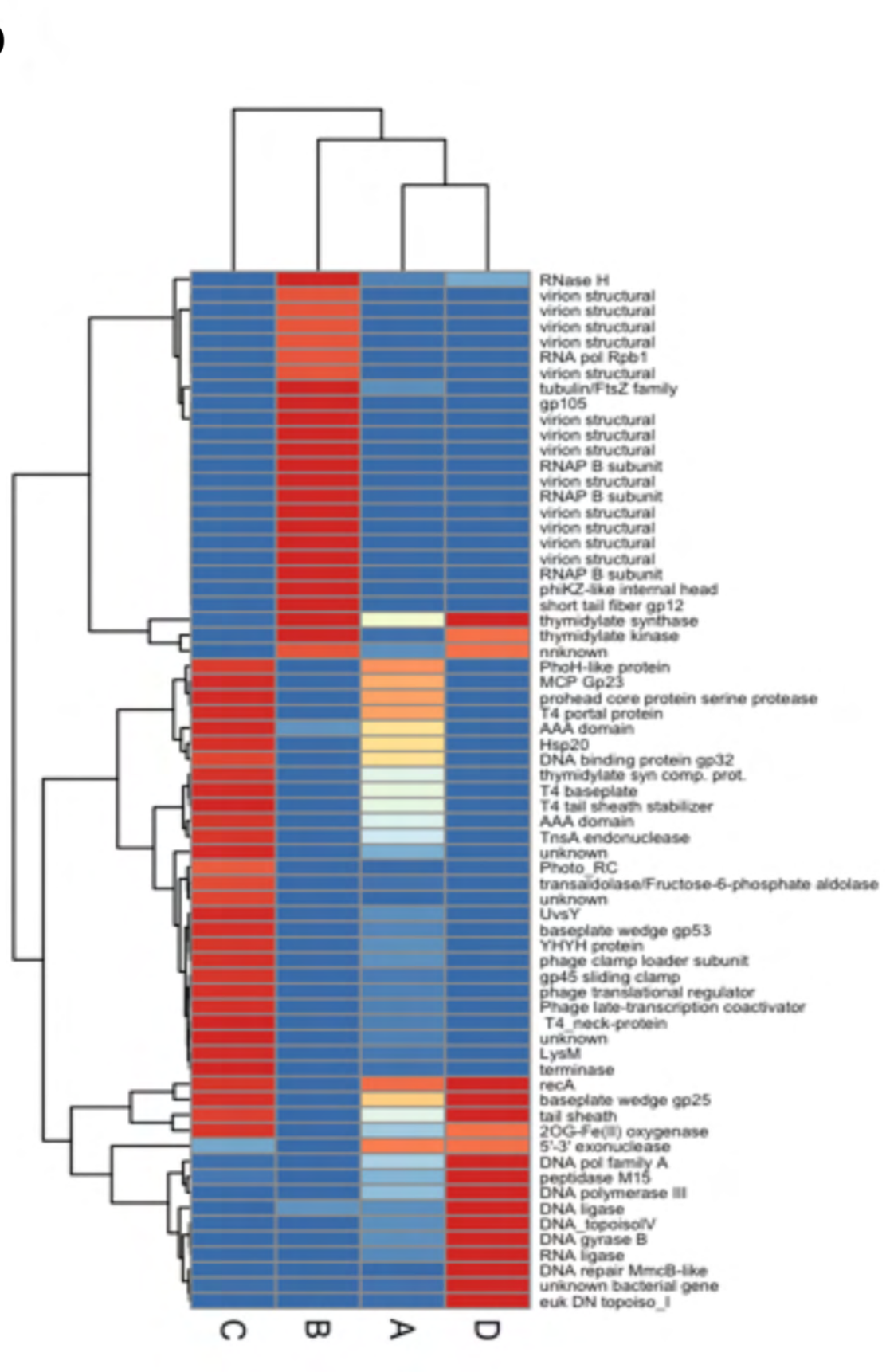
**Figure 5. a,b**) Maps of the relative abundance (**a**) of jumbo phages (in RPKM) and (**b**) number of jumbo populations present in each surface (SRF) sample of the picoplankton size fraction (either 0.22-3µm or 0.22-1.6µm depending on availability). Dots sizes are proportional to the number of populations or RPKM and colored by biome (Coastal - pink, Westerlies - purple, Trades - blue). (**c**) Scatterplot of the mean RPKM of a jumbo population in SRF picoplankton samples versus the number of SRF picoplankton stations it was present. Populations are colored by PGC and size corresponds to putative genome length in 100 kilobases. (**d**) Boxplot of the number of jumbo phage populations in a sample separated by depth sorted by median for

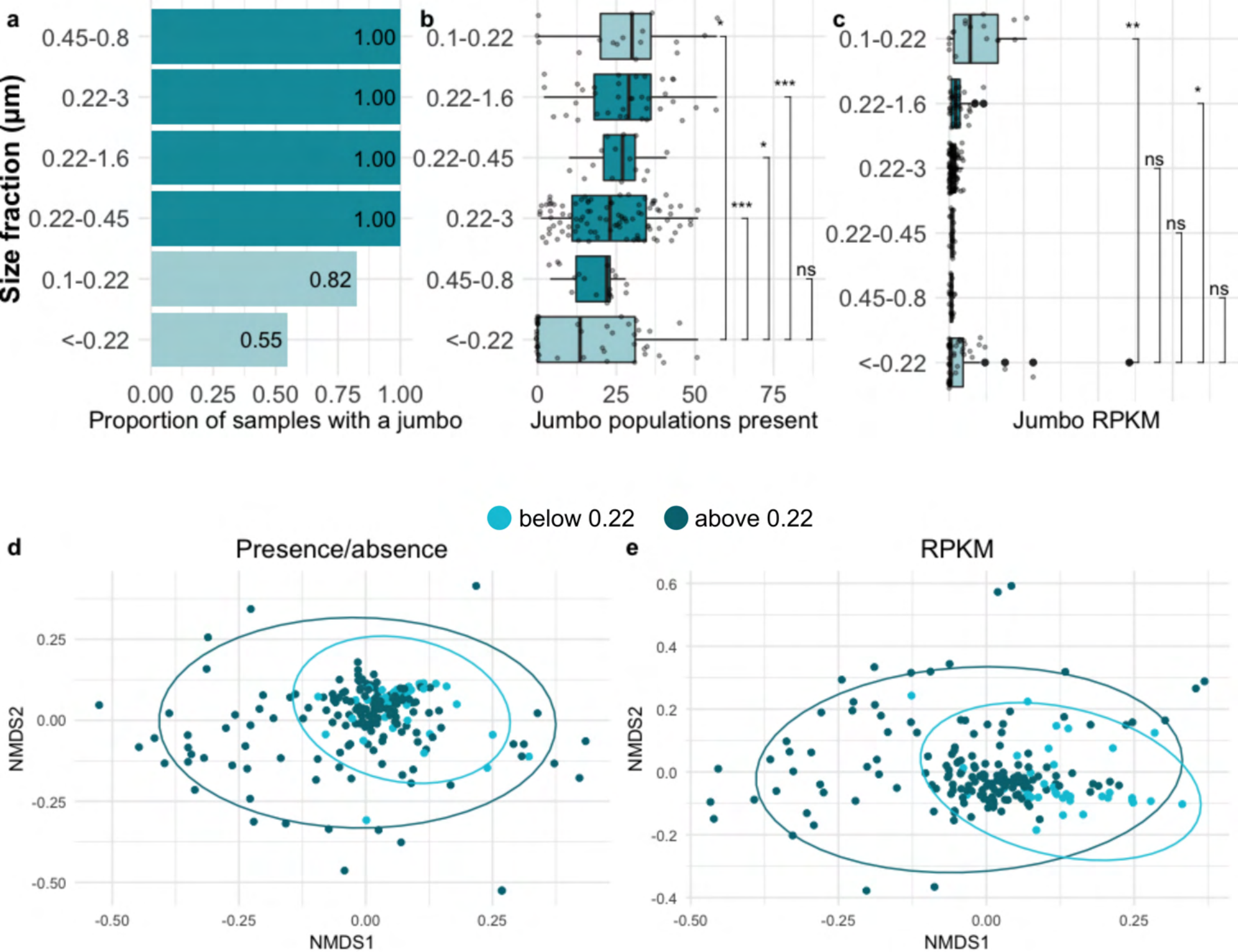
each PGC. Significance bars correspond to Wilcox test, with stars corresponding to p-values < 0.05 and those with pvalues > 0.05 as not significant (stat_compare_means function)

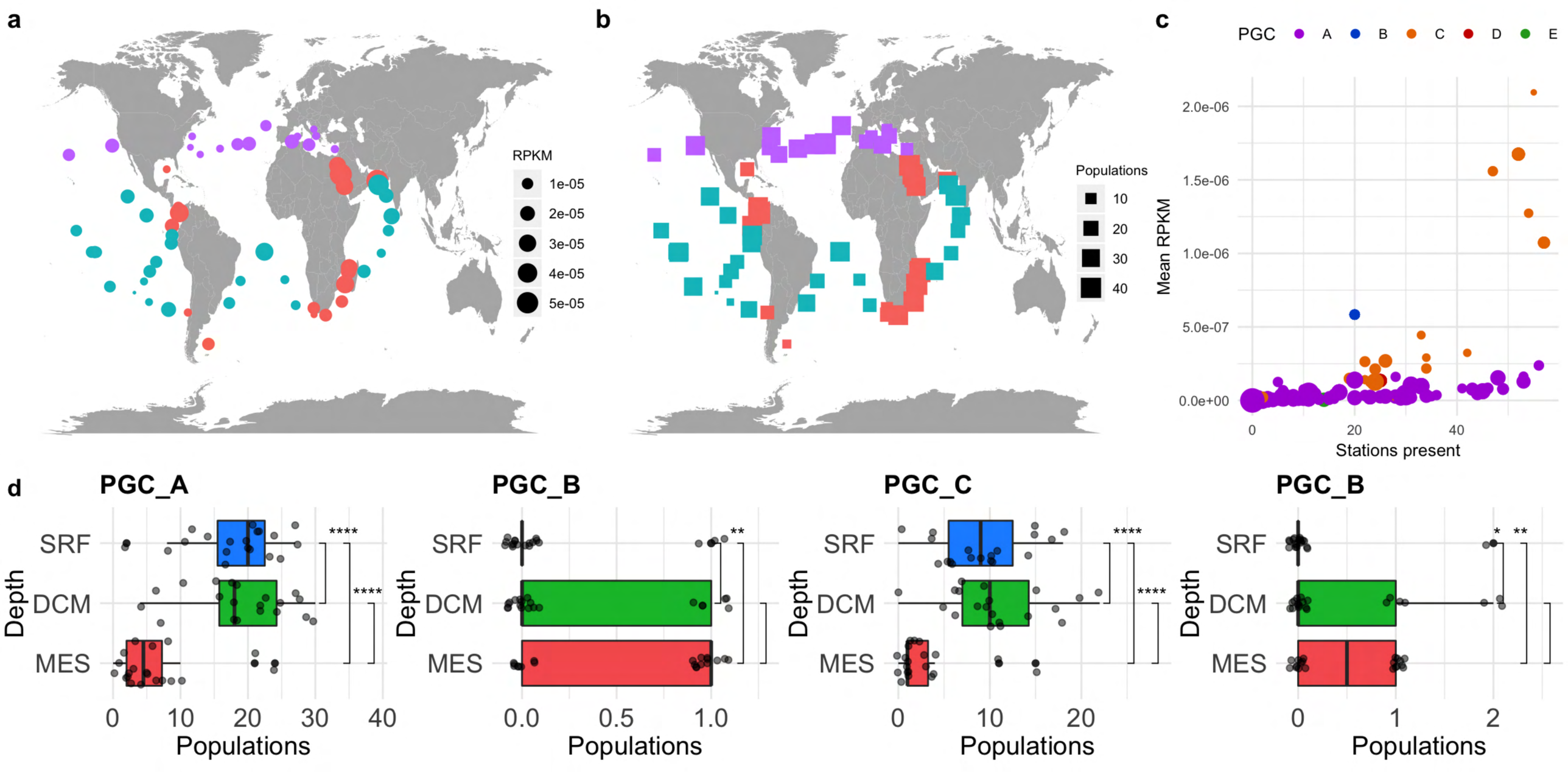

a
b
PGC
No. members
0
200
400
600
3
27
29
35
28
4
40
13
15
11
38
21
18
26
9
37
10
5
6
39
24
30
17
31
34
8
12
1
25
36
7
32
2
23
14
16
20
33
19
22
134
11
90
7
1
PGC
A
B
C
D
E
c
TerL
d
MCP


a
Boxplot of sequence length by PGC
Genome size (kb)
0.006
0.004
0.002
0.000
37 26 7 35 31 29 1 20 25 28 3 24 38 33 18 17 2 36 5 14 22 15 21 8 12 11 19 39 6 13 40 30 16 4 34 23 9 32 27 10
PGC
E
C
B
D
A
b
Dataset origin
Counts
400
300
200
100
0
A B C D E
PGC
Dataset locale
aquatic
engineered
host-associated
marine
reference
terrestrial
This study
c
Host phylum or dataset
Counts
400
300
200
100
0
A B C D E
PGC
Host phylum or dataset
Actinobacteria
Bacteroidetes
Cyanobacteria
Firmicutes
Proteobacteria
Metagenomic


a
1.00
0.75
0.50
0.25
0.00
Mean proportion of genes
A
B
C
D
E
PGC
Category
Amino acid
Carbon utilization
Energy
Information processing
Membrane
Other metabolism
Signaling
Transcription
Translation
Virion structure
Unknown
No hit


b
RNase H
virion structural
virion structural
virion structural
virion structural
RNA pol Rpb1
virion structural
tubulin/FtsZ family
gp105
virion structural
virion structural
virion structural
RNAP B subunit
virion structural
RNAP B subunit
virion structural
virion structural
virion structural
virion structural
RNAP B subunit
phiKZ-like internal head
short tail fiber gp12
thymidylate synthase
thymidylate kinase
nnknown
PhoH-like protein
MCP Gp23
prohead core protein serine protease
T4 portal protein
AAA domain
Hsp20
DNA binding protein gp32
thymidylate syn comp. prot.
T4 baseplate
T4 tail sheath stabilizer
AAA domain
TnsA endonuclease
unknown
Photo_RC
transaldolase/Fructose-6-phosphate aldolase
unknown
UvsY
baseplate wedge gp53
YHYH protein
phage clamp loader subunit
gp45 sliding clamp
phage translational regulator
Phage late-transcription coactivator
T4_neck-protein
unknown
LysM
terminase
recA
baseplate wedge gp25
tail sheath
2OG-Fe(II) oxygenase
5'-3' exonuclease
DNA pol family A
peptidase M15
DNA polymerase III
DNA ligase
DNA_topoisoIV
DNA gyrase B
RNA ligase
DNA repair MmcB-like
unknown bacterial gene
euk DN topoiso_I
C
B
A
D


a
Size fraction (µm)
0.45-0.8
0.22-3
0.22-1.6
0.22-0.45
0.1-0.22
<-0.22
1.00
1.00
1.00
1.00
0.82
0.55
0.00 0.25 0.50 0.75 1.00
Proportion of samples with a jumbo
b
0.1-0.22
0.22-1.6
0.22-0.45
0.22-3
0.45-0.8
<-0.22
*
***
*
***
ns
0 25 50 75
Jumbo populations present
c
0.1-0.22
0.22-1.6
0.22-3
0.22-0.45
0.45-0.8
<-0.22
**
*
ns
ns
ns
Jumbo RPKM
below 0.22
above 0.22
d
Presence/absence
NMDS2
0.25
0.00
-0.25
-0.50
-0.50 -0.25 0.00 0.25
NMDS1
e
RPKM
NMDS2
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.50 -0.25 0.00 0.25
NMDS1


a
RPKM
1e-05
2e-05
3e-05
4e-05
5e-05
b
Populations
10
20
30
40
c
PGC
A
B
C
D
E
Mean RPKM
2.0e-06
1.5e-06
1.0e-06
5.0e-07
0.0e+00
0
20
40
Stations present
d
PGC_A
PGC_B
PGC_C
PGC_B
Depth
SRF
DCM
MES
****
**
*
Populations