Marine DNA viral *macro-* and *micro-*diversity from pole to pole

Ann C Gregory[1,†], Ahmed A Zayed[1,†], Nádia Conceição-Neto[2,3], Ben Temperton[4], Ben Bolduc[1], Adriana Alberti[5,17], Mathieu Ardyna[6,¥], Ksenia Arkhipova[7], Margaux Carmichael[8,17], Corinne Cruaud[9,17], Céline Dimier[6,10,17], Guillermo Domínguez-Huerta[1], Joannie Ferland[11], Stefanie Kandels-Lewis[12,13], Yunxiao Liu[1], Claudie Marec[11], Stéphane Pesant[14,15], Marc Picheral[6,17], Sergey Pisarev[16], Julie Poulain[5,17], Jean-Éric Tremblay[11], Dean Vik[1], *Tara* Oceans coordinators[§], Marcel Babin[11], Chris Bowler[10,17], Alexander I Culley[18], Colomban de Vargas[8,17], Bas E Dutilh[7,19], Daniele Iudicone[20], Lee Karp-Boss[21], Simon Roux[1,‡], Shinichi Sunagawa[22], Patrick Wincker[5,17], & Matthew B Sullivan[1,23,*]

Affiliations:

[1]Department of Microbiology, The Ohio State University, Columbus, Ohio 43210, USA.

[2]Department of Microbiology and Immunology, Rega Institute for Medical Research, Laboratory of Viral Metagenomics, KU Leuven - University of Leuven, Leuven, Belgium.

[3]Department of Microbiology and Immunology, Rega Institute for Medical Research, Laboratory for Clinical and Epidemiological Virology, KU Leuven - University of Leuven, Leuven, Belgium.

[4]School of Biosciences, University of Exeter, Exeter, UK.

[5]Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ Evry, Université Paris-Saclay, 91057 Evry, France.

[6]Sorbonne Université, CNRS, Laboratoire d'Océanographie de Villefanche, LOV, F-06230 Villefranche-sur-mer, France

[7]Theoretical Biology and Bioinformatics, Utrecht University, Utrecht, Netherlands.

[8]Sorbonne Université, CNRS, Station Biologique de Roscoff, AD2M ECOMAP, 29680 Roscoff, France.

[9]CEA - Institut de Biologie François Jacob, Genoscope, Evry, 91057, France.

[10]Institut de Biologie de l'Ecole Normale Supérieure (IBENS), Ecole normale supérieure, CNRS, INSERM, Université PSL, 75005 Paris, France.

[11]Département de biologie, Québec Océan and Takuvik Joint International Laboratory (UMI 3376), Université Laval (Canada) - CNRS (France), Université Laval, Québec, QC, G1V 0A6, Canada.

[12]Structural and Computational Biology, European Molecular Biology Laboratory, 69117 Heidelberg, Germany.

[13]Directors' Research, European Molecular Biology Laboratory, 69117 Heidelberg, Germany.

[14]PANGAEA, Data Publisher for Earth and Environmental Science, University of Bremen, 28359 Bremen, Germany.

[15]MARUM, Bremen University, 28359 Bremen, Germany.

[16]Shirshov Institute of Oceanology of Russian Academy of Sciences, 36 Nakhimovsky prosp, 117997, Moscow, Russia.

[17]Research Federation for the study of Global Ocean Systems Ecology and Evolution, FR2022/*Tara* Oceans GOSEE, 3 rue Michel-Ange, 75016 Paris, France.

[18]Département de biochimie, microbiologie et bio-informatique, Université Laval, Québec, QC, G1V 0A6, Canada.

[19]Centre for Molecular and Biomolecular Informatics, Radboud University Medical Centre, Nijmegen, Netherlands.

[20]Stazione Zoologica Anton Dohrn, Villa Comunale, 80121 Naples, Italy.

[21]School of Marine Sciences, University of Maine, Orono, ME, USA.

[22]Institute of Microbiology, ETH Zurich, Zurich, Switzerland.

[23]Department of Civil, Environmental and Geodetic Engineering, The Ohio State University, Columbus, Ohio 43210, USA.

* Corresponding author. Email: mbsulli@gmail.com

† Equal contributions.

§*Tara* Oceans coordinators and affiliations are listed in the supplementary materials.

¥Present address: Department of Earth System Science, Stanford University, Stanford, CA, 94305, USA.

‡Present address: Department of Energy Joint Genome Institute, Walnut Creek, CA, 94598, USA.

Summary: Microbes drive most ecosystems and are modulated by viruses that impact their lifespan, gene flow and metabolic outputs. However, ecosystem-level impacts of viral community diversity remains difficult to assess due to classification issues and few reference genomes. Here we establish a ~12-fold expanded global ocean DNA virome dataset of 195,728 viral populations, now including the Arctic Ocean, and validate that these populations form discrete genotypic clusters. Meta-community analyses revealed five ecological zones throughout the global ocean, including two distinct Arctic regions. Across the zones, local and global patterns and drivers in viral community diversity were established for both *macro*diversity (*inter*-population diversity) and *micro*diversity (*intra*-population genetic variation). These patterns sometimes, but not always, paralleled those from macro-organisms and revealed temperate and tropical surface waters and the Arctic as biodiversity hotspots and mechanistic hypotheses to explain them. Such further understanding of ocean viruses is critical for broader inclusion in ecosystem models.

Introduction:

Biodiversity is essential for maintaining ecosystem functions and services (reviewed by Tilman *et al.*, 2014). In the oceans, the vast majority of biodiversity is contained within the microbial fraction containing prokaryotes and eukaryotic microbes, which represents ~60% of its biomass (Bar-On *et al.*, 2018). Meta-analyses looking at changes in marine biodiversity show that biodiversity loss increasingly impairs the ocean's capacity to produce food, maintain water quality, and recover from perturbations (Worm *et al.*, 2006). To date, marine conservation efforts have focused on specific organismal communities, such as fisheries or coral reefs, rather than conserving whole ecosystem biodiversity. However, emerging studies across diverse

environments show that the stability and diversity of higher trophic level organisms rely upon diversity throughout the food web (e.g. Soliveres *et al.*, 2016). Despite being the foundation of the food web, most marine microbial biodiversity numbers are based on a few well-studied locations (e.g., Hawaii Ocean Time Series, Bermuda Atlantic Time Series, and San Pedro Ocean Time Series). For ocean microbes and their viruses, global surveys that parallel century-old global terrestrial and decades-old marine macro-organismal global biodiversity surveys (Reiners *et al.*, 2017) are only now emerging (e.g. de Vargas *et al.*, 2015; Sunagawa *et al.*, 2015; Brum *et al.*, 2015; Roux *et al.*, 2016; Ser-Giacomi *et al.*, 2018;**Table S1**). Key to assessing biodiversity changes across marine ecosystems is improving our understanding of current microbial biodiversity levels, distribution patterns, and their ecological drivers.

Despite their tiny size, viruses play a large role in marine ecosystems and food webs. For example, mortality due to viruses is credited with lysing approximately 20-40% of bacteria per day and releasing carbon and other nutrients that impact the food web (reviewed by Suttle, 2007). Beyond mortality, viruses can alter evolutionary trajectories of microbial communities by transferring ~10^{29} genes per day globally (Paul, 1999) and biogeochemical cycling by metabolically reprogramming host photosynthesis, as well as central carbon metabolism and nitrogen and sulfur cycling (reviewed in Hurwitz & U'Ren, 2016). Finally, as the oceans are estimated to capture half of human-caused carbon emissions (Le Quéré *et al.*, 2018), it is notable that genes-to-ecosystems modeling has placed viruses as central players of the ocean 'biological pump' (Guidi *et al.*, 2016). Many of these discoveries are very recent as ocean viral genome sequence space is just now being explored at the level of viral *macro*diversity, i.e., *inter*-population diversity, throughout the global oceans -- at least for the most abundant double-stranded DNA viruses sampled (**Table S2**).

In spite of this progress in studying marine viral *macro*diversity, virtually nothing is known about *micro*diversity, i.e., *intra*-population genetic variation. This is due to the controversy surrounding the existence of viral species (Gregory *et al.*, 2016; Bobay *et al.*, 2018). In eukaryotic organisms, where species boundaries are more widely accepted, such *micro*diversity has been studied and is thought to drive adaptation and speciation to promote and maintain stability in ecosystems (Hughes *et al.*, 2008; Larkin & Martiny, 2017). This is likely also true in viruses since even a few mutations can alter host interactions and ecological and evolutionary dynamics for the genotype (e.g. Marston *et al.*, 2012; Petrie *et al.*, 2018). In nature, viral *micro*diversity measurements have been limited to marker genes (e.g. genes encoding major capsid proteins), which capture neither community-wide variability (Sullivan 2015) nor genome-wide evidence of selection (e.g. Achtman & Wagner 2008). Recently, deeper metagenomic sequencing and population genetic theory-grounded species delimitations (Shapiro *et al.*, 2012; Cadillo-Quiroz *et al.*, 2012) have begun to reveal such *micro*diversity in microbes, and this has elucidated unknown features of speciation, adaptation, pathogenicity and transmission (e.g. Snitkin *et al.*, 2011; Schloissnig *et al.*, 2013; Rosen *et al.*, 2015; Lee *et al.*, 2017; Smillie *et al.*, 2018). Although parallel species delimitations are now available for viruses (Gregory *et al.*, 2016; Bobay *et al.*, 2018), no datasets are yet available to explore genome-wide *micro*diversity in viruses, particularly at the global scale.

Here we leverage the *Tara* Oceans global oceanographic research expedition sampling to establish a deeply-sequenced, global-scale ocean virome dataset and use it to assess the validity of the current viral population definition and to establish and explore baseline *macro*- and *micro*-diversity patterns with their associated drivers across local to global scales. These data have been collected and analyzed in the context of the larger *Tara* Oceans Consortium systematically-

sampled, global-scale, viruses-to-fish-larvae datasets (de Vargas *et al.*, 2015; Sunagawa *et al.*, 2015; Brum *et al.*, 2015; Lima-Mendez *et al.*, 2015; Pesant *et al.* 2015; Roux *et al.*, 2016), and help establish foundational ecological hypotheses for the field and a roadmap for the broader life sciences community to better study viruses in complex communities.

Results & Discussion:

The dataset. The Global Ocean Viromes 2.0 (GOV 2.0) dataset is derived from 3.95 Tb of sequencing across 145 samples distributed throughout the world's oceans (**Fig. 1A** and **Table S3;** see **Methods**). These data build on the prior GOV dataset (Roux *et al.*, 2016) by increased sequencing for mesopelagic samples (defined in our dataset as waters between 150m to 1,000m) and upgrading assemblies, both of which drastically improved sampling of the ocean viruses in these samples (results below). Additionally, we added 41 new samples derived from the *Tara* Oceans Polar Circle (*TOPC*) expedition, which traveled 25,000 km around the Arctic Ocean in 2013. These 41 Arctic Ocean viromes were generated to represent the most significantly climate-impacted region of the ocean, and an extreme environment. No such metagenome-based viral data exist for the Arctic region (Deming & Collins 2017), and more generally, for many planktonic organisms, systematic sampling is uneven throughout the Arctic Ocean (CAFF State of the Arctic Marine Biodiversity Report) due to geopolitical and physical challenges of sampling these regions.

The first step to studying viral biodiversity from the assembled GOV 2.0 dataset (see **Methods** and **Fig. S1A**) was to identify contigs that likely derive from viruses using tools that collectively utilize homology to viral reference databases, probabilistic models on viral genomic features, and viral k-mer signatures (see **Methods**). These putative viral contigs were then assigned to 'populations', which are currently defined as viral contigs ≥10 kb where ≥70% of the shared genes have ≥95% average nucleotide identity (ANI) across its members (Brum *et al.*, 2015; Roux *et al.*, 2016; Roux *et al.*, 2018; population definition also discussed below). This process identified 195,728 viral populations in the GOV 2.0 dataset, which is a ~12-fold increase over the 15,280 identified in the original GOV dataset and assemblies (Roux *et al.*, 2016) and augments prior marine viromic work (**Tables S2**). Of these original GOV viral populations, 12,708 were represented by single contigs and, of these, most (92%) were recovered in GOV 2.0 (**Fig. 1B-inset**), with average lengths increased 2.4-fold from 18 kbp to 44 kbp (**Fig. 1B**). Outside these GOV-known and now improved viral populations, an additional 180,448 new GOV 2.0 viral populations were identified -- derived mostly (58%) from improved assemblies and deeper sequencing of the original GOV samples, and the rest (42%) from the 41 new Arctic Ocean viromes. Finally, new methods to identify shorter viral contigs (see **Methods**) were applied and these identified another 292,402 contigs as viral (5-10 kb length and/or circular), which, when added to the earlier data and clustered at ≥95% ANI, resulted in a total of 488,130 viral populations (N50= 15,395; L50=105,286; mean read depth per population = 17x). Ninety percent of the populations could not be taxonomically classified to a known viral family, but the 10% that could were predominantly dsDNA viral families and bacteriophages (**Fig. 1C, D**).

Although the focus of this study is DNA viruses, a remarkable diversity of RNA viruses has been described in nature, though largely outside of marine systems. For example, transcriptome sequencing from plants (Roossinck *et al*., 2010), arthropods (Shi *et al*., 2016), and birds and bats (reviewed in Greninger, 2018) have shown a genomic and phylogenetic diversity of RNA viruses far beyond those in culture (Shi *et al*., 2018). In the oceans, however, RNA viral diversity and abundance remains largely unknown. The few estimates of marine RNA virus abundance are based on the relative quantification of RNA and DNA from purified viral particles

and genome size extrapolations and suggest that up to half of the viral particles in seawater are RNA viruses (Steward *et al*., 2013, Miranda *et al*., 2016). Direct RNA virus counts are not yet available for any environment due to the lack of RNA-specific stains. To date, our understanding of marine RNA viral diversity is based on single-gene surveys that target subgroups of viruses (reviewed in Culley, 2018) and a few viromes generated from extracellular viral particles (Culley and Steward, 2007; Culley *et al*., 2006; Miranda *et al*., 2016; Steward *et al*., 2013; Urayama *et al*., 2018, Zeigler-Allen *et al.*, 2017) or from RNA viral sequences identified in metatranscriptomes (Carradec *et al*., 2018; Moniruzzaman *et al.*, 2017; Urayama *et al.*, 2018; Zeigler-Allen *et al.*, 2017). Together, these studies suggest that the marine RNA virosphere is composed of a large diversity of positive-polarity ssRNA and dsRNA viruses diverge from established taxa, with an apparent predominance of viruses that infect eukaryotes (Culley, 2018). Due to current methodological limitations, comprehensive, systematic assessments of marine RNA viral diversity on the global scale are not yet available, and are excluded from our analysis.

Validating viral 'population' boundaries. Defining species is controversial for eukaryotes and prokaryotes (Kunz, 2013; Cohan, 2002; Fraser *et al.*, 2009) and even more so for viruses (Bobay *et al.*, 2018), largely because of the paradigm of rampant mosaicism stemming from rapidly evolving ssDNA and RNA viruses, whose evolutionary rates are much higher than dsDNA viruses [reviewed by (Duffy *et al.*, 2008)]. The biological species concept, often referred to as the gold standard for defining species, defines species as interbreeding individuals that remain reproductively isolated from other such groups. To adapt this to prokaryotes and viruses, studies have explored patterns of gene flow to determine whether they might maintain discrete lineages as reproductive isolation does in eukaryotes. Indeed, gene flow and selection define clear boundaries between groups of bacteria, archaea and viruses, though the required scale of data are only available for cyanophages and mycophages among viruses (Shapiro *et al.*, 2012; Cadillo-Quiroz *et al.*, 2012; Gregory *et al.*, 2016; Bobay *et al.*, 2018).

Because measuring gene flow requires extensive datasets not yet available for many groups, the term 'species' is rarely used for prokaryotes or viruses, and instead discrete lineages are described as 'populations'. Separate from these population genetic theory grounded observations, evidence of discrete lineages, or sequence-discrete populations, is to use metagenomic read-mapping to evaluate naturally occurring sequence variation across organisms. Sequence-discrete populations have now been observed for prokaryotes (Konstantinidis & Tiedje 2005) and more recently for some dsDNA viruses (viral-tagged metagenomes and 142 isolate genomes for marine cyanophages; Deng *et al*. 2014, Gregory *et al*. 2016; **Table S4**). Buoyed by this and signatures of at least some dsDNA viruses obeying the biological species concept (Bobay *et al.*, 2018), viral ecologists have established the definition of viral populations described above (Brum *et al.*, 2015; Roux *et al.*, 2016; Roux *et al.*, 2018). Notably, however, only deeply sequenced groups, cyano- and myco-phages, have been evaluated to date (Gregory *et al.*, 2016; Bobay *et al.*, 2018), and an emergent hypothesis suggests that phages evolve with different modes and tempos driven by differing temperate or obligately lytic lifestyles (Mavrich & Hatfull, 2017). Thus, there is a need to evaluate how generalizable this empirically-derived ≥95% ANI cut-off viral population definition is in nature.

To test this, we permissively mapped metagenomic reads against our 488,130 GOV 2.0 viral populations by allowing 'local' matching as low as 18% nucleotide identity, and statistically identifying 'breaks' in the resulting read frequency histograms (see **Methods**). This revealed that, on average, the break occurred such that reads <92% nucleotide identity failed to map (**Fig. 2C; full results Table S5)**, which resulted in a genome-wide signature of ≥95% ANI

for nearly all (99.9% or 487,875) of the GOV 2.0 viral populations, including the smaller <10 kb viral populations (**Fig. 2D**). This implies that the observed viral populations in the dataset are predominantly and detectably sequence-discrete. This result is consistent with data from viral-tagged metagenomes (Deng *et al.*, 2014) and gene-sharing networks of prokaryotic virus genomes (Iranzo *et al.*, 2016*,* Bolduc *et al.*, 2017), which also showed that sampled viral genome sequence space is clustered at each 'species' and 'genus' levels, respectively. Thus, while ssDNA and RNA viruses have variable and elevated genome evolutionary rates that can erode species boundaries [reviewed by (Duffy *et al.*, 2008)], it appears that virtually all metagenome-assembled dsDNA viral populations form discrete genotypic clusters and can be appropriately delineated via a ≥95% genome-wide ANI cut-off.

Meta-community analysis reveals 5 ecological zones. Having organized this global sequence space into discrete and biologically meaningful populations, we next sought to use metagenome-derived abundance estimates to establish patterns and drivers of viral population diversity across the global ocean across multiple levels of ecological organization (**Fig. 3**). This revealed that the 145 GOV 2.0 viral communities robustly assorted into just five meta-communities, denoted ecological zones, whether assessed using Bray-Curtis dissimilarity distances in principal coordinate analysis (**Fig. 4A)**, non-metric multidimensional scaling (**Fig. S2A**), or hierarchical clustering (**Fig. S2B**) and after accounting for variable sample sizes (see **Methods**). We designated these 5 emergent ecological zones as the Arctic (ARC), Antarctic (ANT), bathypelagic (BATHY), temperate and tropical epipelagic (TT-EPI) and mesopelagic (TT-MES), and used these for further study. Depth ranges overlapped with those previously defined (Reygondeau, *et al.* 2018), with epipelagic, mesopelagic, and bathypelagic being waters of depths 0 to 150 meters, 150 to 1,000 meters, and deeper than 2,000 meters, respectively.

Comparison of our virome-inferred ecological zones to those inferred for the oceans in other ways was telling. Our zones differed from traditional oceanographic biogeographical biomes (e.g. Longhurst), where four biomes and ~50 provinces have been designated across surface ocean waters based on annual cycles of nutrient chlorophyll a (Longhurst *et al.* 1995, Longhurst 2007), and from mesopelagic ecoregions and biogeochemical provinces based on biogeography and environmental climatology, respectively (Sutton, *et al.* 2017; Reygondeau, *et al.* 2018). However, they were similar to those observed for marine bacterial communities, which clustered by mid-latitude surface, high-latitude, and deep waters (Ghiglione *et al.*, 2012). This implies that the physicochemical structuring of marine *microbial* communities is likely the most important factor in structuring marine viral communities, perhaps reflecting a relative stability in host range of viruses in the oceans (de Jonge *et al.* 2018). To evaluate this physicochemical structuring, we examined the universal predictors and drivers of viral ecological zones, across one (**Fig. 5A**) and multiple ordination dimensions (**Fig. 5B**; see **Methods**). This suggested that temperature was the major driver structuring these ecological zones, as previously shown from global microbial surveys (Sunagawa *et al.*, 2015) and our own smaller ocean virome surveys, where we posited previously that temperature likely directly impacts microbial community structure, and indirectly viral community structure (Brum *et al*., 2015). Moreover, temperature has been shown to play an important role in virus-host interactions, especially in the Arctic (Maat *et al.*, 2017).

To look for specific viral adaptations in each ecological zone, we identified genes under positive selection by evaluating the ratio of non-synonymous to synonymous mutations observed in gene sequences using the pN/pS equation (Schloissnig *et al*., 2013). Of 1,139,501 genes tested from populations with enough coverage (≥10x mean read depth; mean number of populations

assessed per sample: 14,852 viral populations), 124,882 genes were identified as being under positive selection in at least one sample. Most (82%) of the positively selected genes were functionally unannotatable, with the remaining 18% annotatable as predominantly genes related to structure or DNA metabolism (**Tables S6-S10**). In model systems, such genes are often under strong selective pressures during adaptations to new hosts (Marston *et al.*, 2012; Jian *et al*., 2012; Enav *et al.*, 2018). Thus, we hypothesize that host availability in each ecological zone is a strong selective pressure on our marine viral populations. Given the lack of functional annotations for most of the genes, we clustered all translated GOV 2.0 viral genes into protein clusters (PCs) based on sequence homology (*sensu* Holm & Sander, 1998) to identify positively selected zone-specific PCs. This resulted in 823,193 PCs, of which ~10% (79,588 PCs) appeared under positive selection, with a subset of these specific to a single zone (ARC = 80%; ANT = 33%; BATHY = 37%; TT-EPI = 75%; TT-MES = 69% of positively selected PCs per zone; see **Tables S6-10**). These findings of many zone-specific positively-selected PCs is indicative of niche-differentiation. However, functional stories from these data are challenging as 85% of these zone-specific PCs were of unknown function, with the remaining mostly being the structural and DNA metabolism genes described above. This suggests that we have a lot to learn about the function of genes that most likely drive niche-differentiation across the ecological zones.

Viral *macro-* and *micro-* diversity, and potential drivers, within and between ecological zones. To explore diversity patterns across ecological zones, we calculated per sample diversity using Shannon's *H*' for *macro*diversity and a newly established method for community-wide *micro*diversity. This new method for community-wide microdiversity is limited in that it can only assess well-sampled, abundant populations because it estimates the average nucleotide diversity (or π) from the mean of π from 100 randomly subsampled well-sequenced populations sampled 1,000 times (see **Methods**). These zone-normalized (see **Methods**) comparisons revealed that *macro*diversity was highest in TT-EPI ($p < 0.05$), closely followed by the ARC, and lowest in TT-MES and ANT (**Fig 4B –bottom**), whereas *micro*diversity was highest in TT-MES ($p < 0.05$) and lowest in ARC (**Fig. 4B –left**). At the zonal level, a negative trend between *macro-* and *micro-* diversity emerges (**Fig. 4B-right**), although we note that the small number of zonal points limits our statistical inferences, even in this global dataset.

Recent work suggests that higher *micro*-diversity can impede the maintenance of *macro*-diversity by promoting competitive exclusion (Hart *et al.*, 2016). Thus we posit that, if the zonal level negative *macro/micro* diversity trends are real, this may result from increased *intra*population niche variation that reduces *inter*population niche variation resulting in competitive exclusion by the superior competitors, which may occur slowly and may be why it only appears at this regional scale (**Fig. S5**). Because estimates of *micro*diversity in our dataset and even currently available single virus genomics approaches (Martínez-Hernández *et al.,* 2017) remain limited to only the most abundant populations, testing such a hypothesis awaits critically-needed advances and scalability in single-virus genomics technologies.

At the per-sample level, however, *macro-* and *micro-* diversity were not correlated, even within each zone (**Fig. 4B – right**). Although these are the first data available for viruses, for larger organisms, *macro-* and *micro*-diversity are often correlated across habitats sharing similar species pools, presumably due to habitat characteristics altering immigration, drift, and selection (Vellend & Gerber, 2005). These ecological correlations are generally positive and significantly stronger in discrete habitats (e.g. islands) in contrast to more connected communities like the ocean [reviewed in (Vellend *et al.*, 2014)]. Thus we posit that the lack of correlation between

marine viral *macro-* and *micro-* diversity at this per-sample level is driven by differences in local drivers (**Fig. 4C**). Consistent with this, local potential drivers differed as nutrients strongly (and negatively) correlated with viral *macro*diversity, whereas photosynthetically active radiation (PAR; an indicator of productivity) best (and positively) correlated with viral *micro*diversity in the epipelagic waters (**Fig. 4C**).

Mechanistically, these results suggest several possible hypotheses. We interpret that, at the viral *macro*diversity level, decreased host diversity in algal blooms, which themselves rely on nutrient pulses (Farooq & Malfatti, 2007), could skew viral rank abundance curves towards dominance by increasing abundance of bloom-associated viral populations. Even though algal blooms were not targeted in the *Tara* Oceans expedition, we did find that viral *macro*diversity negatively correlated with chlorophyll a (**Fig. 5C**), and particulate inorganic carbon concentration (PIC; **Fig. 4C**), which is commonly used as a proxy for coccolithophore abundance (Groom & Holligan, 1987). Additionally, viral *macro*diversity negatively correlated with the relative abundance of coccolithophores based on the V9 region of the 18S rRNA genes in the sequencing reads (**Fig. 4C**). For viral *micro*diversity in epipelagic waters, we interpret that PAR is potentially the main driver (**Fig. 4C**). PAR is known to impact host diversity, particularly in nutrient-poor surface waters, by inhibiting photoautotrophs through overwhelming their photosystems with too many electrons that can back up and even damage the photosystems (Feng *et al.*, 2015). Further PAR can inhibit the growth of the dominant heterotroph, SAR11 (Ruiz-González *et al.*, 2013), and can stimulate other key microbes such as *Roseobacter*, *Gammaproteobacteria* and NOR5 (Ruiz-González *et al.*, 2013). We hypothesize that the shorter-term impacts of high PAR in the surface waters on host communities may create new niches for viruses, whereby *micro*diversity increases to enable differentiation of existing viral populations. As above, advances in single-virus genomics would be invaluable for testing this hypothesis.

Viral *macro-* and *micro-* diversity, and potential drivers, against classical ecological gradients. Ecologists have long explored the relationship between diversity and geographic range, which in eukaryotes and bacteria are highly (and positively) correlated and thought to be due to the accumulation of niche-specific selective mutations across populations with large heterogeneous geographic ranges (i.e. the niche variation hypothesis; Van Valen, 1965, Hedrick, 2006, Rosen *et al.*, 2015). No parallel studies have looked at viruses. To explore this for viruses, we determined the geographic range of viral populations based on their distribution within and between ecological zones (**Fig. 6A**) and then calculated their average π (see **Methods**) to assess patterns in *macro-* and *micro-* diversity, respectively. Viral populations were designated as 'multi-zonal' if they were observed in >1 ecological zone, 'zone-specific regional' if they were observed in only one zone, but $\geq$2 viral communities, or 'zone-specific local' if they were observed in only 1 viral community within a single zone.

These analyses first revealed differences in the dominant viral geographic ranges across the different ecological zones. For example, multi-zonal viral populations dominated ANT and BATHY (>60% of viral populations found within zone), both across the zone (**Fig. 6B**) and within each station (**Fig. S6**), whereas zone-specific regional viral populations dominated TT-EPI and ARC and the multi-zonal and zone specific viral populations were approximately equally represented in TT-MES (**Fig. 6B**). The high levels of zone-specific viral populations in TT-EPI and ARC, as well as the high levels of viral *macro*diversity (**Fig. 4B-bottom**), are indicative of high endemism and suggest these regions may be biodiversity hotspots for marine viruses. In contrast, the ANT and BATHY are composed mostly of multi-zonal viral populations suggesting that they may be sink habitats that are more dependent on migration (*sensu*

Watkinson & Sutherland, 1995). However, across all ecological zones, viral population *micro*diversity decreased with virus geographic range (**Fig. 6C;** $p < 0.05$), presumably from varied ecologies providing differing selective niches for the single, widely-distributed population that then drive differentiation through isolation-by-environment processes (*sensu* Shapiro *et al.*, 2012). Such findings are new for viruses, but parallel the results for eukaryotes (Hedrick, 2006) and bacteria (Rosen *et al.*, 2015) and suggest a universality to isolation-by-environment processes across organismal kingdoms and viruses.

Ecologists have also long observed, across most flora and fauna, that there are latitudinal patterns in diversity across both terrestrial and marine environments. Briefly, the latitude diversity gradient suggests that both *macro*- and *micro*-diversity are highest at mid-latitudes and decrease poleward (Pianka 1966, Hillebrand 2004, Mannion *et al.*, 2013, Miraldo *et al.*, 2016). We found that both viral *macro*- and *micro*-diversity followed the latitude diversity gradient except in ARC, where both increased (**Fig. 7A**). This high equatorial *macro*- and *micro*-diversity was consistent across the Indian, Atlantic, and Pacific Oceans as expected (**Fig. 7B & C**). The Arctic Ocean, however, was not only unexpectedly elevated in diversity, but it also displayed a unique pattern. Specifically, two distinct zones – definable by climatology-derived water mass nutrient stoichiometry (N*; **Fig. 7D**; see *Comparing ARC-H and ARC-L* in **Methods**) – emerged as high (ARC-H) and low (ARC-L) diversity regions that were significantly differentiable at both *macro*- and *micro*-diversity levels (**Fig. 7E**). Further, ARC-H was characterized by low nutrient ratios (N*; >9X lower in ARC-H than ARC-L on average; $p < 5E-04$) and drove the divergence from the latitude diversity gradient (**Fig. S7)**.

Mechanistically, we interpret these observations as follows. Prior work in this region has shown (i) strong denitrification in the Bering Strait (Devol *et al.*, 1997), which explains the low N* in the west, and (ii) increasing oligotrophy in the Beaufort Gyre due to increasing vertical stratification, which selects against larger algae and for smaller algae and bacteria in the ARC-H (Li *et al.*, 2009). As above, we hypothesize that shorter-term increased host diversity results in increased viral *macro*- and *micro*-diversity in ARC-H. Though our GOV 2.0 dataset is confounded by seasonality of sampling, we posit that this elevated summer-time *macro*- and *micro*-diversity in ARC may fuel viral ecological differentiation and represent an unrecognized 'cradle' of viral biodiversity beyond the tropics. Though this elevated diversity in the Arctic was surprising, together with a similar deviation seen in mollusks (Valdovinos *et al.*, 2003) and recently reported in ray-finned fish (Rabosky *et al.*, 2018), these results call into question whether this decades-old paradigm needs revisiting and suggests that polar regions may be important biodiversity hotspots for viruses, as well as larger organisms.

Finally, as ocean exploration accelerates, patterns in diversity through the vertical layers of the ocean have become a focus. An emergent depth diversity gradient hypothesis suggests that *macro*diversity decreases with depth (Costello & Chaudhary, 2017), which has been explored across the World Register of Marine Species that includes some microbes and viruses (http://www.marinespecies.org/), but *micro*diversity has not yet been explored for any organism. Overall, our virome-inferred diversity patterns were less obviously consistent with the depth diversity gradient, although deep water ocean data were limited (**Fig. 7F**). Briefly, viral *macro*diversity largely followed the depth diversity gradient with high diversity in the surface waters and decreased diversity with depth, whereas viral *micro*diversity did not as it decreased until 200 m depth, but then sharply increased (**Fig. 7F**). This deep water increase coincided with an increase in bacterial *macro*diversity in the mesopelagic region (**Fig. S8A & B**), and in TT-MES, this bacterial *macro*diversity correlated with viral *micro*diversity (**Fig. S8C**).

If more extensive deep water sampling confirms these patterns, we see several scenarios that could explain these data. First, we hypothesize that viral *micro*diversity may, in part, be driven by an increase in *macro*diversity of zone-specific bacterial populations in TT-MES, which we interpret as an expansion of host 'niches' available for infection that could drive diversification in viruses (Elena *et al*., 2009). Second, we hypothesize that the decrease in viral *macro*diversity may be driven by increased viral *micro*diversity of some viral populations in the mesopelagic region that can promote competitive exclusion (*sensu* Hart *et al.*, 2016) as discussed above. Alternatively, lower cell density in the mesopelagic layer (Sunagawa *et al.* 2015) may result in less encounters between "predator" and "prey", reducing viral speciation (as a function of reduced number of viral generations), but selecting for viruses with broader host range. Again, testing these hypotheses will require technological advances to measure *in situ* host ranges and sensitivities of viruses and cells, respectively, at scales relevant to the diversity in nature.

Conclusions:

This study provides a systematic and global-scale view of patterns and drivers of marine viral *macro-* and *micro-* diversity that reveals three overarching advances. First, five ecological zones emerge for the global ocean, which contrasts known Longhurst biogeographic patterning in other organisms, but is consistent with observations from the largely co-sampled ocean microbiome (Sunagawa *et al.* 2015). Second, patterns and drivers of viral *macro-* and *micro-* diversity differ per-sample and positively correlate to geographic range. These findings offer hints at underlying mechanisms that impact these two levels of diversity that will guide researchers from discovery to hypothesis-testing as technologies, such as scalable single virus genomics and *in situ* host range assays, advance towards sampling scales relevant to those in nature. Third, epipelagic waters and the Arctic Ocean emerge from our work as biodiversity hotspots for viruses. While this is surprising given the latitude diversity gradient paradigm that the tropics rather than the poles are the cradles of diversity, it is in line with other observations in larger organisms (Valdovinos *et al.*, 2003, Rabosky *et al.*, 2018) and emphasizes the importance of these drastically climate-impacted Arctic regions for global biodiversity. Together, these advances, along with the parallel global-scale ecosystem-wide measurements of *Tara* Oceans (e.g. de Vargas *et al.*, 2015; Sunagawa *et al.*, 2015; Brum *et al.*, 2015; Lima-Mendez *et al.*, 2015; Roux *et al.*, 2016) provide the foundation for incorporating viruses into emerging genes-to-ecosystems models (e.g. Guidi *et al.* 2016, Garza *et al.*, 2018) that guide ocean ecosystem management decisions that are likely needed if humans and the Earth System are to survive the current epoch of the planet-altering Anthropocene.

References:

Achtman, M., and Wagner, M. (2008). Microbial diversity and the genetic nature of microbial species. *Nat. Rev. Microbiol.* 6, 431–40.

Bar-On, Y.M., Phillips, R., and Milo, R. (2018). The biomass distribution on Earth. *Proc. Natl. Acad. Sci. USA*, 10.1073/pnas.1711842115.

Bobay, L., and Ochman H. (2018). Biological species in the viral world. *Proc. Natl. Acad. Sci. USA*, 10.1073/pnas.1717593115.

Bolduc, B., Jang, H.B., Doulcier, G., You, Z.Q., Roux, S., and Sullivan, M.B. (2017). vConTACT: an iVirus tool to classify double-stranded DNA viruses that infect Archaea and Bacteria. *PeerJ.* 5, e3243.

Brum, J.R., Ignacio-Espinoza, J.C., Roux, S., Doulcier, G., Acinas, S.G., Alberti, A., Chaffron, S., Cruaud, C., de Vargas, C., Gasol, J.M. *et al.* (2015). Patterns and ecological drivers of ocean viral communities. *Science*. 348, 1261498.

Cadillo-Quiroz, H., Didelot, X., Held, N.L., Herrera, A., Darling, A., Reno, M.L., Krause, D.J., and Whitaker, R.J. (2012). Patterns of Gene Flow Define Species of Thermophilic Archaea. *PLOS Biol.* 10, e1001265.

Carradec, Q., Pelletier, E., Da Silva, C., Alberti, A., Seeleuthner, Y., Blanc-Mathieu, R., Lima-Mendez, G., Rocha, F., Tirichine, L., Labadie, K., et al. (2018). A global ocean atlas of eukaryotic genes. *Nat. Commun.* 9.

Cohan, F.M. (2002). What are bacterial species? *Annu. Rev. Microbiol.* 56, 457-487.

Conservation of Arctic Flora and Fauna (2017). *State of the Arctic Marine Biodiversity Report*. Conservation of Arctic Flora and Fauna.

Costello, M.J., and Chaudhary, C. (2017). Marine biodiversity, biogeography, deep-Sea gradients, and conservation. *Curr. Biol.* 27, 2051.

Culley, A. (2018). New insight into the RNA aquatic virosphere via viromics. *Virus Res*. 244, 84–89.

Culley, A.I., and Steward, G.F. (2007). New genera of RNA viruses in subtropical seawater, inferred from polymerase gene sequences. *Appl. Environ. Microbiol.* 73, 5937–5944.

Culley, A.I., Lang, A.S., and Suttle, C.A. (2006). Metagenomic Analysis of Coastal RNA Virus Communities. *Science.* 312, 1795–1798.

de Jonge, P.A., Nobrega, F.L., Brouns, S.J.J., and Dutilh, B.E. (2019). Molecular and evolutionary determinants of bacteriophage host range. *Trends Microbiol.* 27, 51-63.

de Vargas, C., Audic, S., Henry, N., Decelle, J., Mahé, F., Logares, R., Lara, E., Berney, C., Le Bescot, N., Probert, I., *et al.* (2015). Eukaryotic plankton diversity in the sunlit ocean. *Science.* 348, 1261605.

Deming, J. W., and Collins, E. (2017). Sea ice as a habitat for Bacteria, Archaea and Viruses.In: Thomas D.N. (ed). Sea ice. John Wiley and sons, Ltd. 3rd edition.

Deng, L., Ignacio-Espinoza, J.C., Gregory, A.C., Poulos, B.T., Weitz, J.S., Hugenholtz, P., and Sullivan, M.B. (2014). Viral tagging reveals discrete populations in Synechococcus viral genome sequence space. *Nature.* 513, 242–245.

Devol, A.H., Codispoti, L.A., and Christensen, J.P. (1997). Summer and winter denitrification rates in western Arctic shelf sediments. *Cont. Shelf Res.* 17.9, 1029-1033.

Duffy, S., Shackelton, L.A., and Holmes, E.C. (2008). Rates of evolutionary change in viruses: patterns and determinants. *Nat. Rev. Genet.* 9, 267–276.

Elena, S.F., Agudelo-Romero, P., Lalić, J. (2009) The evolution of viruses in multi-host fitnness landscapes. *Open Virol.* J. 3, 1-6.

Enav, H., Kirzner S., Lindell, D., Mandel-Gutfreund, and Y., Béjá, O. (2018). Adaptation to sub-optimal hosts is a driver of viral diversification in the ocean. *Nature Comm.* 9, 4698.

Farooq, A., and Malfatti, F. (2007). Microbial structuring of marine ecosystems. *Nat. Rev. Microbiol.* 5.10., 782-791.

Feng, J., Durant, J.M, Stige, L.C., Hessen, D.O., Hjermann, D.Ø., Zhu, L., Llope, M., and Stenseth, N.C. (2015). Contrasting correlation patterns between environmental factors and chlorophyll levels in the global ocean. *Global Biogeochem. Cycles.* 29.12, 2095-2107.

Fraser, C., Alm, E.J., Polz, M.F., Spratt, B.G., and Hanage, W.P. (2009). The bacterial species challenge: making sense of genetic and ecological diversity. *Science.* 323, 741-746.

Garza, D.R., van Verk, M.C., Huynen, M.A., and Dutilh, B.E. (2018). Towards predicting the environmental metabolome from metagenomics with a mechanistic model. *Nat. Microbiol.* 3, 456-460.

Ghiglione, J.F., Galand, P.E., Pommier, T., Pedrós-Alió, C., Maas, E.W., Bakker, K., Bertilson, S., Kirchmanj, D.L., Lovejoy, C., Yager, P.L. *et al.* (2012). Pole-to-pole biogeography of surface and deep marine bacterial communities. *Proc. Natl. Acad. Sci. USA.* 109, 17633–17638.

Gregory, A.C., Solonenko, S.A., Ignacio-Espinoza, J.C., LaButti, K., Copeland, A., Sudek, S., Maitland, A., Chittick, L., Dos Santos, F., Weitz, J.S. *et al.* (2016). Genomic differentiation among wild cyanophages despite widespread horizontal gene transfer. *BMC Genomics*. 17, 930.

Greninger, A.L. (2018). A decade of RNA virus metagenomics is (not) enough. *Virus Res*. 244, 218–229.

Groom, S.B., and Holligan, P.M. (1987). Remote sensing of coccolithophore blooms. *Adv. Space Res.* 7, 73–78.

Guidi, L., Chaffron, S., Bittner, L., Eveillard, D., Larhlimi, A., Roux, S., Darzi, Y., Audic, S., Berline, L., Brum, J., *et al.* (2016). Plankton networks driving carbon export in the oligotrophic ocean. *Nature*. 532, 465–470.

Hart, S.P, Schreiber, S.J., and Levine, J.M. (2016). How variation between individuals affects species coexistence. *Ecol. Lett.* 19.8, 825-838.

Hedrick, P.W. (2006). Genetic Polymorphism in Heterogeneous Environments: The Age of Genomics. *Annu. Rev. Ecol. Evol. Syst.* 37, 67–93.

Hillebrand, H. (2004) On the generality of the latitudinal diversity gradient: *Am. Nat.* 163:192–211.

Holm, L. and Sander, C. (1998). Removing near-neighbour redundancy from large protein sequence collections. *Bioinformatics.* 14, 423–429.

Hughes, A.R., Inouye, B.D., Johnson, M.T. J., Underwood, N., and Vellend, M. (2008). Ecological consequences of genetic diversity. *Ecol. Lett.* 11, 609–623.

Hurwitz, B.L., and U'Ren, J.M. (2016). Viral metabolic reprogramming in marine ecosystems. *Curr Opin Microbiol.* 31, 161-168.

Iranzo, J., Koonin, E.V., Prangishvili, D., and Krupovic, M. (2016). Bipartite network analysis of the archaeal virosphere: evolutionary connections between viruses and capsid-less mobile elements. *J. Virol.* 90.24, 11043-11055.

Jian, H., Xu, J., Xiao, X., and Wang, F. (2012). Dynamic modulation of DNA replication and gene transcription in deep-sea filamentous phage SW1 in response to changes of host growth and temperature. *PLoS One* 7.8, e41578.

Konstantinidis, K.T., and Tiedje, J. (2005) Genomic insights that advance the species definition for prokaryotes. *Proc. Natl. Acad. Sci. USA.* 102, 2567-2572.

Kunz, W. (2013). *Do species exist?: Principles of taxonomic classification*. John Wiley & Sons.

Larkin, A.A., and Martiny, A.C. (2017). Microdiversity shapes the traits, niche space, and biogeography of microbial taxa. *Environ. Microbiol.* Rep. 9, 55–70.

Le Quéré, C., Andrew, R. M., Friedlingstein, P., Sitch, S., Pongratz, J., Manning, A.C.,Korsbakken, J. I., Peters, G. P., Canadell, J. G., Jackson, R., *et al.* (2018). Global carbon budget 2017. *Earth System Science Data* 10.1, 405-448.

Lee, S.T.M., Kahn, S.A., Delmont, T.O., Shaiber, A., Esen, Ö.C., Hubert, N.A., Morrison, H.G., Antonopoulos, D.A., Rubin, D.T., and Eren, A.M. (2017). Tracking microbial colonization in fecal microbiota transplantation experiments via genome-resolved metagenomics. *Microbiome. J.* 5, 50.

Leibold, M.A., Holyoak, M., Mouquet, N., Amarasekare, P., Chase, J.M., Hoopes, M.F., Holt, R.D., Shurin, J.B., Law, R., Tilman, D. *et al.* (2004). The metacommunity concept: a framework for multi-scale community ecology. *Ecol. Lett.* 7, 601–613.

Li, W.K.W., McLaughlin, F.A., Lovejoy, C., and Carmack, E.C. (2009). Smallest algae thrive as the Arctic Ocean freshens. *Science.* 326, 539.

Lima-Mendez, G., Faust, K., Henry, N., Decelle, J., Colin, S., Carcillo, F., Chaffron, S., Ignacio-Espinosa, J.C., Roux, S., Vincent, F., *et al.* (2015). Determinants of community structure in the global plankton interactome. *Science*. 348, 1262073.

Longhurst, A.R. (2007) Ecological geography of the sea (Boston, MA: Academic Press).

Longhurst, A., Sathyendranath, S., Platt, T., and Caverhill, C. (1995). An estimate of global primary production in the ocean from satellite radiometer data. *J. Plankton Res.* 17, 1245-1271.

Maat, D.S., Biggs, T., Evans, C., van Bleijswijk, J.D.L., van der Wel, N.N., Dutilh, B.E., Brussaard, C.P.D. (2017) Characterization and temperature dependence of Arctic Micromonas polaris viruses. *Viruses* 9.6, 134.

Mannion, P.D., Upchurch, P., Benson, R.B.J., Goswami, A. (2013) The latitudinal biodiversity gradient through deep time. *Trends Ecol. Evol.* 29: 42-50.

Marston, M.F., Pierciey, F.J. Jr., Shepard, A., Gearin G., Qi, J., Yandava, C., Schuster, S.C., Henn, M.R., and Martiny, J.B.H. (2012). Rapid diversification of coevolving marine *Synechococcus* and a virus. *Proc. Natl. Acad. Sci. USA.* 109, 4544–4549.

Martínez-Hernández, F., Fornas, O., Lluesma Gomez, M., Bolduc, B., de la Cruz Peña, M.J., Martínez, J.M., Antón, J., Gasol, J.M., Rosselli, R., Rodríguez-Valera, F., *et al.* (2017). Single-virus genomics reveals hidden cosmopolitan and abundant viruses. *Nature Comm.* 8, 15892.

Mavrich, T.N., and Hatfull G.F. (2017). Bacteriophage evolution differs by host, lifestyle and genome. *Nat. Microbiol.* 2, 17112.

Miraldo, A., Li, S., Borregaard, M.K., Flórez-Rodríguez, A., Gopalakrishnan, S., Rizvanovic, M., Wang, Z., Rahbek, C., Marske, K.A., and Nogués-Bravo, D. (2016). An Anthropocene map of genetic diversity. *Science.* 353, 1532–1535.

Miranda, J.A., Culley, A.I., Schvarcz, C.R., and Steward, G.F. (2016). RNA viruses as major contributors to Antarctic virioplankton. *Environ. Microbiol.* 18, 3714–3727.

Moniruzzaman, M., Wurch, L.L., Alexander, H., Dyhrman, S.T., Gobler, C.J., and Wilhelm, S.W. (2017). Virus-host relationships of marine single-celled eukaryotes resolved from metatranscriptomics. *Nat. Commun.* 8, 1–10.

Paul, J.H. (1999). Microbial gene transfer: an ecological perspective. *J Mol Microbiol Biotechnol.* 1, 45-50.

Pesant, S., Not, F., Picheral, M., Kandels-Lewis, S., Le Bescot, N., Gorsky, G., Iudicone, D., Karsenti, E., Speich, S., Troublé, R., *et al*. (2015). Open science resources for the discovery and analysis of Tara Oceans data. *Sci Data.* 2, 150023.

Petrie, K.L., Palmer, N.D., Johnson, D.T., Medina, S.J., Yan, S.J., Li, V., Burmeister, A.R., and Meyer, J.R. (2018) Destabilizing mutations encode nongenetic variation that drives evolutionary innovation. *Science.* 359, 1542-1545.

Pianka, E.R. (1966). Latitudinal Gradients in Species diversity: A Review of Concepts. *Am. Nat.* 100, 33–46.

Rabosky, D.L., Chang, J., Title, P.O., Cowman, P.F., Sallan, L., Friedman, M., Kaschner, K., Garilao, C., Near, T.J., Coll, M. *et al.* (2018). An inverse latitudinal gradient in speciation rate for marine fishes. *Nature.* 559, 392-395.

Reiners, W.A., Lockwood, J.A., Reiners, D.S., and Prager, S.D. (2017). 100 years of ecology: what are our concepts and are they useful? *Ecol. Monograph.* 87, 260–277.

Reygondeau, G., Guidi, L., Beaugrand, G., Henson, S.A., Koubbi, P., MacKenzie, B.R., Sutton, T.T., Fioroni, M., and Maury, O. (2018). Global biogeochemical provinces of the mesopelagic zone. *J. Biogeogr.* 45.2, 500-514.

Roossinck, M.J., Saha, P., Wiley, G.B., Quan, J., White, J.D., Lai, H., Chavarría, F., Shen, G., and Roe, B.A. (2010). Ecogenomics: Using massively parallel pyrosequencing to understand virus ecology. *Mol. Ecol.* 19, 81–88.

Rosen, M.J., Davison, M., Bhaya, D., and Fisher, D.S. (2015). Fine-scale diversity and extensive recombination in a quasisexual bacterial population occupying a broad niche. *Science.* 348, 1019–1023.

Roux, S., Adriaenssens, E.M., Dutilh. B.E., Koonin, E.V., Kropinski, A.M., Krupovic, M., Kuhn, J.H., Lavigne, R., Brister, R., Varsani, A. *et al.* (2018). Minimum Information about an Uncultivated Virus Genome (MIUViG). *Nature Biotechnol.* nbt.4306.

Roux, S., Enault, F., Hurwitz, B.L., and Sullivan, M.B. (2015). VirSorter: mining viral signal from microbial genomic data. *PeerJ*. **3**, e985.

Roux, S., Brum, J.R., Dutilh, B.E., Sunagawa, S., Duhaime, M.B., Loy, A., Poulos, B.T., Solonenko, N., Lara, E., Poulain, J. *et al.* (2016). Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature.* 537, 689–693.

Ruiz-González, C., Simó, R., Sommaruga, R., and Gasol, J.M. (2013). Away from darkness: a review on the effects of solar radiation on heterotrophic bacterioplankton activity. *Front. Microbiol.* 4, 131.

Schloissnig, S., Arumugam, M., Sunagawa, S., Mitreva, M., Tap, J., Zhu, A., Waller, A., Mende, D.R., Kultima, J.R., Martin, J. *et al.* (2013). Genomic variation landscape of the human gut microbiome. *Nature.* 493, 45–50.

Ser-Giacomi, E., Zinger, L., Malviya, S., De Vargas, C., Karsenti, E., Bowler, C., De Monte, S. (2018). Ubiquitous abundance distribution of non-dominant plankton across the global ocean. *Nat. Ecol. Evol.* 2, 1243–1249.

Shapiro, B.J., Friedman, J., Cordero, O.X., Preheim, S.P., Timberlake, S.C., Szabó, G., Polz, M.F., and Alm, E.J. (2012). Population genomics of early events in the ecological differentiation of bacteria. *Science.* 336, 48–51.

Shi, M., Lin, X.D., Tian, J.H., Chen, L.J., Chen, X., Li, C.X., Qin, X.C., Li, J., Cao, J.P., Eden, J.S., et al. (2016). Redefining the invertebrate RNA virosphere. *Nature.* 540, 539–543.

Shi, M., Zhang, Y.Z., and Holmes, E.C. (2018). Meta-transcriptomics and the evolutionary biology of RNA viruses. *Virus Res.* 243, 83–90.

Smillie, C.S., Sauk, J., Gevers, D., Friedman, J., Sung, J., Youngster, I., Hohmann, E.L., Staley, C., Khoruts, A., Sadowsky, M.J, *et al.* (2018). Strain tracking reveals the determinants of bacterial engraftment in the human gut following fecal microbiota transplantation. *Cell Host Microbe* 23, 229-240.

Snitkin, E.S., Zelazny, A.M., Montero, C.I., Stock, F., Mijares, L., NISC Comparative Sequence Program, Murray, P.R., and Segre, J.A. (2011). Genome-wide recombination drives diversification of epidemic strains of *Acinetobacter baumannii*. *Proc. Natl. Acad. Sci. USA.* 108, 13758-13763.

Soliveres, S., van der Plas, F., Manning, P., Prati, D., Gossner, M.M., Renner, S.C., Alt, F., Arndt, H., Baumgartner, V., Binkenstein, J., *et al.* (2016). Biodiversity at multiple trophic levels is needed for ecosystem multifunctionality. *Nature.* 536, 456-459.

Steward, G.F., Culley, A.I., Mueller, J.A., Wood-Charlson, E.M., Belcaid, M., and Poisson, G. (2013). Are we missing half of the viruses in the ocean? *ISME J.* 7, 672–679.

Sullivan, M.B. (2015). Viromes, not gene markers, for studying double-stranded DNA virus communities. *J. Virol.* 89.5, 2459-2461.

Sunagawa, S., Coelho, L.P., Chaffron, S., Kultima, J.R., Labadie, K., Salazar, G., Djahanschiri, B., Zeller, G., Mende, D.R., Alberti, A. *et al.* (2015). Structure and function of the global ocean microbiome. *Science.* 348, 1261359.

Suttle, C. A. (2007). Marine viruses — major players in the global ecosystem. *Nat. Rev. Microbiol.* 5, 801–812.

Sutton, T.T., Clark, M.R., Dunn, D.C., Halpin, P.N., Rogers, A.D., Guinotte, J., Bograd, S.J., Angel, M.V., Perez, J.A.A., Wishner, K., *et al*. (2017). A global biogeographic classification of the mesopelagic zone. *Deep-Sea Res. I.* 126, 85-102.

Tilman, D., Isbell, F., and Cowles, J.M. (2014). Biodiversity and ecosystem functioning. *Annu. Rev. Ecol. Evol. Syst.* 45, 471-493.

Urayama, S., Takaki, Y., Nishi, S., Yoshida-Takashima, Y., Deguchi, S., Takai, K., and Nunoura, T. (2018). Unveiling the RNA virosphere associated with marine microorganisms. *Mol. Ecol. Resour.* 1–12.

Valdovinos, C., Navarrette, S.A., and Marquet, P.A. (2003). Mollusk species diversity in the Southeastern Pacific: Why are there more species towards the pole? *Ecography*. 26, 139-144.

Van Valen, L. (1965). Morphological variation and width of ecological niche. *Am. Nat.* 99, 377-389.

Vellend, M., and Geber, M.A. (2005). Connections between species diversity and genetic diversity. *Ecol. Lett.* 8, 767–781.

Vellend, M., Lajoie, G., Bourret, A., Múrria, C., Kembel, S.W., and Garant, D. (2014). Drawing ecological inferences from coincident patterns of population- and community-level biodiversity. *Mol. Ecol.* 23, 2890–2901.

Watkinson, A.R., and Sutherland, W.J. (1995). Sources, sinks, and pseudo-sinks. *J. Anim. Ecol.* 64.1, 126-130.

Worm, B., Barbier, E.B., Beaumont, N., Duffy, J.E., Folke, C., Halpern, B.S., Jackson, J.B., Lotze, H.K., Micheli, F., Palumbi, S.R., *et al.* (2006). Impacts of biodiversity loss on ocean ecosystem services. *Science.* 314, 787-790.

Zeigler Allen, L., McCrow, J.P., Ininbergs, K., Dupont, C.L., Badger, J.H., Hoffman, J.M., Ekman, M., Allen, A.E., Bergman, B., and Venter, J.C. (2017). The Baltic Sea Virome: Diversity and Transcriptional Activity of DNA and RNA Viruses. *mSystems* 2, e00125-16.

Main Text Figure Legends:

Fig. 1. The Global Ocean Viromes 2.0. (A) Arctic projection of the global ocean highlighting the new sampling stations of viromes in the GOV 2.0 dataset. Datasets from non-arctic samples were previously published in (Brum *et al.*, 2015; Roux *et al.*, 2016). **(B)** Histograms of the average assembled contig lengths for viral populations >10 kb shared between GOV and GOV 2.0. **B-inset.** More than 92% of the unbinned GOV viral populations were reassembled and identified in GOV 2.0 >10 kb populations. **(C)** Pie charts showing how many of the 488,130 total viral populations comprising GOV 2.0 can be annotated and, of those, their viral family level taxonomy. **(D)** Barplot showing the host affiliations for each viral population at the domain level.

Fig. 2. GOV 2.0 viral population have discrete population boundaries. (A) Barplots showing the read mapping results for the most abundant viral population >10kb in length for each of the top four viral families. Despite differences in read boundaries across the representative viral populations, there is no difference in the average read boundaries across the different viral families. **(B)** Histogram showing the read distribution frequency break (i.e. read boundary) between spuriously mapped reads and legitimate reads mapping to the genome. **(C)** Histograms showing the average percent identity of reads mapped to each genome after removing spuriously mapped reads.

Fig. 3. Ecological levels of organization. Schematic showing the different ecological levels of organization studied in this paper.

Fig. 4. Viral communities partition into five ecological zones with different *macro-* and *micro-* diversity levels. (A) Principal coordinate analysis (PCoA) of a Bray-Curtis dissimilarity matrix calculated from GOV 2.0. Analyses show that viromes significantly (Permanova p = 0.001) structure into five distinct global ecological zones: ARC, ANT, BATHY, TT-EPI, and TT-MES zones. Ellipses in the PCoA plot are drawn around the centroids of each group at 95% (inner) and 97.5% (outer) confidence intervals. Four outlier viromes that did not cluster with their ecological zones were removed (**Fig. S3A**) and all the sequencing reads were used (see **Fig. S3B** and **Methods**). **(B – right)** Scatterplots showing correlations between *macro-* (Shannon's *H*') and *micro-* (average π for viral populations with ≥ 10x median read depth coverage; see **Methods**) diversity values for each sample across GOV 2.0. The larger circles represent the average per zone. **(B – left)** Boxplots showing median and quartiles of average *micro*diversity per ecological zone. (**B – bottom**) Boxplots showing median and quartiles of *macrodiversity* for each ecological zone. Zonal samples were randomly downsampled to n = 5 to account for zone sampling difference. All pairwise comparisons shown were statistically significant (p<0.01) using two-tailed Mann-Whitney U-tests. **(C)** Positive (blue) and negative (red) Pearson's correlation results comparing *macro-* (upper) and *micro-* (lower) diversity with different biogeographical and biogeochemical parameters at the global scale (see **Fig. S4**, **Table S3** for all abbreviations, and **Methods**). The significance of the correlations is indicated by the size of the black circles on top of the bars, and the variables on the x-axis are ordered from the strongest to the weakest correlation with *macro*diversity (except for the top four variables correlating with *micro*diversity for readability).

Fig. 5. Ecological drivers of global viral *macro*diversity. (A) Regression analysis between the first coordinate of a PCoA (**Fig. 4A**) and temperature showed that samples were separated by

their local temperatures with an r^2 of 0.82. **(B)** Potential ecological drivers & predictors of beta-diversity across GOV 2.0 for the first two dimensions (Goodness of fit r^2 using a generalized additive model) and across all dimensions (Mantel test based on Spearman's correlation). Temperature was uniformly reported as the best predictor of viral beta-diversity globally. **(C)** Regression analysis between viral *macro*diversity at the deep chlorophyll maximum (DCM) layer and areal chlorophyll a concentration (after cube transformation) showed that the negative correlation between viral *macro*diversity and nutrients (**Fig. 4C**) is mediated (at least partially) by primary productivity. The untransformed values are provided on the lower axis for reference. The Shannon's *H* outlier 32_DCM (**Fig. S3**) and a chlorophyll a concentration outlier (173_DCM; **Fig. 5D**) have been excluded from the regression analysis. **(D)** Boxplot analysis of areal chlorophyll a concentrations showing a single outlier concentration that fell above the fourth quantile of the data points (function geom_boxplot of ggplot).

Fig. 6. Size of geographic range positively correlates with *micro*diversity. (A) Venn diagram showing the number of viral populations found only in one zone (zone-specific) and those that are shared between and among the five ecological zones (multi-zonal). **(B)** Stacked barplots showing the number of multi-zonal, regional, and local viral populations found within the species pool of each ecological zone. **(C)** Boxplots showing median and quartiles of *micro*diversity (average π for viral populations with $\geq$ 10x median read depth coverage) per populations found within each zone defined as multi-zonal, regional, or local. Statistics were the same as in Fig. 2.

Fig. 7. Viral *macro-* and *micro-* diversity global biodiversity trends. (A) Loess smooth plots showing the latitudinal distributions of *macro-* and *micro-*diversity. **(B & C)** Equirectangular projections of the globe showing *macro-* and *micro-*diversity levels within each sample, respectively, across the global ocean. Samples collected at different depths from the same latitude and longitude are overlaid and the colors representing their *macro-* and *micro-* diversity values are merged. **(D)** Arctic projection of the global ocean showing the geographical division between ARC-H and ARC-L stations. The patterns are largely concordant with the Arctic division by climatology-derived N*. While we did sample across different seasons, the calculated N* values are not dependent on the season (see *impact of the coast, depth, and seasons* in **Methods**). **(E)** Boxplots showing median and quartiles of *macro-* (left) and *micro-* (right) diversity of the ARC-H and ARC-L regions. Statistics were the same as in Fig. 2. **(F)** Loess smooth plots showing the depth distributions of *macro-* and *micro-* population diversity. On all the smooth plots, the line represents the Loess best fit, while the lighter band corresponds to the 95% confidence window of the fit. Abbreviations: N*, the departure from dissolved N:P stoichiometry in the Redfield ratio and a geochemical tracer of Pacific and Atlantic water mass (see **Methods**).

Main Text Figures:

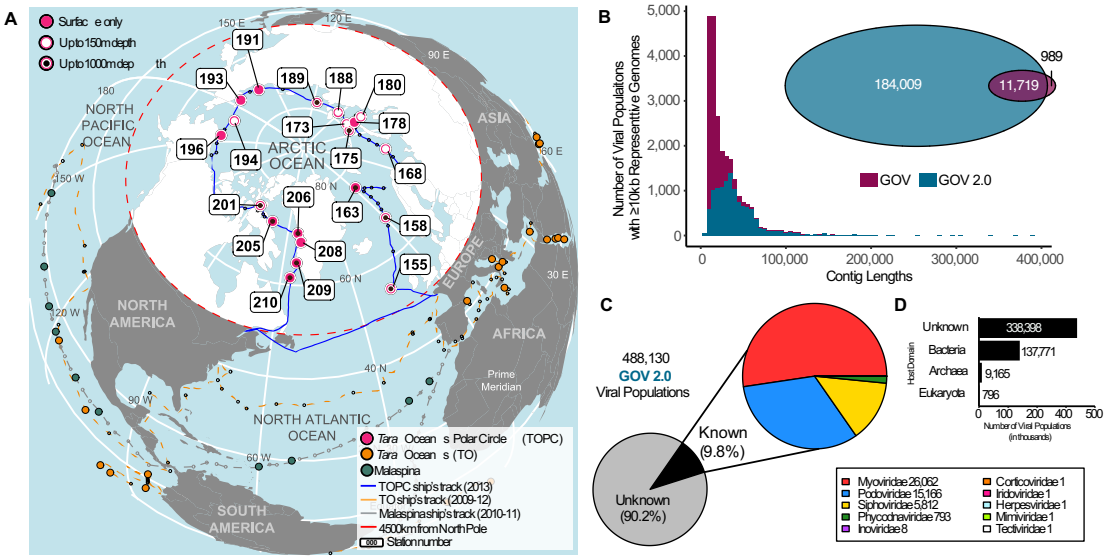


Fig. 1. The Global Ocean Viromes 2.0.

755

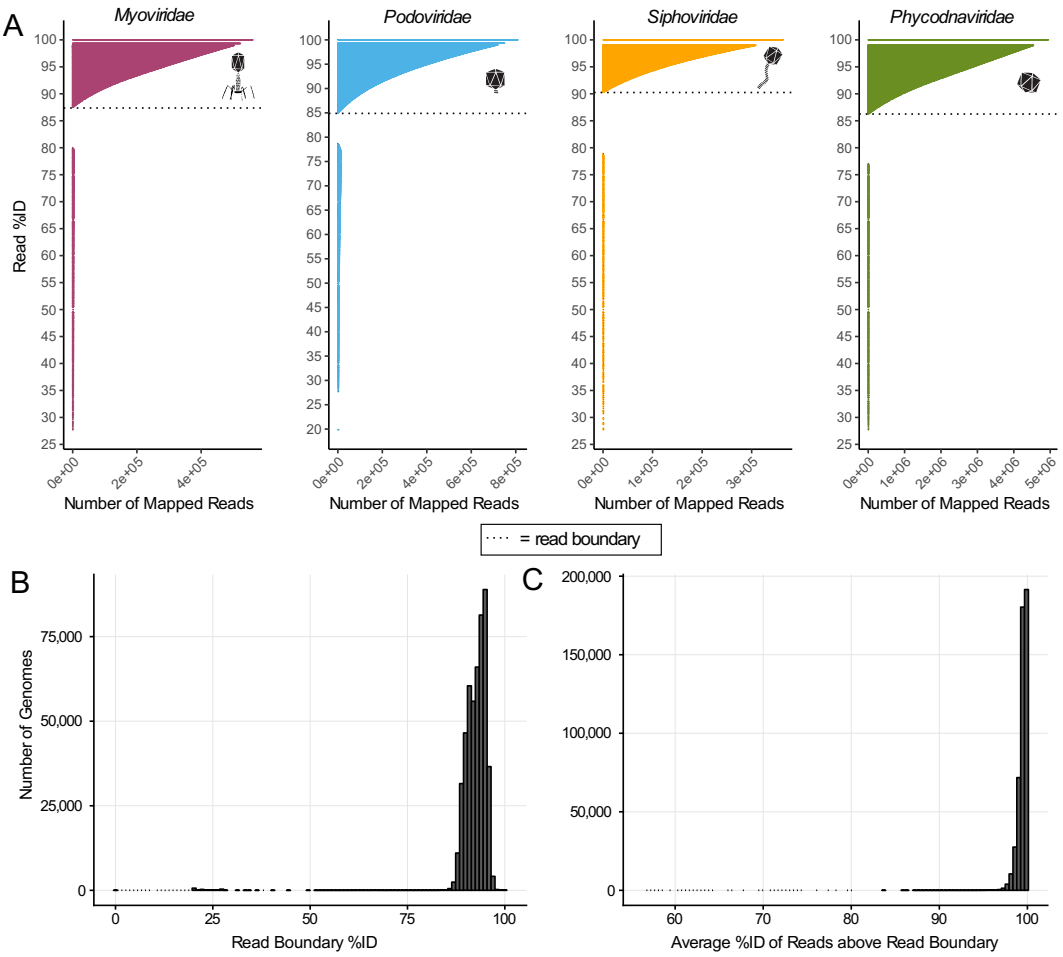
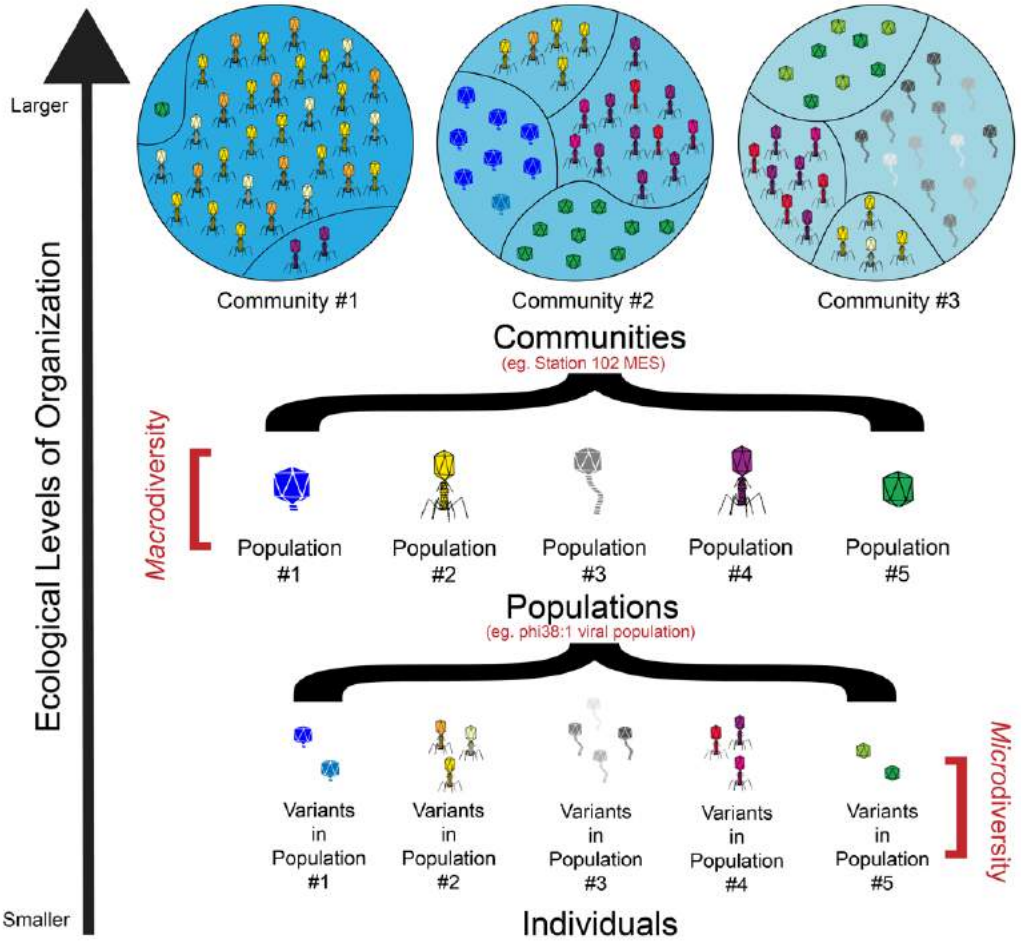


Fig. 2. GOV 2.0 viral population have discrete population boundaries.



760

Fig. 3. Ecological levels of organization.

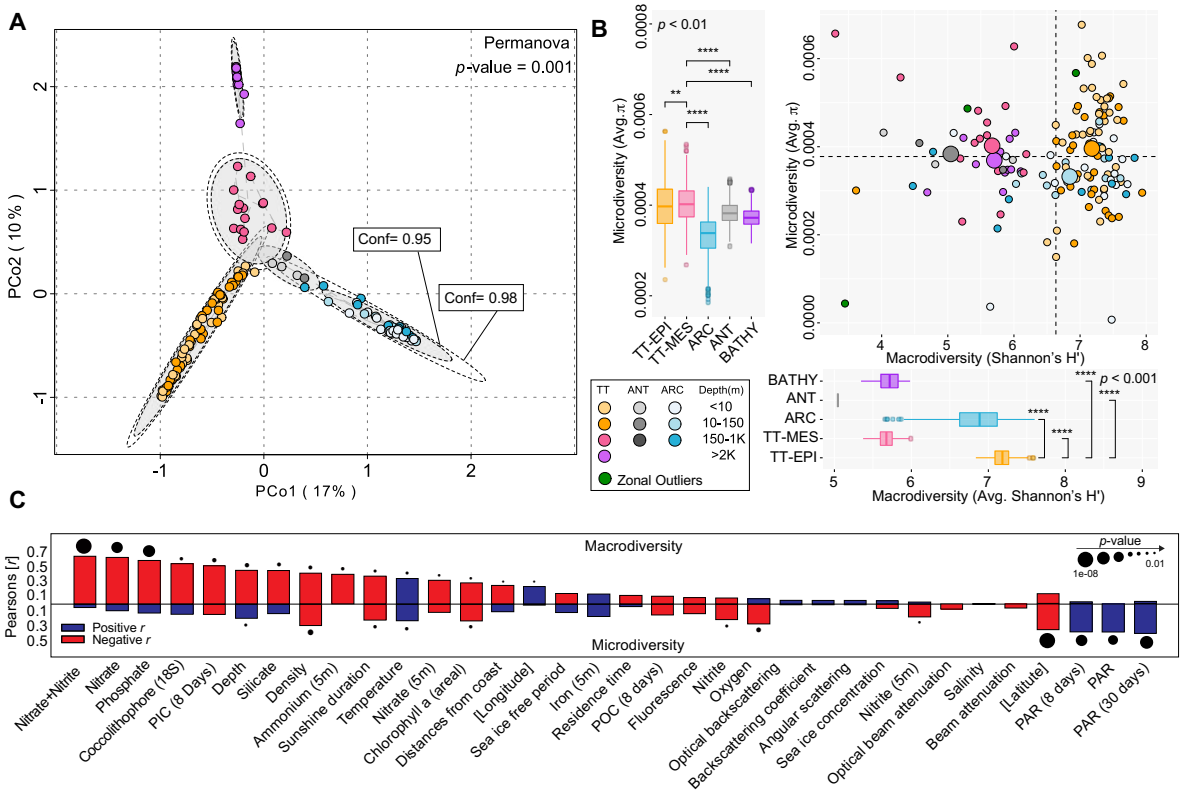
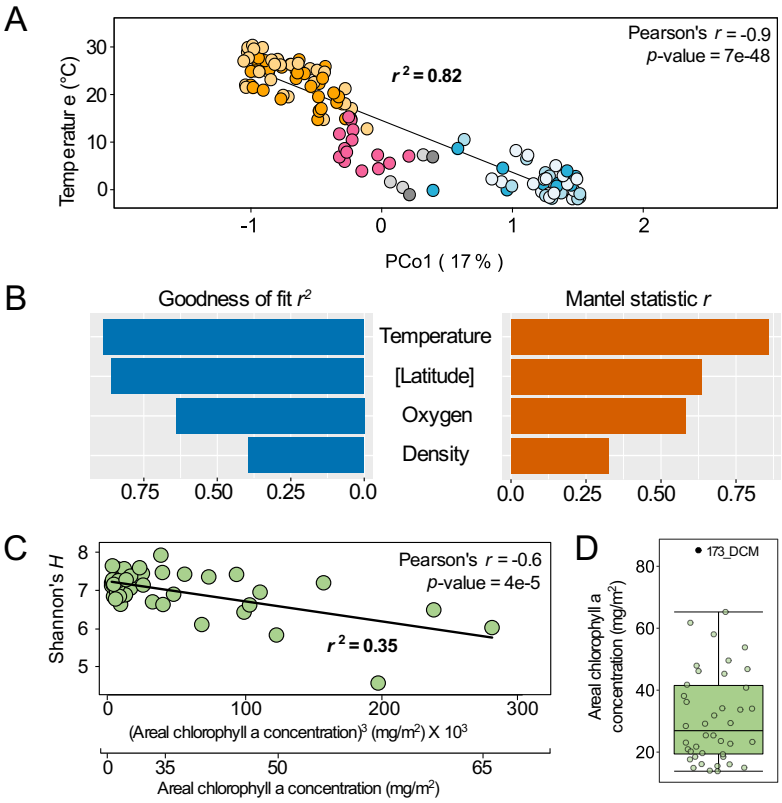
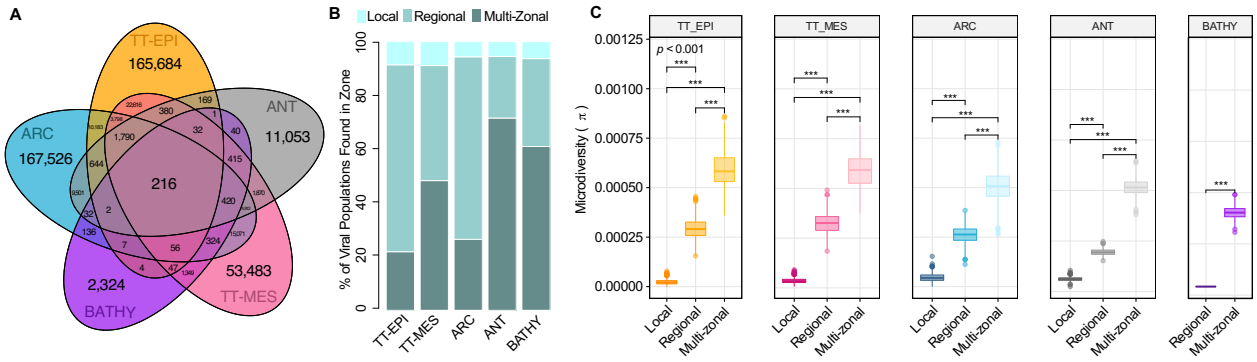


Fig. 4. Viral communities partition into five ecological zones with different *macro-* and *micro-* diversity levels.



770 **Fig. 5. Ecological drivers of global viral *macro*diversity.**



775 **Fig. 6. Size of geographic range positively correlates with *micro*diversity.**

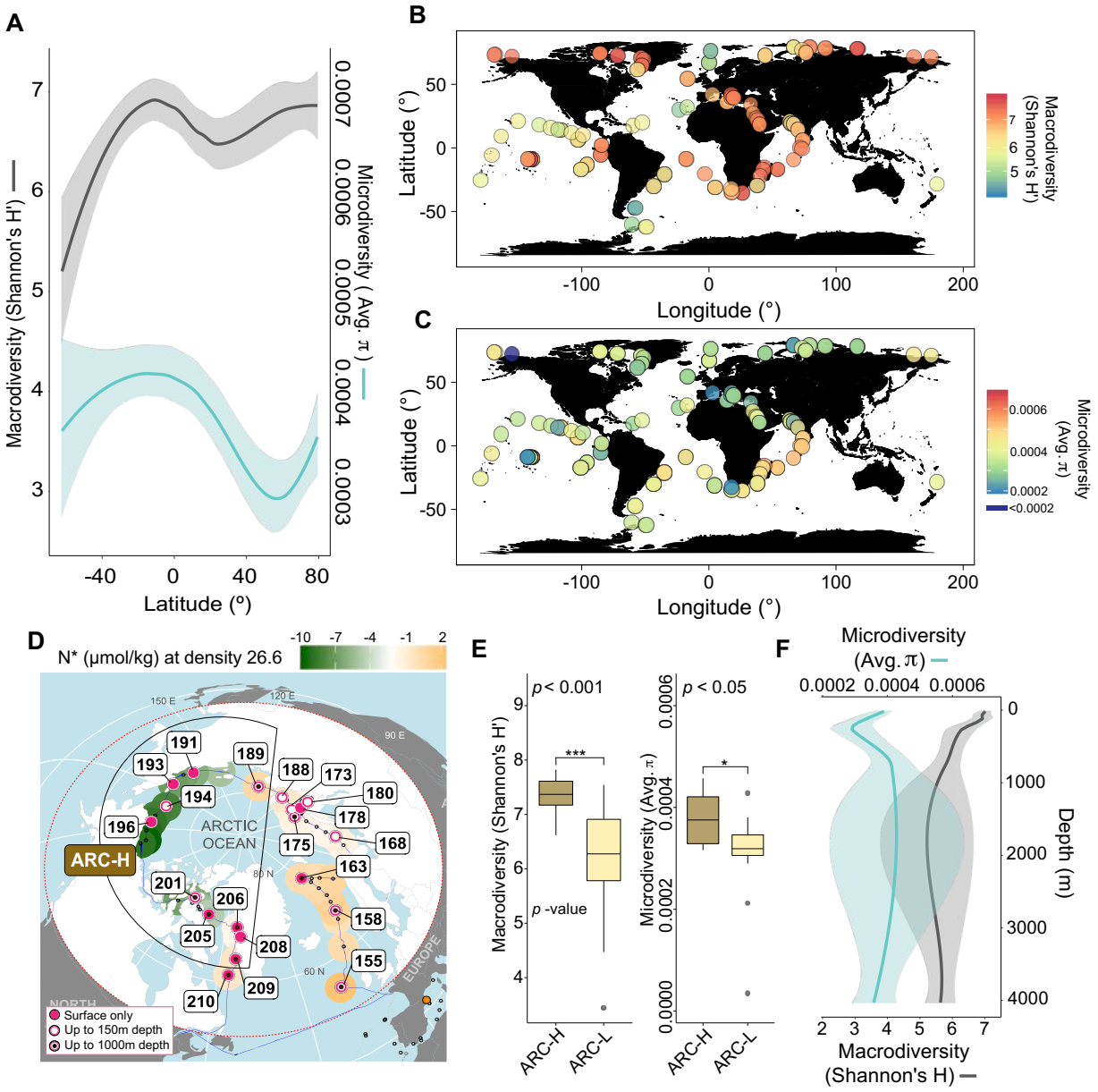


Fig. 7. Viral *macro-* and *micro-* diversity global biodiversity trends.

780

STAR Methods Text

Key Resources Table

785

Reagent or Resource	Source	Identifier(s)
Sequencing Reagents and Kits		
NEBNext DNA Sample Prep Master Mix	New England Biolabs, Ipswich, MA	Cat n° E6040S
NEXTflex PCR free barcodes	Bioo Scientific, Austin, TX	Cat n° NOVA-514110
Kapa Hifi Hot Start Library Amplification kit	KAPA Biosystems, Wilmington, MA	Cat n° KK2611
DNA SMART ChIPSeq Kit	Takara Bio USA, Mountain View, CA	Cat N° 634865
Deposited Data		
Tara Oceans Viromes Raw Reads	Brum *et al.*, 2015; Roux *et al.*, 2016	European Nucleotide Archive (ENA) - see **Table S3** for details
Tara Oceans Polar Circle Raw Reads	This paper	European Nucleotide Archive (ENA) - see **Table S3** for details
Malaspania Viromes Raw Reads	Roux *et al.*, 2016	Integrated Microbial Genomes (IMG) with Joint Genome Institute - see **Table S3** for details
16S rRNA gene *Tara* Oceans data	Logares *et al.*, 2014	Supplementary materials in Logares *et al.*, 2014
Biogeographical and Physicochemical data	Pesant *et al.*, 2015	PANGAEA (Data Publisher for Earth & Environmental Science) - see **Table S3** for details
N* Arctic Data	This paper	**Table S3**

Software and Algorithms		
nucmer (MUMmer3.23)	Kurtz *et al.*, 2004	https://sourceforge.net/projects/mummer/
bbmap 37.57	https://jgi.doe.gov/data-and-tools/bbtools/	https://jgi.doe.gov/data-and-tools/bbtools/
metaSPAdes 3.11	Nurk *et al.*, 2017	https://github.com/ablab/spades/releases
prodigal 2.6.1	Hyatt *et al.*, 2010	https://github.com/hyattpd/Prodigal
diamond	Buchfink *et al.*, 2014	https://github.com/bbuchfink/diamond
VirSorter v2	Roux *et al.*, 2015	https://github.com/simroux/VirSorter
VirFinder	Ren *et al.*, 2017	https://github.com/jessieren/VirFinder
CAT	Cambuy *et al.*, 2016	https://github.com/dutilh/CAT
blast 2.4.0+	ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/	ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/
vConTACT2	Jang *et al.*, *in press* 2018	https://bitbucket.org/MAVERICLab/vcontact2
bowtie2	Langmead & Salzberg, 2012	https://github.com/BenLangmead/bowtie2
BamM	https://github.com/Ecogenomics/BamM	https://github.com/Ecogenomics/BamM
Bedtools	Quinlan & Hall, 2010	https://github.com/arq5x/bedtools2/blob/master/docs/content/overview.rst
Vegan (R package)	Dixon, 2003	https://cran.r-project.org/web/packages/vegan/index.html
BiodiversityR (R package)	https://cran.r-project.org/web/packages/Biodiversity	https://cran.r-project.org/web/packages/Bi

	R/index.html	odiversityR/index.html
heatmap3 (R package)	https://cran.r-project.org/web/packages/heatmap3/index.html	https://cran.r-project.org/web/packages/heatmap3/index.html
ggplot2 (R package)	https://cran.r-project.org/web/packages/ggplot2/index.html	https://cran.r-project.org/web/packages/ggplot2/index.html
ggpubr (R package)	https://cran.r-project.org/web/packages/ggpubr/index.html	https://cran.r-project.org/web/packages/ggpubr/index.html
Analyses scripts (per Figure)	This paper	https://bitbucket.org/MAVERICLab/GOV2

Contact for Reagent and Resource Sharing

Further information and requests for resources and reagents should be directed to and will be fulfilled by the corresponding contact, Matthew Sullivan (mbsulli@gmail.com).

Experimental Model and Subject Details

Not applicable.

Methods Details

Tara Oceans Polar Circle (TOPC) expedition sample collection, processing, and sequencing

Between June 2013 and December 2013, 41 samples were collected at different depths from 20 different sites near or within the Arctic Ocean (see full list of samples in **Table S3**). Physicochemical measurements, sample collection, and DNA extractions were performed using the methods described in (Roux *et al.*, 2016). Extracted DNA was prepared for sequencing using library preparation method described in (Alberti *et al.*, 2017) for viral samples collected during the *TOPC* campaign (section 4.2) and sequenced using the HiSeq 2000 system (101 bp, paired end reads). Importantly, our sample collection and library preparation methods have known bias towards <0.2um dsDNA viruses (Roux *et al.*, 2017). The *TOPC* samples were combined with the previously published viromes in (Brum *et al.*, 2015; Roux *et al.*, 2016). Of the previously published dataset, the mesopelagic samples at (*Tara* stations 37, 39, 56, 68, 70, 76, 78, 111, 122, 137, 138) and the Southern Ocean samples (*Tara* stations 82_DCM, 84, 85) were sequenced deeper. These combined samples comprise the GOV 2.0 dataset. The number of reads found in each sample can be found in **Table S3**.

Due to different library preparation for the *TOPC* samples than the original *Tara* Oceans samples, the previously sequenced mesopelagic samples (*Tara* stations 68, 78, 111, 137) were prepped using the *TOPC* library preparation to determine if it impacted our ability to assemble viral populations. We found no significant difference between library preparations in terms of the number of viral genomes assembled and the average genome length (**Fig. S9A & B**). Additionally, to directly assess the impact of experimental variation between *Tara* Oceans and *TOPC* on our ecological interpretations, we applied hierarchical clustering on a Bray-Curtis

dissimilarity matrix of our viromes and we found that all of the mesopelagic samples prepared using the *TOPC* protocols clustered with their respective samples prepared using the original *Tara* Ocean protocols, and the variation between them was far less than the ecological variation across our viromes (see distances in hierarchical clustering in **Fig. S9D**). For two surface samples (*Tara* Stations 100 and 102), we also re-prepped the DNA using the DNA SMART ChIP-Seq kit which allows us to catch ssDNA in the library preparation (Takara) and further sequenced these two samples using the HiSeq 2000 system.

While the *Tara* Oceans and *Malaspina* expeditions used the same sampling and storage approaches (described in Roux *et al.*, 2016), the sequencing reads were longer for the latter (101 bp for *Tara* and 151 bp for *Malaspina*). Given this, we have performed further analyses to evaluate whether the contribution of this experimental method variation surpasses the ecological variation presented in this study or not. These analyses, which are further described below, showed that ecological variation much better explained the data than experimental methods. To evaluate this, we compared the deep ocean samples collected from the *Tara* Oceans and *Malaspina* expeditions to assess their power to predict the correct ecological zone (mesopelagic or bathypelagic) based on the depth of collection (ecological variation) and the sequencing read length (experimental variation). Using three different metrics, namely the r^2 value in a univariate regression analysis, the bayesian information criterion (BIC) of such constructed univariate model, and the *p*-value associated with different components in a multivariate regression analysis, we found that the depth of collection, rather than the experimental variation, best predicts the ecological zone (higher r^2), with a better model fit (lower BIC), and lower *p*-value (**Fig. S9C**). Additionally, we have one *Malaspina* sample from the mesopelagic ecological zone (the rest are *Tara* samples), and there is no significant difference between the *Malaspina* sample and Tara samples in the mesopelagic (**Fig. S3C** and **D**). Together these findings demonstrate that the differences between the samples collected during the different expeditions are predominantly the result of ecology and community structure rather than experimental artifact.

All the remaining STAR Methods we used are quantifications and statistical analyses. All the details related to these STAR Methods are therefore provided in the following section,

Quantification and Statistical Analyses

Quantification and Statistical Analyses

Viral contig assembly, identification, and dereplication

All samples in the GOV 2.0 dataset (Roux *et al.*, 2016) as well as the previously sequenced *TOPC* library-prepped mesopelagic samples and the DNA SMART ChIP-Seq kit surface samples were individually assembled using metaSPAdes 3.11.1 (Nurk *et al.*, 2017). Prior to assembly, *Malaspina* samples from GOV 2.0 were further quality controlled. Briefly, adaptors and Phix174 reads were removed and reads were trimmed using bbduk.sh (https://jgi.doe.gov/data-and-tools/bbtools/; minlength=30 qtrim=rl maq=20 maxns=0 trimq=14 qtrim=rl). Following assembly, contigs ≥1.5kb were piped through VirSorter (Roux *et al.*, 2015) and VirFinder (Ren *et al.*, 2017) and those that mapped to the human, cat or dog genomes were removed. Contigs ≥5kb or ≥1.5kb and circular that were sorted as VirSorter categories 1-6 and/or VirFinder score ≥0.7 and $p < 0.05$ were pulled for further investigation. Of these contigs, those sorted as VirSorter categories 1 and 2, VirFinder score ≥0.9 and $p < 0.05$ or were identified as viral by both VirSorter (categories 1-6) and VirFinder (score ≥0.7 and $p < 0.05$) were classified as viral. The remaining contigs were run through CAT (Cambuy *et al.*, 2016) and those with <40%

(based on an average gene size of 1000) of the genome classified as bacterial, archaeal, or eukaryotic were considered viral. In total, 848,507 viral contigs were identified. Viral contigs were grouped into populations if they shared ≥95% nucleotide identity across ≥80% of the genome (*sensu* Brum *et al.*, 2015) using nucmer (Kurtz *et al.*, 2004). This resulted in 488,130 total viral populations found in GOV 2.0 (see **Table S5** for VirSorter, VirFinder, and CAT results), of which 195,728 were ≥10kb.

Viral taxonomy

For each viral population, ORFs were called using Prodigal (Hyatt *et al.*, 2010) and the resulting protein sequences were used as input for vConTACT2 (Jang *et al.*, *in press* 2018) and for blastp. Viral populations represented by contigs >10kb were clustered with Viral RefSeq release 85 viral genomes using vConTACT2. Those that clustered with a virus from RefSeq based on amino acid homology based on diamond (Buchfink *et al.*, 2015) alignments were able to be assigned to a known viral taxonomic genus and family. For GOV 2.0 viral populations that could not be assigned taxonomy or were <10kb, family level taxonomy was assigned using a majority-rules approach, where if >50% of a genome's proteins were assigned to the same viral family using a blastp bitscore ≥50 with a Viral RefSeq virus, it was considered part of that viral family.

Viral population boundaries

To determine if our viral populations had discrete sequence boundaries, all reads across the GOV 2.0 dataset (excluding the *Tara* stations 68, 78, 111, 137 prepped using the *TOPC* library preparation methods and the DNA SMART ChIP-Seq kit prepped libraries) were pooled and mapped non-deterministically to our viral populations using the 'very-sensitive-local' setting in bowtie2 (Langmead & Salzberg, 2012). The percent nucleotide identity (% ID) of each mapped read and the positions in the genome where the read mapped were determined. The frequency of reads mapping at a specific % IDs were weighted based on the length of each read mapped across the genomes. Frequencies of reads mapping at specific % IDs were smoothed using Loess smooth functions (span = 1 to be more permissive of lower % ID reads) to create read frequency histograms (% ID vs. frequency). To determine break in the distribution of read frequencies between the different % IDs, Euclidean distances calculated were calculated between % ID frequencies and then hierarchically clustered in R.

Calculating viral population relative abundances, average read depths, and population ranks

To calculate the relative abundances of the different viral populations in each sample, reads from each GOV 2.0 virome were first non-deterministically mapped to the GOV 2.0 viral population genomes using bowtie2. BamM (https://github.com/ecogenomics/BamM) was used to remove reads that mapped at <95% nucleotide identity to the contigs, bedtools genomecov (Quinlan & Hall, 2010) was used to determine how many positions across each genome were covered by reads, and custom Perl scripts were used to further filter out contigs without enough coverage across the length of the contig. For downstream *macro*diversity calculations, contigs ≥5kb in length that had <5kb coverage or less than the total length of the contig covered for contigs <5kb were removed. For downstream *micro*diversity calculations, all contigs with <70% of the contig covered were removed. BamM was used to calculate the average read depth ('tpmean' -minus the top and bottom 10% depths) across each contig. For the *macro*diversity calculations, the average read depth was used as a proxy for abundance and normalized by total

read number per metagenome to allow for sample-to-sample comparison. The rank abundance of all the viral populations was calculated using the normalized abundances and the 'rankabundance' in the BiodiversityR R package.

Subsampling reads

Unequal sequencing depth can have large impacts on diversity measurements, specifically α-diversity measurements (Lemos *et al.*, 2011). Due to 5x more sequencing depth in *TOPC* samples and the deeply sequenced mesopelagic and Southern Ocean samples (**Table S3**), all viromes in the GOV 2.0 dataset were randomly subsampled without replacement to 20M reads for *Tara* or 10M reads for *Malaspina* (as many *Malaspina* samples were <20M reads and there was no significant difference between the 10M and 20M reads assemblies; p = 1) using reformat.sh from bbtools suite (https://sourceforge.net/projects/bbmap/). The subsampled read libraries were assembled using metaSPAdes 3.11.1. Contigs ≥1.5kb that shared ≥95% nucleotide identity across ≥80% of the genome with the 488,130 viral populations in GOV 2.0 were pulled out and grouped into populations to be used as the subsampled GOV 2.0 viral populations. In total, there were 46,699 viral populations. Relative abundances were calculated per sample as aforementioned for *macro*diversity calculations, but using the subsampled GOV 2.0 viral populations and the subsampled reads.

Macrodiversity calculations

The *macro*diversity α- (Shannon's *H*) and β- (Bray-Curtis dissimilarity) diversity statistics were performed using vegan in R (Dixon, 2003). The α-diversity calculations were based on the relative abundances produced from the subsampled reads. Loess smooth plots with 95% confidence windows in ggplot2 in R were used to look at changes in Shannon's *H* across latitude (**Fig. 7A)** and depth (**Fig. 7F)**. For the β-diversity, both the subsampled and the total reads abundances were used to look at community structure (**Fig. S3**). Principal Coordinate analysis (function capscale of vegan package with no constraints applied) and NMDS analysis (function metaMDS; K=2 and trymax=100) were used as the ordination methods on the Bray-Curtis dissimilarity matrices from both the subsampled and total reads calculated from GOV 2.0 (function vegdist; method "bray") after a cube root transformation (function nthroot; n=3). The ecological zones that emerged were verified using a permanova test (function "adonis") and the confidence intervals were plotted using function "ordiellipse" at the specified confidence limits (95% and 97.5%) using the standard deviation method. There were no significant differences in clustering between the subsampled and all reads Bray-Curtis dissimilarity PCoA plots (**Fig. S3**). Hierarchical clustering (function pvclust; method.dist="cor" and method.hclust="average") was conducted on the same Bray-Curtis dissimilarity matrices using 1000 bootstrap iterations and only the approximately unbiased (AU) bootstrap values were reported. The heatmaps were generated using the heatmap3 package with appropriate rotations of the branches in the dendrograms. Samples that did not cluster with their ecological zone (*Tara* mesopelagic stations 72, 85, and 102 and *Tara* surface station 155) were considered outliers and removed from further analyses (**Fig. S3A & C**).

Microdiversity calculations

Viral populations with an average read depth of ≥10x across 70% of their representative contig in at least one sample in the GOV 2.0 dataset were flagged for *micro*diversity analyses. We used 10x as the minimum coverage because population genetic statistics were found to be

relatively consistent down to 10x based on previous downsampling coverage analyses (Schloissnig *et al.*, 2013). BAM files containing reads mapping at ≥95% nucleotide identity were filtered for just the flagged viral populations. Samtools mpileup and bcftools were used to call single nucleotide variants (SNVs) across these populations. SNV calls with a quality call > 30 threshold were kept. Coverage for each allele for each SNV locus was summed across all the metagenomes. For each SNV locus, the consensus allele was re-verified and those with alternative alleles that had a frequency >1% (1000 Genomes Project Consortium, 2012), the classical definition of a polymorphism, and supported by at least 4 reads were considered SNP loci (Schloissnig *et al.,* 2013). Nucleotide diversity (π) per genome were calculated using equation from (Schloissnig *et al.*, 2013). Due to the variable coverage across the genome, coverage was randomly downsampled to 10x coverage per locus in the genome. For the downsampling, if there was not the target 10x coverage for the locus, all of the alleles were sampled. Nucleotide diversity (π) was calculated for each genome with an average read depth ≥10x across 70% of their contig in each sample. For each sample, π values of 100 viral populations were randomly selected and averaged. This was repeated 1000x and the average of the all 1000 subsamplings was used as the final microdiversity value for each sample. Loess smooth plots with 95% confidence windows in ggplot2 in R were used to look at changes in average π across latitude (**Fig. 7A)** and depth (**Fig. 7F)**.

Annotating Genes & Making Protein Clusters

Genes were annotated by translating the sequences into proteins and running a combination of reciprocal best blast hit analyses against the KEGG database (Kanehisa *et al.*, 2002), and blast against the UniProt Reference Clusters database (Suzek *et al.*, 2007), searching for matches against the InterPro protein signature database using InterProScan (Zdobnov *et al.*, 2001), and running HMM searches against Pfams (Bateman *et al.*, 2004). A diamond 'blastall' alignment search (Buchfink *et al.*, 2015) of all the protein sequences was performed against all the protein sequence was performed and the protocol "Clustering similarity graphs encoded in BLAST results" with a granularity of I=2 from the MCL website (https://micans.org/mcl/; Enright *et al.*, 2002) was used to create protein clusters.

Selection Analyses

Natural selection (pN/pS) was calculated using the method from (Schloissnig *et al.*, 2013). The pN/pS method compares the expected ratio of non-synonymous and synonymous substitutions based on a uniform model of occurrence of mutations across the genome with the observed ratio of non-synonymous and synonymous substitutions. The original method treats each SNP locus as independent from each other. Thus, if two SNPs occur in the same codon, the alternate codon produced from each SNP would be considered in the pN/pS calculation. Thus, if two SNPs occur in one codon, the effect of the SNPs could potentially cancel each other out or amplify a non-synonymous signal leading to false positive selection calls. In order to minimize this bias, SNPs found within the same codon in the same gene were tested for linkage in each metagenome. If SNP alleles from loci within the same codon had depth coverage within 15% of each other within each metagenome, they were considered linked in that sample.

For each codon with SNP loci in a gene, the minimum coverage was identified based on the lowest read depth coverage among the three base pair position. The initial number of the consensus codon was determined based on the lowest coverage of the consensus alleles at the SNP locus or loci if linked. The initial numbers of potential alternate codons was based on the

coverage of the alternate allele at that position or the lowest coverage between two linked SNPs. The final coverage of the each codon per SNP locus was calculated by taking the rounded down number of the product of the initial number x (initial number/ minimum coverage for the codon). These codons then subsampled down to 10x. The number of observed non-synonymous and synonymous substitutions were counted and pN/pS was calculated. Genes were considered under positive selection if pN/pS was >1.

Drivers of Macro- and Micro-diversity

Regression analysis between the first coordinate of the PCoA (**Fig. 5A)** and available temperature measurements was conducted using the lm function in R. The environmental variables were fitted to the first two dimensions of the PCoA using a generalized additive model (function envfit; permutations=9999 and na.rm = TRUE). Then, they were correlated with all the PCoA dimensions using a mantel test (function mantel; permutations=9999 and method="spear") after scaling (function scale) and calculating their distance matrices (function vegdist; method "euclid" and na.rm = TRUE). Finally, they were correlated with Shannon's *H* and π using Pearson's correlation (function cor; use="pairwise.complete.obs") after removing Shannon's *H* outliers based on a boxplot analysis (**Fig. S4**).

Subsampling macro- and micro- diversity

Due to unequal sampling across each ecological zone, we chose to normalize the number of samples between each ecological zone by subsampling the down to lowest zone sample size (ANT; n = 5). Shannon's *H* outliers were not included in the subsampling. Five samples within each zone were randomly subsampled without replacement and their *macro*- and *micro*- diversity values averaged, respectively. We subsampled 1000x and plotted the averages and assessed for significant differences using Mann-Whitney U-tests in ggboxplot from the R package ggpubr (**Fig. 4B**).

Classifying multi-zonal, regional, and local viral populations

To determine geographic range, viral populations were evaluated for their distributions across the five ecological zones and plotted using the VennDiagram package in R (**Fig. 6A**). If present in ≥1 sample in more than one ecological zone, it was considered multi-zonal (58% GOV 2.0 viral populations). If present only in samples found within a single zone, it was considered zone-specific (48% GOV 2.0 viral populations). Zone-specific viral populations were further divided into regional (≥2 samples within a zone) and local (only 1 sample within a zone). The proportion of multi-zonal, regional, and local viral populations found across each zone (**Fig. 6B**) and across each station (**Fig. S6**) were calculated by dividing the number of each type by the total number of viral populations found across a zone or station, respectively. To assess the impact of geographic range on *micro*diversity per zone, stations were randomly subsampled without replacement as described above. Within each sample, π values of 50, 100, and 20 viral populations of each geographic distribution (multi-zonal, regional, and local, respectively) were randomly selected and averaged. All the viral populations with a geographic range were sampled and averaged in samples that lacked enough deeply-sequenced viral populations with particular geographic range. This was repeated 1000x and the averages plotted and assessed for significant differences using Mann-Whitney U-tests in ggboxplot from the R package ggpubr (**Fig. 6C**).

Comparing ARC-H and ARC-L

The ARC-H and ARC-L regions were defined based on their biogeography; the ARC-H stations were located in the Pacific Arctic region, the Arctic Archipelago, and the Davis-Baffin Bay, in addition to one station (Station 189) in the Kara-Laptev sea, which was separated by a land mass from the rest of the stations in the same area (**Fig. 7D**). The ARC-L stations were located in the Kara-Laptev Sea (except Station 189), the Barents Sea, and subpolar areas (stations 155 and 210). The departure from the dissolved N:P stoichiometry in the Redfield ratio (N*) was calculated as in (Tremblay *et al.*, 2015) to represent the deficit in dissolved inorganic nitrogen (DIN) in the ratio and as a geochemical tracer of pacific and atlantic water masses. *Macro*- and *micro*- diversity values for each station in ARC-H and ARC-L were plotted and assessed for significant differences using Mann-Whitney U-tests in ggboxplot from the R package ggpubr (**Fig. 7E**).

Comparing GOV to GOV 2.0

Viral populations assembled in the GOV (Roux *et al.*, 2016) were compared to the GOV 2.0 viral populations (**Fig. 1B**) using blastn. Unbinned GOV viral populations with a nucleotide alignment to a GOV 2.0 viral populations with ≥95% nucleotide identity and an alignment length ≥50% the length were considered present in the GOV 2.0. These results were plotted in a venn diagram using the VennDiagram package in R. The frequency of contig lengths of viral populations that were shared across both samples were plotted using ggplot2 (function "geom_histogram"; binwidth =5000).

Calculating 16S OTU Macrodiversity

Previously published 16S OTU data were taken from (Logares *et al.*, 2014). The *macro*diversity α- (Shannon's *H*) statistics were performed using vegan in R (Dixon, 2003). Loess smooth plots with 95% confidence windows in ggplot2 in R were used to look at changes in bacterial Shannon's H down the depth gradient. Differences between surface, deep chlorophyll maximum, and mesopelagic bacterial samples were compared using Mann-Whitney U-tests and plotted in ggboxplot from the R package ggpubr. Finally, viral *micro*diversity was correlated with bacterial Shannon's *H* using Pearson's correlation (function cor; use="pairwise.complete.obs") and a linear regression (**Fig. S8C**).

Impact of the coast, depth, and seasons

GOV 2.0 samples are largely open ocean samples. Even though the arctic samples were more coastal, we didn't observe any significant coastal impact on the global *macro*diversity (Pearson's r = -0.25; Bonferroni-corrected p-value = 0.18) and *micro*diversity (Pearson's r = 0.1; p-value = 0.16) levels (**Fig. 4C**). Although nitrate and phosphate levels generally increase with depth, we observed higher negative correlations and significantly lower p-values for these nutrients with *macro*diversity levels than between depth and *macro*diversity (**Fig. 4C**) which suggests an impact of nutrients on viral diversity via primary production (**Fig. 5C**). Additionally, since the sampling was largely at discrete depth layers with different densities in the TT region (epipelagic, mesopelagic, and bathypelagic), rather than sampling gradients, we discerned a clearer signal for the separation between these ecological zones (**Fig. 4A**). On the other hand, all the arctic epipelagic and mesopelagic samples fell within the same ecological zone due to the absence of a pycnocline in this area (**Fig. 4A**). Finally, the circumnavigation of the Arctic Ocean spanned multiple seasons (spring, summer, and fall). Based on our previous observation from a time-series data in a sub-arctic system (Hurwitz & Sullivan, 2013), our viral *macro*diversity is expected to be lowest during the spring and summer and increase towards the winter season.

However, our calculated N* values are not dependant on the season and represent the largest magnitude of change among all of the environmental variables that correlated with *macro*diversity between the ARC-H and ARC-L regions.

Assessment of microbial contamination

To quantifying microbial contamination across our samples, we screened our metagenomic reads using singleM (github.com/wwood/singlem) for 16S sequences using the dedicated 16S SingleM package. We found that our viromes are exceptionally clean. Specifically, the number of 16S sequences in our samples ranged from 0-40 per million reads **(Table S3)**, and hence the samples are considered to have "likely negligible bacterial contamination" according to the metric proposed by authors evaluating such signals in published viromes (threshold was 200 16S sequences per million; Roux *et al.*, 2013). In spite of our viromes being exceptionally clean, we sought to evaluate the impact of any variation in 16S, and hence bacterial contamination, however small, on our findings. We found that even though microbial contamination increases with depth (most probably due to the decrease in cell size; linear regression $r^2 = 0.89$), this increase was driven mainly by the bathypelagic samples. Briefly, the average contamination in BATHY was 28.7 per million reads (standard deviation = 6.8) as compared to the rest of the samples (average contamination = 1.7 per million reads and standard deviation = 2). These bathypelagic samples were not included in any of the ecological driver analyses due to the unavailability of the environmental data to us. Further, it is clear that our estimates of diversity were not influenced by the minor variations in the negligible contamination in our viroomes as a linear regression between Shannon's *H* and the number of 16S reads from deep ocean samples resulted in a negligible r^2 value (0.06). These data (used for conducting the regression analysis) represent a large range of diversity (3.3-7.8) and the full range of contamination (0-40), but avoid the convolution from the ecological difference between the surface and deep ocean layers. Thus, we conclude that the diversity observations we make in this study are driven by ecological variation far greater than microbial contamination.

Data and Software Availability

Code availability

Scripts used in this manuscript are available on the Sullivan laboratory bitbucket under GOV 2.0.

Data availability

All raw reads are available through ENA (*Tara* Oceans and *TOPC*) or IMG (Malapsina) using the identifiers listed in **Table S3**. Processed data are available through iVirus, including all assembled contigs, viral populations and genes.

Author contributions:

MC, CD, JF, SK-L, CM, SPe, MP, SPi, JP, and *Tara* Oceans coordinators conceptualized and organized sampling efforts for the *Tara* Oceans Polar Circle expedition. SPe annotated, curated, and managed all biogeochemical data. AA, CC, and PW coordinated all sequencing efforts. ACG, AAZ, NC-N, BT, BB, KA, GD-H,YL, DV, J-ET, MB, CB, CdV, AC, BED, DI, LK-B, SR, SS, PW, and MBS created the study design, analyzed the data, and wrote the manuscript. All authors approved the final manuscript. **Competing interests:** The authors declare no competing interests.

Acknowledgments:

This global sampling effort was enabled by countless scientists and crew who sampled aboard the *Tara*, as well as the leadership of the *Tara* Expeditions Foundation. Computational support was provided by an award from the Ohio Supercomputer Center (OSC) to MBS. Study design and manuscript comments from Bonnie T. Poulos, Ho Bin Jang, M. Consuelo Gazitúa, Olivier Zablocki, Janaina Rigonato, Damien Eveillard, Frédéric Mahé, Federico Ibarbalz, and Hisashi Endo are gratefully acknowledged. Funding was provided by the Gordon and Betty Moore Foundation (#3790) and NSF (OCE#1536989 and OCE#1829831) to MBS, Oceanomics (ANR-11-BTBR-0008) and France Genomique (ANR-10-INBS-09) to Genoscope, ETH and Helmut Horten Foundation to SS, a Netherlands Organization for Scientific Research (NOWO) Vidi grant 864.14.004 to BED, and an NIH T32 training grant fellowship (AI112542) to ACG.

Materials & Methods References:

- 1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature.* **491**, 56-65.
- Alberti, A., Poulain, J., Engelen, S., Labadie, K., Romac, S., Ferrera, I., Albini, G., Aury, J.M., Belser, C., Bertrand, A., *et al.* (2017). Viral to metazoan marine plankton nucleotide sequences from the *Tara* Oceans expedition. *Sci. Data.* **4**, 170093.
- Angly, F.E., Felts, B., Breitbart, M., Salamon, P., Edwards, R.A., Carlson, C., Chan, A.M., Haynes, M., Kelley, S., Liu, H., *et al.* (2006). The marine viromes of four oceanic regions. *PLOS Biol.* **4.11**, e368.
- Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E.L., *et al.* (2004). The Pfam protein families database. *Nucleic Acids Res.* **32**, D138-141.
- Buchfink, B., Chao, X., Huson, D.H. (2015) Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12.1**, 59-60.
- Cambuy, D.D., Coutinho, F.H., and Dutilh, B.E. (2016). Contig annotation tool CAT robustly classifies assembled metagenomic contigs and long sequences. *BioRxiv*, 072868.
- Dixon, P. (2003). VEGAN, a package of R functions for community ecology. *J. Veg. Sci.* **14.6**, 927-930.
- Enright, A.J., Van Dongen S., and Ouzounis C.A. (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* 30(7), 1575-1584.
- Hurwitz, B.L., and Sullivan, M.B. (2013). The Pacific Ocean virome (POV): a marine viral metagenomic dataset and associated protein clusters for quantitative viral ecology. *PLOS One*. **8.2**, e57355.
- Hyatt, D., Chen, G.L., Locascio, P.F., Land, M.L., Larimer, F.W., and Hauser, L.J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinform.***11**, 119.
- Jang, H-B., Bolduc, B., Zablocki, O., Kuhn, J.H., Adriaenssens, E.M., Krupovic, M., Brister, R., Kropinski, A.M., Koonin, E.V., Turner, D., *et al.* (2018). Gene sharing networks to automate genome-based prokaryotic viral taxonomy, *Nature Biotechnol.* (*in press*).
- Kanehisa, M., Goto, S., Kiwashima, S., and Nakaya, A. (2002). The KEGG databases at GenomeNet. *Nucleic Acids Res.* 30, 42-46.

- Kurtz, S., Phillippy, A., Delcher, A.L., Smoot, M., Shumway, M., Antonescu, C., and Salzberg, S.L. (2004). Versatile and open software for comparing large genomes. *Genome Biol.* **5.2**, R12.
- Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods*. **9.4**, 357-359.
- Lemos, L.N., Fulthorpe, R.R., Triplett, E.W., and Roesch, L.F. (2011). Rethinking microbial diversity analysis in the high throughput sequencing era. *J. Microbial. Methods.* **86.1**, 42-51.
- Logares, R., Sunagawa, S., Salazar, G., Cornejo-Castillo, F.M., Ferrera, I., Sarmento, H., Hingamp, P., Ogata, H., de Vargas, C., Lima-Mendez, G., *et al.* (2014). Metagenomic 16S rDNA Illumina tags are a powerful alternative to amplicon sequencing to explore diversity and structure of microbial communities. *Environ. Microbiol.* **16.9**, 2659-2671.
- Marston, M.F., and Amrich, C.G. (2009). Recombination and microdiversity in coastal marine cyanophages. *Environ. Microbiol.* **11.11**, 2893-2903 (2009).
- Marston, M.F., and Martiny, J.B. (2016). Genomic diversification of marine cyanophages in stable ecotypes. *Environ. Microbiol.* **18.11**, 4240-4253.
- Nurk, S., Meleshko, D., Korobeynikov, A., and Pevzner, P.A. (2017). metaSPAdes: a new versatile metagenomic assembler. *Genome Res.*, gr-213958.
- Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. **26.6**, 841-842.
- Ren, J., Ahlgren, N.A., Lu, Y.Y., Fuhrman, J.A., and Sun, F. (2017). VirFinder: a novel *k*-mer based tool for identifying viral sequences from assembled metagenomic data. *Microbiome*. **5**, 69.
- Roux, S., Emerson, J.B., Eloe-Fadrosh, E.A., and Sullivan, M.B. (2017). Benchmarking viromics: an in silico evaluation of metagenome-enabled estimates of viral community composition and diversity. *PeerJ.* **5**, e3817.
- Roux, S., Krupovic, M., Debroas, D., Forterre, P., and Enault, F. (2013). Assessment of viral community functional potential from viral metagenomes may be hampered by contamination with cellular sequences. *Open Biol.* **3**:130160.
- Sul, W.J., Oliver, T.A., Ducklow, H.W., Amaral-Zettler, L.A., and Sogin, M.L. (2013). Marine bacteria exhibit a bipolar distribution. *Proc. Natl. Acad. Sci. USA*. **110**, 2342-2347.
- Suzek, B.E., Wang, Y., Huang, H., McGarvey, P.B., Wu, C.H., and UniProt Consortium. (2015). UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics.* 31, 926-932.
- Tremblay, J-É., Anderson, L.G., Matrai, P., Coupel, P., Bélanger, S., Michel, C., and Reigstad, M. (2015). Global and regional drivers of nutrient supply, primary production and CO_2 drawdown in the changing Arctic Ocean. *Prog. Oceanogr.* **193**, 171-196.
- Zdobnov, E.M., and Apweiler, R. (2001). InterProScan--an integration platform for the signature-recognition methods in InterPro. *Bioinformatics*. 17, 847-849.
- Zeigler-Allen, L., McCrow, J.P., Ininbergs, K., Dupont, C.L., Badger, J.H., Hoffman, J.M., Ekman, M., Allen, A.E., Bergman, B., and Venter, J.C. (2017). The Baltic Sea virome: diversity and transcriptional activity of DNA and RNA viruses. *mSystems.* **2.1**, e00125-16.
- Zinger, L., Amaral-Zettler, L.A., Fuhrman, J.A., Horner-Devine, M.C., Huse, S.M., Welch, D.B., Martiny, J.B., Sogin, M., Boetius, A., and Ramette, A. (2011). Global

patterns of bacterial beta-diversity in seafloor and seawater ecosystems. *PLOS One*. **6.9**, e24570.

List of Supplementary Materials:

Tara Oceans Coordinators and Affiliations

Figures S1-S9

Tables S1-S10

Supplementary Materials:

Tara Oceans Coordinators and Affiliations

Silvia G. Acinas[1], Marcel Babin[2], Peer Bork[3,4], Emmanuel Boss[5], Chris Bowler[6,29], Guy Cochrane[7], Colomban de Vargas[8,29], Michael Follows[9], Gabriel Gorsky[10,29], Nigel Grimsley[11,12,29], Lionel Guidi[10,29], Pascal Hingamp[13,29], Daniele Iudicone[14], Olivier Jaillon[15,29], Stefanie Kandels-Lewis[3,16,], Lee Karp-Boss[5], Eric Karsenti[6,16,29], Fabrice Not[17,29], Hiroyuki Ogata[18], Stéphane Pesant[19,20], Nicole Poulton[21], Jeroen Raes[22,23,24], Christian Sardet[10,29], Sabrina Speich[25,26,29], Lars Stemmann[10,29], Matthew B. Sullivan[27], Shinichi Sunagawa[28], Patrick Wincker[15,29]

[1]Department of Marine Biology and Oceanography, Institute of Marine Sciences (ICM)-CSIC, Pg. Marítim de la Barceloneta 37-49, E08003 Barcelona, Spain.

[2]Département de biologie, Québec Océan and Takuvik Joint International Laboratory (UMI 3376), Université Laval (Canada) - CNRS (France), Université Laval, Québec, QC, G1V 0A6, Canada.

[3]Structural and Computational Biology, European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany.

[4]Max-Delbrück-Centre for Molecular Medicine, 13092 Berlin, Germany.

[5]School of Marine Sciences, University of Maine, Orono, ME 04469, USA.

[6]Institut de Biologie de l'Ecole Normale Supérieure (IBENS), Ecole normale supérieure, CNRS, INSERM, Université PSL, 75005 Paris, France.

[7]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Welcome Trust Genome Campus, Hinxton, Cambridge, UK.

[8]Sorbonne Université, CNRS, Station Biologique de Roscoff, AD2M ECOMAP, 29680 Roscoff, France.

[9]Department of Earth, Atmospheric, and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139, USA.

[10]Sorbonne Université, CNRS, Laboratoire d'Océanographie de Villefanche, LOV, F-06230 Villefranche-sur-mer, France.

[11]CNRS UMR 7232, Biologie Intégrative des Organismes Marins, Avenue du Fontaulé, 66650 Banyuls-sur-Mer, France.

[12]Sorbonne Universités Paris 06, OOB UPMC, Avenue du Fontaulé, 66650 Banyuls-sur-Mer, France.

[13]Aix Marseille Univ., Université de Toulon, CNRS, IRD, MIO UM 110, 13288, Marseille, France.

[14]Stazione Zoologica Anton Dohrn, Villa Comunale, 80121 Naples, Italy.

[15]Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ Evry, Université Paris-Saclay, 91057 Evry, France.

[16]Directors' Research European Molecular Biology Laboratory Meyerhofstr. 1 69117 Heidelberg, Germany.

[17]Sorbonne Université, CNRS - UMR7144 - Ecology of Marine Plankton Group, Station Biologique de Roscoff, Place Georges Teissier, 29680 Roscoff, France.

[18]Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto 611-0011, Japan.

[19]PANGAEA, Data Publisher for Earth and Environmental Science, University of Bremen, 28359 Bremen, Germany.

[20]MARUM, Center for Marine Environmental Sciences, University of Bremen, 28359 Bremen, Germany.

[21]Bigelow Laboratory for Ocean Sciences, East Boothbay, ME, 04544, USA.

[22]Department of Microbiology and Immunology, Rega Institute, KU Leuven, Herestraat 49, 3000 Leuven, Belgium.

[23]Center for the Biology of Disease, VIB KU Leuven, Herestraat 49, 3000 Leuven, Belgium.

[24]Department of Applied Biological Sciences, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium.

[25]Department of Geosciences, Laboratoire de Météorologie Dynamique (LMD), Ecole Normale Supérieure, 24 rue Lhomond 75231 Paris, Cedex 05, France.

[26]Ocean Physics Laboratory, University of Western Brittany, 6 avenue Victor-Le-Gorgeu, BP 809, Brest 29285, France.

[27]Departments of Microbiology and Civil, Environmental and Geodetic Engineering, The Ohio State University, Columbus, OH, 43210, USA.

[28]Institute of Microbiology, ETH Zurich, Zurich, Switzerland.

[29]Research Federation for the study of Global Ocean Systems Ecology and Evolution, FR2022/*Tara* Oceans GOSEE, 3 rue Michel-Ange, 75016 Paris, France.

Supplementary Figures

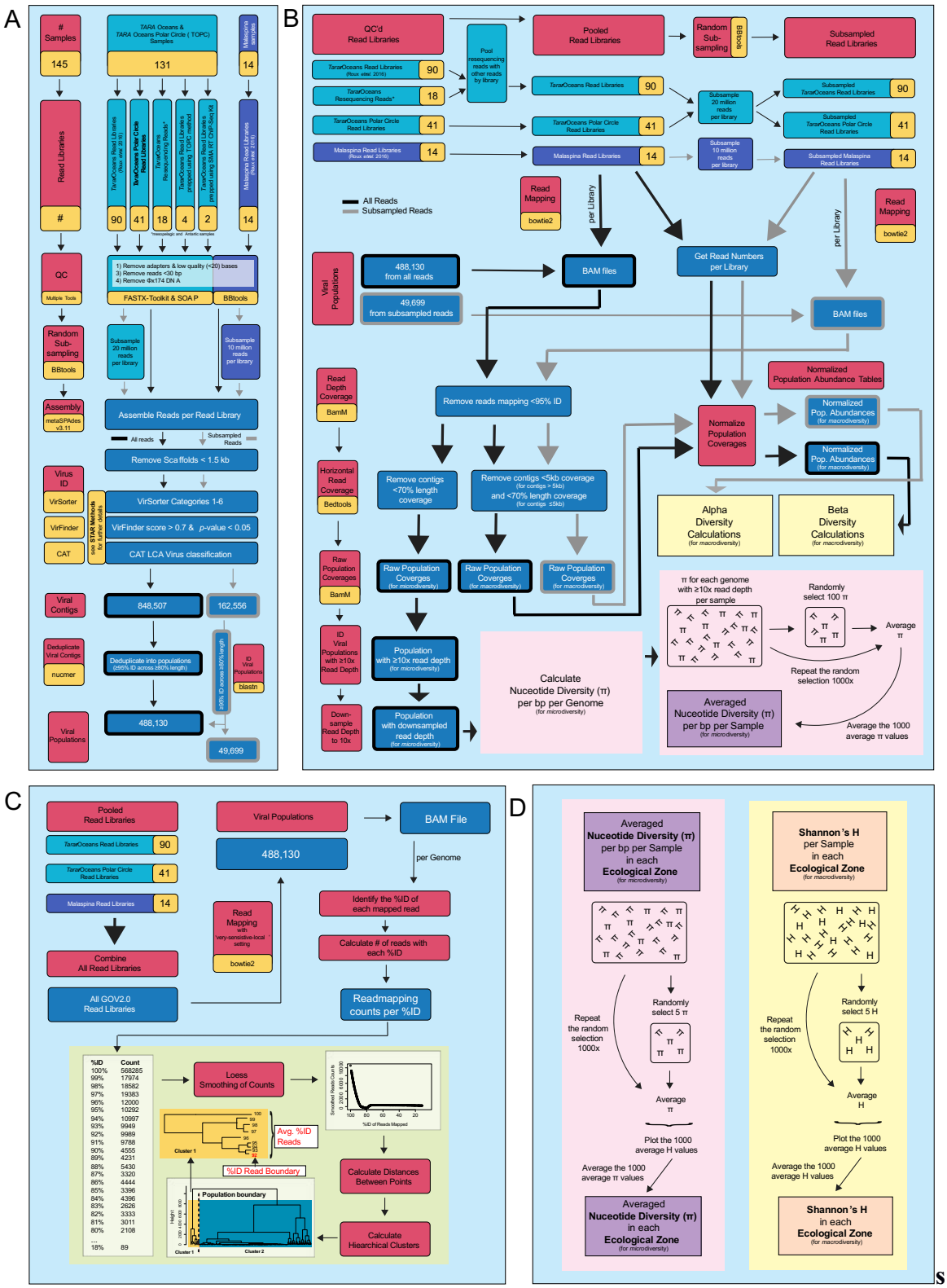


Fig. S1. Bioinformatic workflow. Flow diagrams showing the bioinformatic workflow for **(A)** the assembly and identification of viral populations, **(B)** the population coverages and

abundances and how they were used to calculate *macro*- and *micro*-diversity calculations, **(C)** prediction of population boundaries, and **(D)** how average *macro*- and *micro*-diversity calculations per ecological zone were calculated.

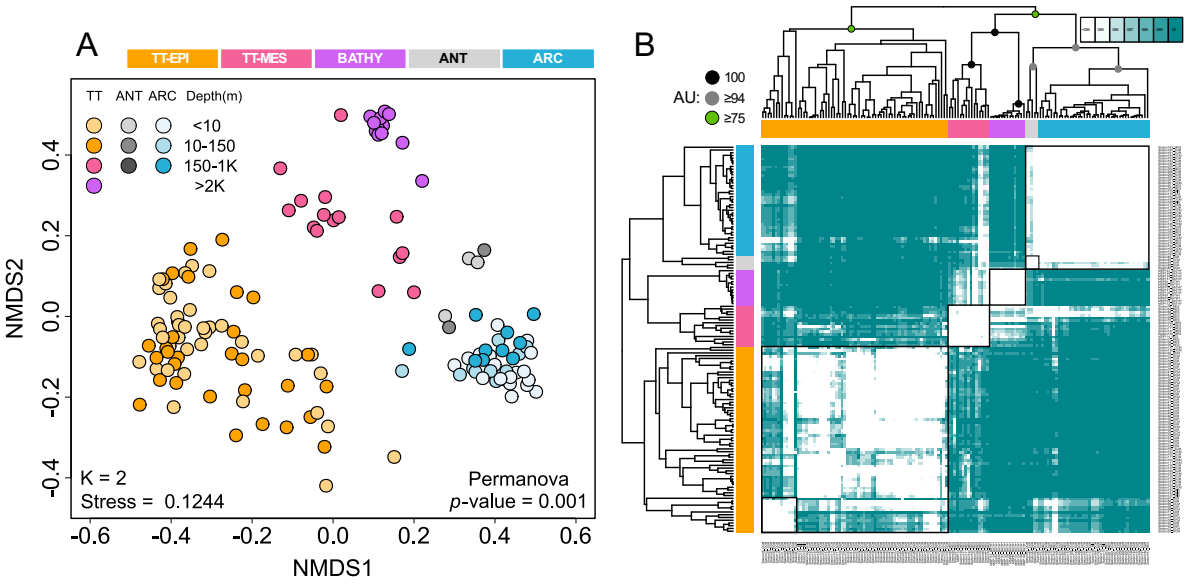


Fig. S2. Non-metric multidimensional scaling (NMDS) and hierarchical clustering of GOV 2.0. As observed with the Principal Coordinate analysis **(Fig. 4A)**, NMDS analysis **(A)** and correlation-based hierarchical clustering **(B)** of a Bray-Curtis dissimilarity matrix calculated from GOV 2.0 structured the viromes into five distinct global ecological zones with an approximately unbiased (AU) bootstrap value $\geq$ 77 in the hierarchical clustering. Four outlier viromes were removed and all the sequencing reads were used, with justification provided in (**Fig. S3, C and D**), respectively. Abbreviations: ARC, Arctic; ANT, Antarctic; BATHY, bathypelagic; TT-EPI, temperate and tropical epipelagic; TT-MES, temperate and tropical mesopelagic.

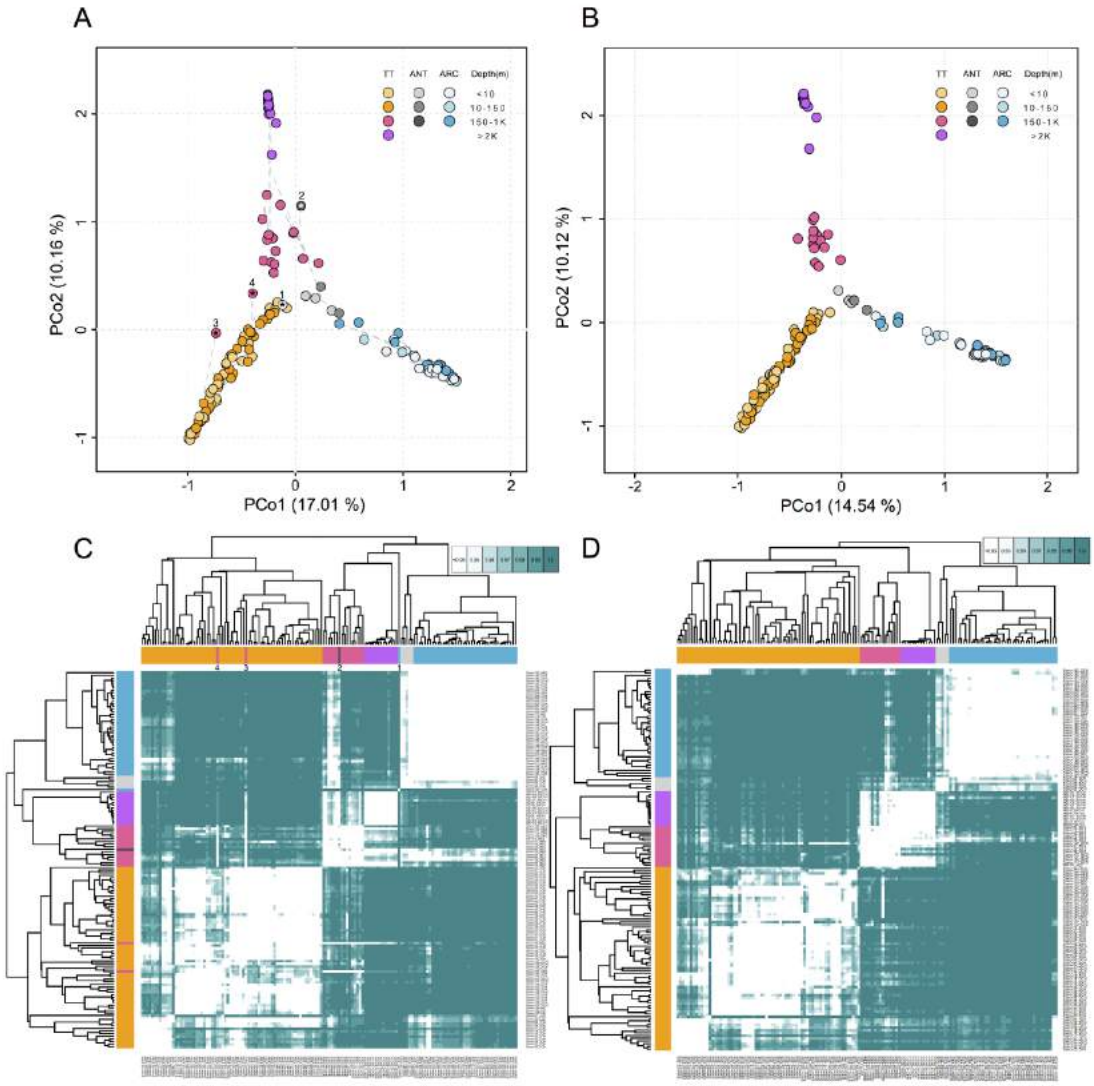


Fig. S3. Beta-diversity of the total reads and subsampled reads GOV 2.0 dataset. PCoA of a Bray-Curtis dissimilarity matrix calculated from GOV 2.0 using all the sequencing reads **(A)** and after randomly subsampling the reads to the same sequencing depth **(B)**. The dissimilarity matrices from **(A)** and **(B)** were used to conduct hierarchical clustering on the samples as shown in **(C)** and **(D)**, respectively. The four viromes which were removed from (**Fig. 4)** and **(Fig. S2)** are highlighted with asterisks; sample 1 (station 155_SUR) is the only surface sample in the North Atlantic Drift Province and could have been influenced by the warm surface currents going northward due to the Atlantic Meridional Overturning Circulation; sample 2 (station 85_MES) is the only mesopelagic sample from the Southern Ocean and could have been influenced by the upwelling of ancient deep ocean water (which is also congruent with the similarity observed between deep water bacterial communities of polar and lower latitude (Ghiglione *et al.*, 2012)); sample 3 (station72_MES) fell outside the 97.5% confidence intervals of all the ecological zones; sample 4 (station102_MES) was located in El Niño-Southern Oscillation region and could have been influenced by the upwellings and downwellings in this area. Additionally, samples 1, 3, and 4 were among the Shannon's *H* outliers (**Fig. S4**). Viral

communities still partitioned into five ecological zones after subsampling the reads as shown by the PCoA **(B)** and hierarchical clustering **(D)** plots.

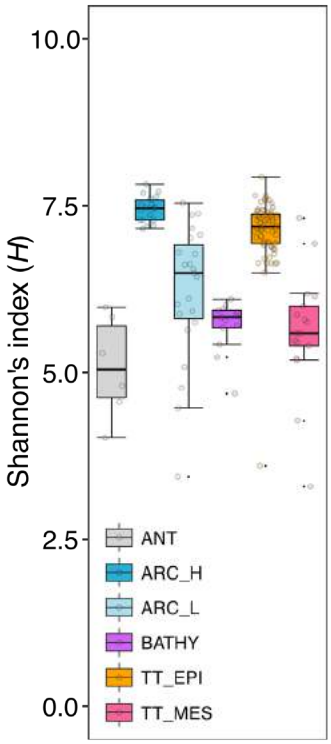


Fig. S4. Boxplot analysis of viral *macro*diversity across GOV 2.0 ecological zones. Outliers that fell below the first quantile or above the fourth quantile (function geom_boxplot of ggplot) of each ecological zone were removed before examining the predictors of viral *macro*diversity (**Fig. 4C**). Outliers: 32_SUR, 155_SUR, 56_MES, 70_MES, 72_MES, 102_MES, MSP131, and MSP144.

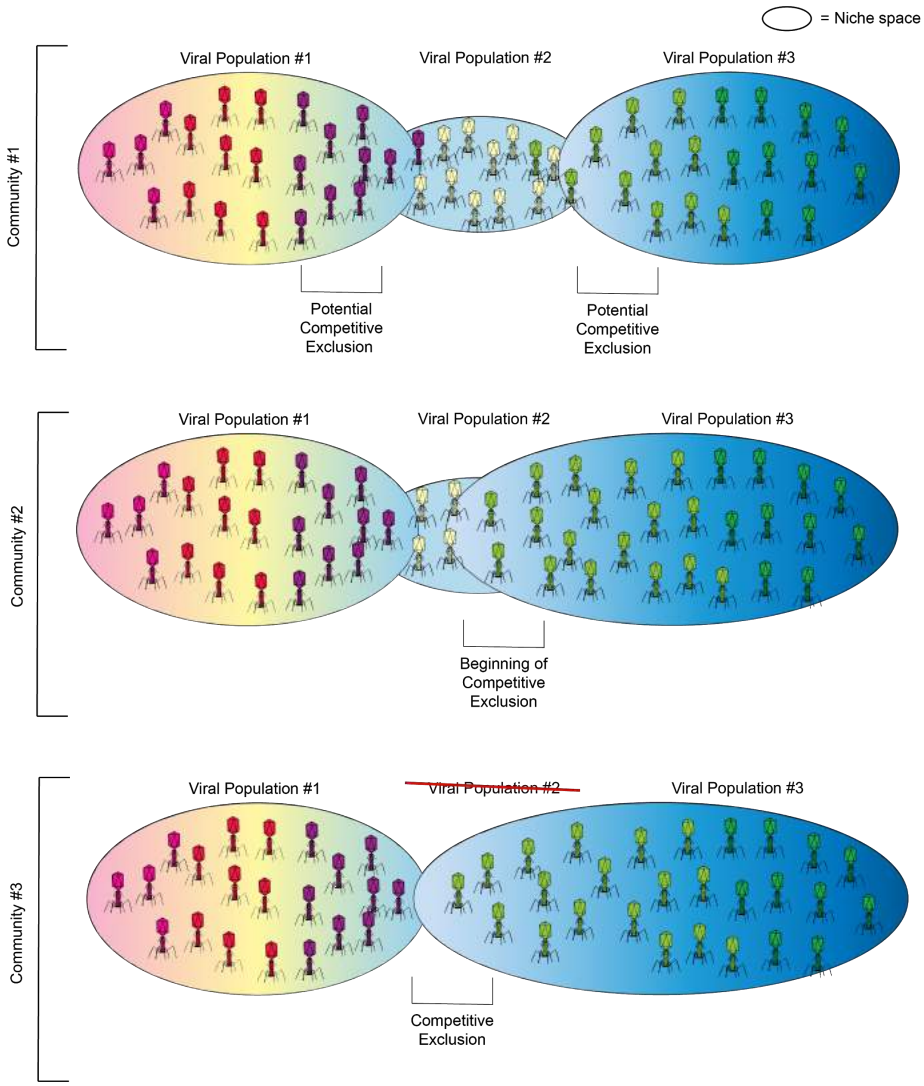
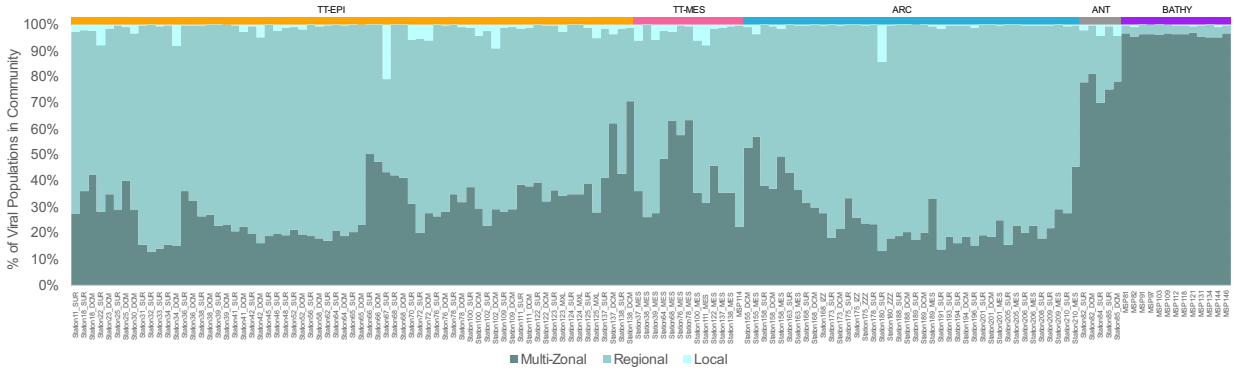


Fig. S5. Schematic showing the interplay of increased *micro*diversity and competitive exclusion. Viral populations with more *micro*diversity usually have larger niche sizes and therefore can outcompete viral populations with smaller overlapping niche sizes. This process of competitive exclusion may not be visible in each community as seen across the three communities. Thus, the average of communities such as across ecological zones can better show this relationship.



1355 **Fig. S6.** Stacked barplots showing the number of multi-zonal, regional, and local viral populations found within the species pool of each station. Ecological zone outliers (see **Fig. S3**) are excluded.

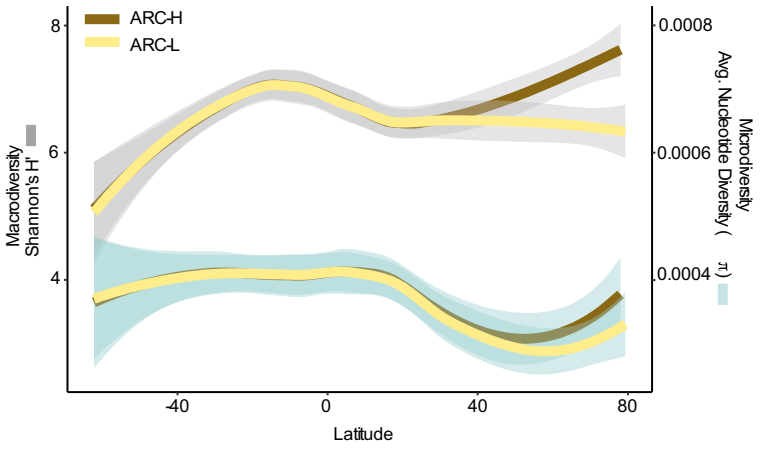


Fig. S7. ARC-H drives the divergence from the latitude diversity gradient. Loess smooth plots showing the latitudinal distributions of *macro*- and *micro*- population diversity with ARC-H and ARC-L regions. The line represents the loess best fit, while the lighter band corresponds to the 95% confidence window of the fit.

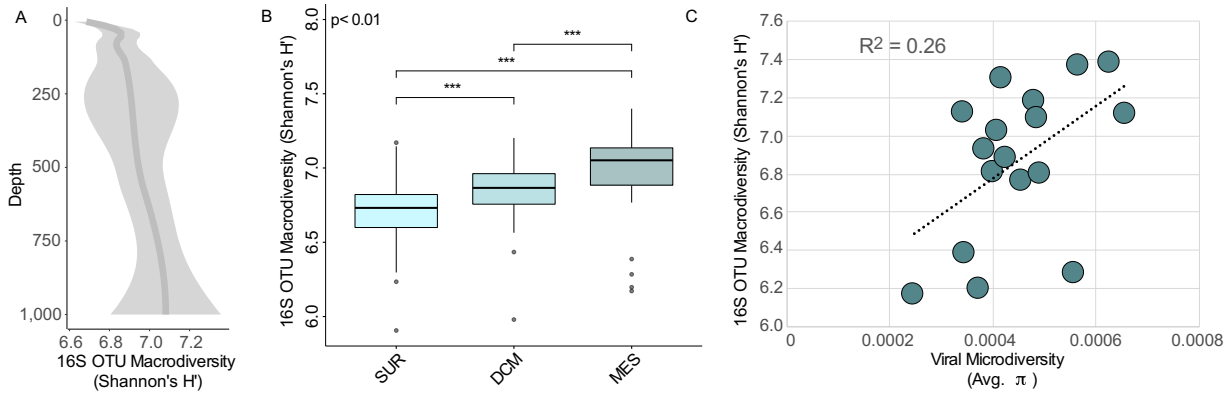


Fig. S8. Microbial 16S OTUs biodiversity deviate from the depth diversity gradient and positively correlates with viral *micro*diversity in the mesopelagic. (A) Loess smooth plots showing 16S OTUs (Logares *et al.*, 2014) *macro*diversity distributions down the depth gradient. The line represents the loess best fit, while the lighter band corresponds to the 95% confidence window of the fit. **(B)** Boxplots showing median and quartiles of surface, deep chlorophyll maximum (DCM), and mesopelagic 16S OTU data taken from (Logares *et al.*, 2014). All pairwise comparisons shown were statistically significant (p<0.05) using two-tailed Mann-Whitney U-tests. **(C)** Scatterplot showing the positive correlation (Pearson's correlation r = 0.51; p-value = 0.036) and linear regression (r^2 = 0.26) between *Tara* Oceans mesopelagic samples shared between the 16S OTU samples in (Logares *et al.*, 2014) and our viral samples in GOV 2.0.

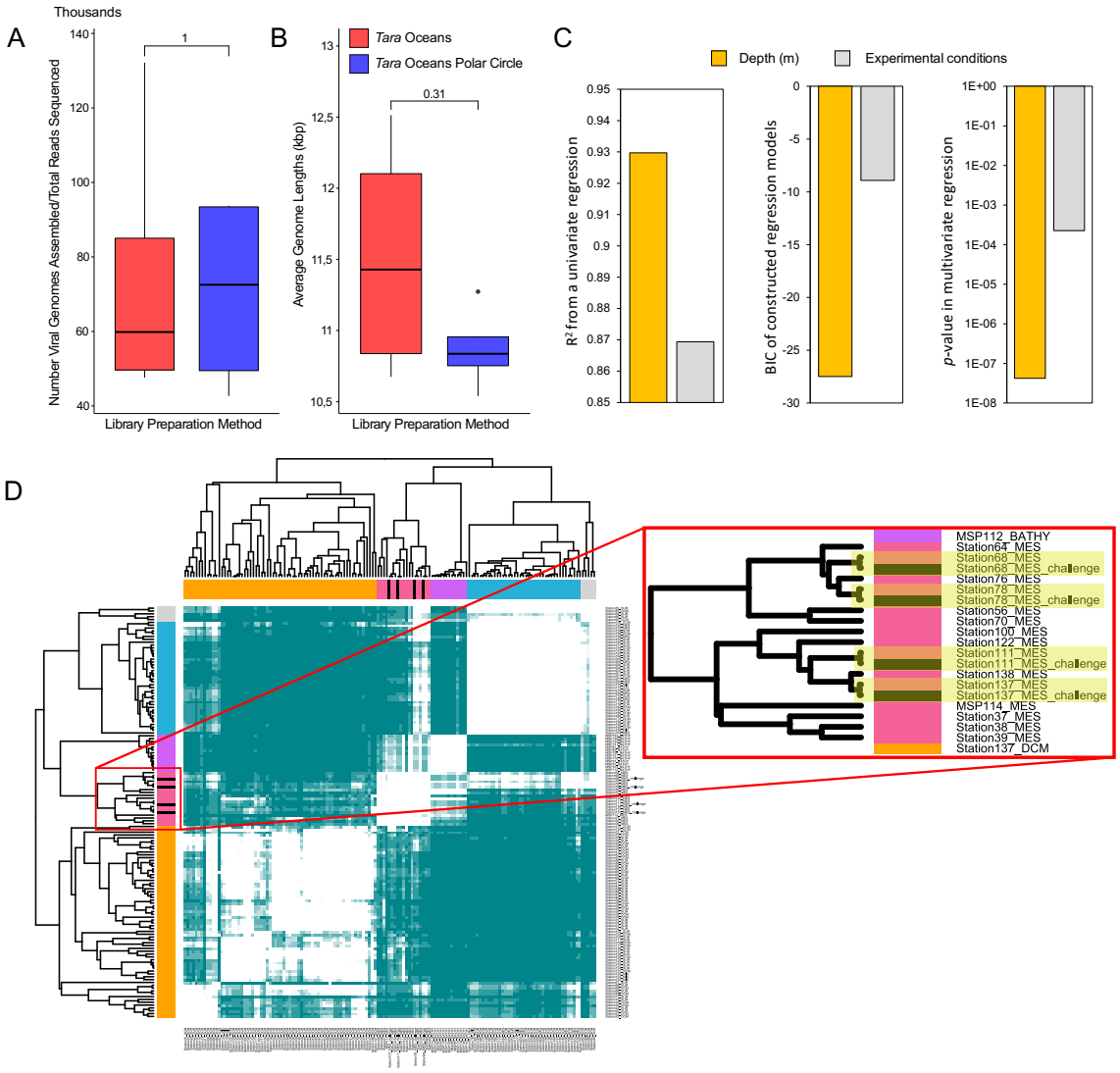


Fig. S9. Library preparation and experimental conditions comparisons. (A & B) Boxplots showing median and quartiles of the number of assembled viral genomes per total reads sequenced and the average genome lengths in *TO* and *TOPC* preparations of *Tara* mesopelagic stations 68, 78, 111, and 137, respectively. All pairwise comparisons shown were not statistically significant using two-tailed Mann-Whitney U-tests. **(C)** Depth (as an ecological variable) predicts the ecological zone of the deep ocean (mesopelagic or bathypelagic) better than experimental variation between *Tara* and *Malaspina* expeditions, with a higher r^2 (left), lower BIC (middle), and lower *p*-value (right). The first two metrics were calculated from a univariate regression analysis (using depth alone or experimental variation alone as a predictor of the ecological zone), while the third metric was calculated from a multivariate multiple regression analysis that uses both depth and experimental variation as predictors. **(D)** Hierarchical clustering of a Bray-Curtis dissimilarity matrix calculated from GOV 2.0 viromes to which four additional viromes (black bars) have been added to control for the impact of experimental variation between the *Tara* Oceans and *Tara* Oceans Polar Circle expeditions. The four viromes prepared using the *Tara* Oceans Polar Circle protocols clustered with their respective original

samples, which were prepared using the *Tara* Oceans protocols indicating that experimental variation was far less than ecological variation.

1400