Diversity and ecological footprint of Global Ocean RNA viruses

Guillermo Dominguez-Huerta, Ahmed Zayed, James Wainaina, Jiarong Guo, Funing Tian, Akbar Adjie Pratama, Benjamin Bolduc, Mohamed Mohssen, Olivier Zablocki, Eric Pelletier, et al.

Title: Diversity and ecological footprint of Global Ocean RNA viruses

Authors: Guillermo Dominguez-Huerta[1,2,3†], Ahmed A. Zayed[1,2,3†], James M. Wainaina[1,3], Jiarong Guo[1,2,3], Funing Tian[1,3], Akbar Adjie Pratama[1,2], Benjamin Bolduc[1,2,3], Mohamed Mohssen[1,3,4], Olivier Zablocki[1,2,3], Eric Pelletier[5,6], Erwan Delage[6,7], Adriana Alberti[5,6§], Jean-Marc Aury[5,6], Quentin Carradec[5,6], Corinne da Silva[5,6], Karine Labadie[5,6], Julie Poulain[5,6], ***Tara* Oceans Coordinators**[‡], Chris Bowler[6,8], Damien Eveillard[6,7], Lionel Guidi[6,9], Eric Karsenti[6,8,10], Jens H. Kuhn[11], Hiroyuki Ogata[12], Patrick Wincker[5,6], Alexander Culley[13], Samuel Chaffron[6,7], and Matthew B. Sullivan[1,2,3,4,14*]

Affiliations:

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

[2]EMERGE Biology Integration Institute, The Ohio State University; Columbus, Ohio 43210, USA

[3]Center of Microbiome Science, The Ohio State University; Columbus, Ohio 43210, USA

[4]The Interdisciplinary Biophysics Graduate Program, The Ohio State University; Columbus, Ohio 43210, USA

[5]Génomique Métabolique, Genoscope, Institut François-Jacob, CEA, CNRS, Univ Evry, Université Paris-Saclay; 91000 Evry, France

[6]Research Federation for the Study of Global Ocean Systems Ecology and Evolution, FR2022/*Tara* Oceans GOSEE; 75016 Paris, France

[7]Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004; F-44000 Nantes, France

[8]Institut de Biologie de l'Ecole Normale Supérieure, Ecole Normale Supérieure, CNRS, INSERM, Université PSL; 75005 Paris, France

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

[10]Directors' Research European Molecular Biology Laboratory Meyerhofstr; 1 69117 Heidelberg, Germany

[11]Integrated Research Facility at Fort Detrick, National Institute of Allergy and Infectious Diseases, National Institutes of Health; Fort Detrick, Frederick, MD 21702, USA.

[12]Institute for Chemical Research, Kyoto University, Gokasho, Uji; Kyoto 611-0011, Japan

[13]Département de Biochimie, Microbiologie et Bio-informatique, Université Laval; Québec, QC G1V 0A6, Canada

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

[†]These authors contributed equally to this work

[§]Present address: Université Paris-Saclay, CEA, CNRS, Institute for Integrative Biology of the Cell (I2BC); 91198, Gif-sur-Yvette, France

[‡]The Tara Oceans Coordinators are listed in the Supplementary Text

*Corresponding author. Email: sullivan.948@osu.edu

Abstract:

DNA viruses are increasingly recognized to influence marine microbes and microbe-mediated biogeochemical cycling. However, little is known about global marine RNA virus diversity, ecology, and ecosystem roles. Here we uncover patterns and predictors of marine RNA virus community- and "species"-level diversity, and contextualize their ecological impacts from pole to pole. Our analyses revealed four ecological zones, latitudinal and depth diversity patterns, and environmental correlates for RNA viruses. Our findings only partially parallel those of co-sampled plankton and show unexpectedly high polar ecological interactions. The influence of RNA viruses on ecosystems appears to be large, as predicted hosts are ecologically important. Moreover, the occurrence of auxiliary metabolic genes indicates RNA viruses cause reprogramming of diverse host metabolisms, including photosynthesis and carbon cycling, and RNA virus abundances significantly predict ocean carbon export.

One-Sentence Summary: Community- and "species"-level analyses reveal unexpected ecological patterns and roles of RNA viruses in the Global Ocean

Main Text:

The Global Ocean is dominated by plankton communities that are essential to sustain life on Earth. Plankton are at the base of the food web for marine and terrestrial organisms and drive planetary biogeochemical cycles (*1*, *2*). Because nearly half of Earth's primary production derives from ocean plankton, carbon cycling and biodiversity studies have long been a focus in oceanography (*3*). In addition, marine plankton are central to the biological carbon pump as their activity determines whether dissolved carbon dioxide is assimilated into biomass that can be sequestered to the deep ocean or recycled in surface waters and likely released to the atmosphere (*4*, *5*). Thus, understanding ocean biodiversity, carbon export, and related chemical transformations are critical to predicting the changing role of the ocean in the Anthropocene.

Plankton are susceptible to virus infection. Double-stranded DNA (dsDNA) viruses have been increasingly recognized as major ecosystem players (*6*), whereas RNA viruses have been less well-studied owing to methodological challenges (*7*). It is clear, however, that marine RNA viruses are likely important in marine ecosystems, as they (i) are abundant (*8*, *9*), (ii) infect protists and invertebrates that are central to ocean biogeochemical cycling (*10*), and (iii) have been statistically associated with termination of algal blooms (*11*, *12*) and modulation of host diversity (*13*). Despite literature increasingly presenting RNA viruses as a likely major force behind biogeochemistry (*6*, *14*, *15*), empirical data are challenging to obtain. Recent sequencing surveys, including from the oceans, have identified thousands of previously unknown RNA viruses, comprising genus- or subfamily-rank taxa (*16*–*18*) as well as phylum-rank taxa (*19*). However, research on the ecology of RNA viruses has been limited to small spatial scales among pelagic waters and/or viruses associated with larger plankton of a few species (**table S1**). This lack of ecological context, particularly over large scales, limits the incorporation of RNA viruses into predictive models.

Previously, we analyzed 771 metatranscriptomes (provided by *Tara* Oceans Expeditions) that span diverse ocean waters, depths, organismal size fractions, and sequencing library approaches (**Fig. 1A; fig. S1; table S2** for sample metadata; **Materials and Methods**) to identify and quantify RNA viruses (*19*). This effort led to the identification of 44,779 RNA virus contigs that were de-replicated to 5,504 "species"-level virus operational taxonomic units (vOTUs), for which we established taxonomy, evolutionary origins, and biogeography. Here we leverage these data to generate and test several existing hypotheses about RNA virus diversity and their ecological roles throughout the Global Ocean.

RNA virus meta-community analyses reveal distinct ecological zones

Given the importance of marine plankton (*2*), scientists have long sought to understand their ecological patterns and drivers through space and/or time. Temporal studies have revealed seasonal-, depth-, and nutrient-related local or regional drivers of plankton species diversity and community composition, whereas systematic surveys sought to examine these ecological patterns and drivers on a global scale (**table S3**). However, none of these global studies included RNA viruses. Hence, we used our previously generated RNA vOTUs (*19*), pre-clustered at 90% average nucleotide identity across 80% of the shorter sequence length and 1-kb minimum contig length (see **Materials and Methods**), and their relative abundances, estimated by means of metatranscriptomic read mapping (see **Materials and Methods**), to investigate marine RNA virus ecology globally.

Using a statistical method that non-linearly deconvolutes high-dimensional data into two-dimensional space (**Fig. 1B**; t-distributed Stochastic Neighbor Embedding, **fig. S2A–C**) and classical hierarchical clustering techniques (**fig. S2D**) on Bray-Curtis dissimilarity matrices of

RNA vOTU relative abundances (**Materials and Methods**), we show that Global Ocean RNA virus communities can be assigned to four ecological zones: Arctic, Antarctic, Temperate and Tropical Epipelagic, and Temperate and Tropical Mesopelagic. Assortment into only four ecological zones contrasts the 56 biogeochemical provinces classically described for the surface oceans where nutrients and primary productivity drive plankton community composition (*20*). However, the four ecological zone assignments are nearly identical (115 of 118 shared samples) to those inferred for prokaryotic dsDNA viruses (*21*) (see **Materials and Methods**; note that the fifth Bathypelagic zone inferred from dsDNA virus analyses was not sampled here), and largely parallel to those from broader *Tara* Oceans Consortium work on prokaryotes (*22*). Before this study, these ecological zone analyses had not been performed for eukaryotes or eukaryotic RNA viruses. Also previously, transport or migration of eukaryotic plankton across ocean surface biomes and layers was thought to erode the boundaries between these ecological zones (*23*). Our and other recent eukaryotic data (*24*), challenge this hypothesis.

Investigation of ecological parameters that potentially drive community structure at large scale revealed that temperature alone could explain most RNA virus community composition variation along the first ordination axis (**Fig. 1C**). Other ecological drivers, including oxygen, depth, and nutrient availability, may shape plankton community composition (**table S3 A10–14**), but these often co-vary with temperature. Limited sampling in these previous, geographically constrained studies led to the hypothesis that depth is the main driver of plankton community composition. With global data now available, it is apparent that temperature variance potentially drives stratification in non-polar regions (fig. S2E–F) and selects for cold-adapted communities in polar regions. A temperature-driven RNA virus community composition complements that for dsDNA viruses (*21*), prokaryotes (*22*), eukaryotes (*24*), and their interactions (*25*).

Differential predictors of RNA virus global and local "species"-level diversity

Comparison of the diversity patterns of RNA (this study) and dsDNA (*21*) viruses revealed highly concordant large-scale patterns, including previously-identified (*21*) high- and low-diversity regions of the Arctic Ocean (ARC-H and ARC-L; **Fig. 2**). However, local diversity comparisons (i.e., per-sample comparisons) showed that the concordance, despite being significant (p <0.02), was modest (r ≈0.25 per each Pearson's and Spearman's tests), which suggests that small-scale diversity drivers may differ for DNA and RNA viruses. When examining the large suite of environmental variables available for our samples (**table S4**) for possible correlations with RNA and dsDNA virus diversity, we accounted for collinearity using a systems biology network analysis framework to reduce environmental factor dimensionality into fewer environmental "modules" (**Fig. 3**; see **Materials and Methods**).

We found, first, similar to dsDNA viruses (*21*), temperature (cyan module in **Fig. 3**) was not the best predictor of RNA virus diversity. Instead, nutrients (white module in **Fig. 3**) were prominent predictors of species diversity for both dsDNA and RNA viruses, along with other signatures of primary productivity (violet module in **Fig. 3**). Second, in our previous study on dsDNA viruses (*21*), we showed that the link between dsDNA virus diversity and nutrients might be through primary productivity, because photosynthetic coccolithophores' abundance and particulate inorganic carbon [PIC] concentration co-varies with dsDNA virus diversity (light green module in **Fig. 3**). More recently, the relationship of dsDNA viruses and PIC has been posited to be abiotic via direct virus-mediated mineral precipitation (*26*). Unlike dsDNA virus diversity, RNA virus diversity does not correlate with the PIC module, but does still correlate with primary productivity pigment concentrations, such as chlorophyll b (yellow module in **Fig. 3**), indicating, as expected, that dsDNA and RNA viruses infect different hosts. This, and other biological features of RNA viruses, such as their shorter and faster-evolving genomes, higher

burst sizes, lytic lifestyles, and eukaryotic hosts, are hypothesized to drive virus–host interaction and ecosystem impact differences from dsDNA viruses (*27*). Models, based on known RNA virus biological features, also lend support to this idea (*6*, *7*, *27*, *28*). We interpret the small-scale differences in diversity patterns, despite high concordance at the large scale, to also derive from varied biological features across RNA and dsDNA viruses.

Together these findings indicate that the underlying large-scale potential drivers for virus community composition (which encompasses the identity and abundance of vOTUs) and species diversity (which encompasses the vOTUs' richness and distribution evenness) act similarly for the RNA viruses of eukaryotes and the dsDNA viruses of prokaryotes. For virus community composition, perhaps this is not surprising given that likely host community compositions (planktonic prokaryotes and microbial eukaryotes) also appear to be mainly driven by temperature (*22*, *24*, *29*). For virus diversity, the relationship with host diversity can be more complex (see next section). Locally, the varying biological features of RNA viruses are hypothesized (*7*, *28*) to drive virus-host interaction and ecosystem impact differences between largely prokaryotic dsDNA viruses and eukaryotic RNA viruses. For local diversity predictors, our findings are consistent with this hypothesis.

RNA virus "species"-level diversity along ecological gradients

The physico-chemical tolerances, or ecological gradients, of RNA viruses are not understood. Organismal diversity typically decreases with depth (*30*), as does dsDNA virus diversity (*21*), and we found RNA virus diversity also decreases with depth (**Fig. 4A** and **fig. S3**). Latitudinal diversity gradients are characterized by relatively low polar and high equatorial diversity for most terrestrial flora and fauna (*31*, *32*) and oceanic plankton (*33*). However, paradoxically, prokaryotic dsDNA virus diversity tends to increase in the Arctic (*21*, *33*), unlike

their hosts' diversity (*34*, *35*). Thus, to establish baseline paradigms for RNA viruses, we assessed how RNA virus diversity varied with latitude and how it compares with eukaryotic diversity across our Global Ocean dataset. This revealed no obvious latitudinal pattern for RNA virus diversity, regardless of the size fraction (**Fig. 4B**; **fig. S3**; also see **fig. S4–5** for other sensitivity analyses), reminiscent of the deviation seen for dsDNA viruses (*21*). This disconnect of virus and host diversity also has a precedent among non-viruses (see eukaryotic photosynthetic intracellular symbionts and their eukaryotic hosts (*33*)). We hypothesize that this disconnect is caused by the differential impacts of temperature, allowing (i) viral particles to be better preserved in cold temperatures and/or (ii) more viruses of distinct species to interact with the same host organism in polar waters. The former hypothesis has some support in literature (*36*), whereas the latter is untested.

To test the latter hypothesis, we built an abundance-based co-occurrence network integrating RNA viruses, prokaryotes, and eukaryotes (see **Methods**) to predict hosts for these RNA viruses (*sensu* ref. (*25*)). Assuming that the overall topology of the network is relatively representative, even if specific predictions are not accurate (see the predicted hosts section below), we compared the average number of connections per taxon (i.e., mean degree) in polar and non-polar samples. This comparison showed significantly more connections in polar samples than non-polar samples, and this feature was solely driven by RNA viruses (**Fig. 4C**). This result was unexpected, but is in line with a recent ecological network theory prediction that used data from 511 mammal-infecting viruses to show a non-linear relationship between host and virus diversity (*37*), which was interpreted to be a result of host-sharing among different sets of viruses of separate species.

Hence, although the ecological zones and potential ecological drivers of marine RNA viruses (**Fig. 1B–C**) and their expected eukaryotic hosts (*24*) were similar in our datasets, the

species diversity relationships of RNA viruses and their hosts can be more complex on a global scale.

Marine RNA viruses and inferred local and global ecological impact

First, we sought to place RNA virus diversity data into an ecosystem context by assessing local- to global-scale impacts by means of infected plankton hosts or altered metabolisms (local-scale) versus systems-level ecosystem impact (global-scale). We predicted hosts for our vOTUs using three approaches: (i) host information available for viruses of established taxa, (ii) abundance-based co-occurrence, and/or (iii) endogenous virus element (EVE) signatures (**fig. S6**). Although these results provide only broad taxon rank host predictions, since *in silico* host inferences for RNA viruses are not well-established, they indicated infection of diverse organisms of ecological interest, predominantly protists and fungi, and, to a lesser extent, invertebrate metazoans (**table S5**). We also explored alternative eukaryotic genetic codes for host prediction, which revealed 11 known alternative, eukaryotic genetic codes in 6.8% of the vOTUs and indicated microbial eukaryotes (including mitochondria of yeast, mold, protozoans, and chlorophyceans and nuclear codes of several ciliates) and metazoans (mitochondria of invertebrates) as putative hosts (**table S5**). Notably, these inferred hosts are associated with diverse ecological functions, including phototrophy (e.g., bacillariophytes), phagotrophy (e.g., ciliates), mixotrophy (e.g., dinophyceaens), saprotrophy (e.g., ascomycetes), parasitism (e.g., alveolates), grazing (e.g., arthropods), and filter-feeding (e.g., annelids). Furthermore, several, including certain invertebrate metazoans, and particularly, protists and fungi, are also recognized as critical contributors to the biological carbon pump. Although host prediction is challenging, these findings add support to prior work at smaller scale (**table S1**) that indicate RNA viruses are central ecological players in the oceans. These findings also indicate that, while prokaryotic cells

outnumber eukaryotic organisms in the oceans, few RNA viruses (only 3.4% of the vOTUs) infect bacteria, a result consistent with previous marine virome and virus isolate reports (*7*).

Second, ecosystem impact might be inferred from "cellular" protein sequences we identified in the RNA virus genomes, which we posited may parallel the "auxiliary metabolic genes" (AMGs) that are ecologically important in marine prokaryotic dsDNA viruses (*38*). Although such "cellular" protein sequences are uncommon in RNA virus genomes, either as independent open reading frames (ORFs) or as parts of larger virus proteins, we found 72 functionally distinct AMGs in 95 vOTUs (**table S6**). Together these may hint at how RNA viruses manipulate host physiology to maximize virus production (**Fig. 5**). Although chimeric assemblies might artifactually link AMGs to virus RNA-directed RNA polymerases (RdRP) sequences, several lines of evidence argue against this possibility as follows: (i) 15 AMG–RdRP linkages were observed at multiple sampling sites (**Fig. 5**), and (ii) even though RNA viruses are rarely represented in metatranscriptomes (*16*), long-read sequencing captured three AMG–RdRP linkages (**Data S1**). In addition, no AMG was present in any of the 14 virus contigs putatively derived from EVEs (**Data S2**; see **Materials and Methods**). Mechanistically, we presume such AMGs were acquired by RNA virus genomes through copy-choice recombination with cellular RNAs, as originally suggested for ubiquitin in togaviruses (*39*). We identified 12 previously reported cases of such RdRP-linked AMGs, but only three studies assessed their functional context in virus infection (**table S6**). Thus, we used this larger dataset to explore the possible biology such AMGs might offer to RNA viruses and ecosystems.

Functionally, the 72 AMG types were diverse, with only four cases overlapping with the 12 previously reported AMGs in RNA virus genomes (**table S6**; **Data S1**). The most common functional type of AMGs (15.8%) was involved in RNA modifications (RtcB, AlkB, and RNA 2'-phosphotransferase) and post-translational modifications (NADAR and OARD1), which may

reflect the common need of viruses to evade host antiviral responses through repair of virus RNAs and proteins (*40*, *41*). Given that viruses must reprogram cells towards virus progeny production and that RNA viruses have relatively short genomes, it was not surprising to see that protein kinases were abundant (14.8%), since they would allow broad reprogramming capability through limited genetic capacity. The frequency of AMGs suggested that a suite of other processes is impacted by marine RNA viruses, including carbohydrate metabolism (10.9%), translation (8.9%), nutrient transport (7.9%), photosynthesis (5.9%), and vacuolar digestion (4.0%). We posit that many of these AMGs represent ocean-specific RNA virus adaptations that help cellular "virus factories" maximize output in the often ultra-limiting nutrient conditions of seawater.

Finally, recent experimental work has emerged to assess how DNA viruses impact ocean carbon export over small scales (*42*, *43*). We sought to complement these efforts through Global Ocean assessment of RNA viruses by using previously developed machine learning and ecosystem modeling approaches (*10*) (see **Materials and Methods**) to evaluate *in silico* whether RNA viruses might impact ocean carbon export. This revealed that RNA virus abundances were strongly predictive of ocean carbon flux and identified specific vOTUs that were most significant for these predictions (**fig. S7; table S7**). Specifically, from 5,504 vOTUs, 1,243 were identified as part of four highly significant subnetworks (p-values ≈0) of RNA viruses that strongly predicted carbon flux variation (**fig. S7A**). Notably, subnetwork-specific topology interrogation by partial least squares regression modelling and leave-one-out cross-validation techniques (see **Materials and Methods**) showed that these subnetworks represent predictive community biomarkers for carbon export (cross-validated r^2 up to 0.79, and, critically, in a 1:1 ratio, which implies capturing the correct magnitude in the models; **fig. S7A**). Further, these techniques very conservatively identified 11 RNA viruses that were most predictive of carbon flux (i.e., VIP

score; **table S7**; **fig. S7B**) and offer ideal targets for follow-on hypothesis testing. Chlorophytes and haptophytes could be assigned as hosts for two of these viruses (**fig. S7B**). These algal hosts are thought to be critical components in the biological carbon pump (**table S3 A17–19**).

Conclusions

For decades, extensive studies have focused on plankton dynamics and activity to infer the pairwise links among plankton and carbon export, including recent experimental work with viruses (*42*, *43*). Because these seminal studies were focused on narrow geographic ranges or oceanic provinces, we sought here to instead explore Global Ocean signals by taking advantage of the uniform *Tara* Oceans strategy for sampling plankton and sinking particles to broadly investigate oceanic conditions and ecosystem biota (*10*). Hence, although limited by single-time-points or "snapshot" sampling, combining these measurements with a robust statistical framework (i.e., network-based, cross-validated, multivariate-aware correlation analysis) enables statistical exploration to establish hypotheses about key ecosystem players. For this, we can leverage the context of hypothesized interactions (*25*) instead of using the more traditional pairwise correlations (for example, of a member of specific taxon and an ecosystem output) from classical studies.

Notably, previous *Tara* studies have revealed prokaryotic and eukaryotic DNA virus abundances to provide biological proxies for estimating carbon export (*10*, *44*), and one even identified eukaryotic virus abundances as predictive for carbon export efficiency (*44*). However, the RNA virus diversity and abundance analyses presented here represent major advances as follows: (i) our ecological unit and abundance calculation methods (from contigs to high-quality genomes) were extensively evaluated and found to be robust and suitable for sensitive ecological analyses (**fig. S4–5**; a novel evaluation in RNA virus ecology); (ii) our analyses were composed purely of

RNA viruses, due to capturing 25-fold more data that are not dominated by eukaryotic dsDNA viruses; and (iii) our analyses included polar waters, which are critical for carbon export (**fig. S8**). Together these findings provide a roadmap for studying RNA viruses in nature, as well as evidence that RNA viruses play important roles in the ocean ecosystem.

References and Notes

1. C. Costello, L. Cao, S. Gelcich, M. Á. Cisneros-Mata, C. M. Free, H. E. Froehlich, C. D. Golden, G. Ishimura, J. Maier, I. Macadam-Somer, T. Mangin, M. C. Melnychuk, M. Miyahara, C. L. de Moor, R. Naylor, L. Nøstbakken, E. Ojea, E. O'Reilly, A. M. Parma, A. J. Plantinga, S. H. Thilsted, J. Lubchenco, The future of food from the sea. *Nature*. **588**, 95–100 (2020). https://doi.org/10.1038/s41586-020-2616-y
2. P. G. Falkowski, T. Fenchel, E. F. Delong, The microbial engines that drive Earth's biogeochemical cycles. *Science*. **320**, 1034–1039 (2008). https://doi.org/10.1126/science.1153213
3. C. B. Field, M. J. Behrenfeld, J. T. Randerson, P. Falkowski, Primary production of the biosphere: Integrating terrestrial and oceanic components. *Science*. **281**, 237–240 (1998).
4. P. W. Boyd, H. Claustre, M. Levy, D. A. Siegel, T. Weber, Multi-faceted particle pumps drive carbon sequestration in the ocean. *Nature*. **568**, 327–335 (2019). https://doi.org/10.1038/s41586-019-1098-2
5. A. Z. Worden, M. J. Follows, S. J. Giovannoni, S. Wilken, A. E. Zimmerman, P. J. Keeling, Rethinking the marine carbon cycle: Factoring in the multifarious lifestyles of microbes. *Science*. **347**, 1257594 (2015). https://doi.org/10.1126/science.1257594
6. C. A. Suttle, Marine viruses – Major players in the global ecosystem. *Nat. Rev. Microbiol.*

5, 801–812 (2007). https://doi.org/10.1038/nrmicro1750

7. A. Culley, New insight into the RNA aquatic virosphere via viromics. *Virus Res.* **244**, 84–89 (2018). https://doi.org/10.1016/j.virusres.2017.11.008

8. G. F. Steward, A. I. Culley, J. A. Mueller, E. M. Wood-Charlson, M. Belcaid, G. Poisson, Are we missing half of the viruses in the ocean? *ISME J.* **7**, 672–679 (2013). https://doi.org/10.1038/ismej.2012.121

9. J. A. Miranda, A. I. Culley, C. R. Schvarcz, G. F. Steward, RNA viruses as major contributors to Antarctic virioplankton. *Environ. Microbiol.* **18**, 3714–3727 (2016). https://doi.org/10.1111/1462-2920.13291

10. L. Guidi, S. Chaffron, L. Bittner, D. Eveillard, A. Larhlimi, S. Roux, Y. Darzi, S. Audic, L. Berline, J. R. Brum, L. P. Coelho, J. C. I. Espinoza, S. Malviya, S. Sunagawa, C. Dimier, S. Kandels-Lewis, M. Picheral, J. Poulain, S. Searson, L. Stemmann, F. Not, P. Hingamp, S. Speich, M. Follows, L. Karp-Boss, E. Boss, H. Ogata, S. Pesant, J. Weissenbach, P. Wincker, S. G. Acinas, P. Bork, C. De Vargas, D. Iudicone, M. B. Sullivan, J. Raes, E. Karsenti, C. Bowler, G. Gorsky, Plankton networks driving carbon export in the oligotrophic ocean. *Nature*. **532**, 465–470 (2016). https://doi.org/10.1038/nature16942

11. M. Moniruzzaman, L. L. Wurch, H. Alexander, S. T. Dyhrman, C. J. Gobler, S. W. Wilhelm, Virus-host relationships of marine single-celled eukaryotes resolved from metatranscriptomics. *Nat. Commun.* **8**, 1–10 (2017). https://doi.org/10.1038/ncomms16054

12. Y. Tomaru, N. Hata, T. Masuda, M. Tsuji, K. Igata, Y. Masuda, T. Yamatogi, M. Sakaguchi, K. Nagasaki, Ecological dynamics of the bivalve-killing dinoflagellate Heterocapsa circularisquama and its infectious viruses in different locations of western Japan. *Environ. Microbiol.* **9**, 1376–1383 (2007). https://doi.org/10.1111/j.1462-2920.2007.01252.x

13. L. Zeigler Allen, J. P. McCrow, K. Ininbergs, C. L. Dupont, J. H. Badger, J. M. Hoffman, M. Ekman, A. E. Allen, B. Bergman, J. C. Venter, The Baltic Sea Virome: Diversity and Transcriptional Activity of DNA and RNA Viruses. *mSystems*. **2**, e00125–16 (2017). https://doi.org/10.1128/mSystems.00125-16

14. J. A. Gustavsen, D. M. Winget, X. Tian, C. A. Suttle, High temporal and spatial diversity in marine RNA viruses implies that they have an important role in mortality and structuring plankton communities. *Front. Microbiol.* **5**, 703 (2014). https://doi.org/10.3389/fmicb.2014.00703

15. E. P. Starr, E. E. Nuccio, J. Pett-Ridge, J. F. Banfield, M. K. Firestone, Metatranscriptomic reconstruction reveals RNA viruses with the potential to shape carbon cycling in soil. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 25900–25908 (2019). https://doi.org/10.1073/pnas.1908291116

16. M. Shi, X.-D. Lin, J.-H. Tian, L.-J. Chen, X. Chen, C.-X. Li, X.-C. Qin, J. Li, J.-P. Cao, J.-S. Eden, J. Buchmann, W. Wang, J. Xu, E. C. Holmes, Y.-Z. Zhang, Redefining the invertebrate RNA virosphere. *Nature*. **540**, 539–543 (2016). https://doi.org/10.1038/nature20167

17. Y. I. Wolf, D. Kazlauskas, J. Iranzo, A. Lucía-Sanz, J. H. Kuhn, M. Krupovic, V. V. Dolja, E. V. Koonin, Origins and Evolution of the Global RNA Virome. *Mbio*. **9**, e02329-18 (2018). https://doi.org/10.1128/mbio.02329-18

18. C.-X. Li, M. Shi, J.-H. Tian, X.-D. Lin, Y.-J. Kang, L.-J. Chen, X.-C. Qin, J. Xu, E. C. Holmes, Y.-Z. Zhang, Unprecedented genomic diversity of RNA viruses in arthropods reveals the ancestry of negative-sense RNA viruses. *Elife*. **4**, e05378 (2015). https://doi.org/10.7554/elife.05378

19. A.A. Zayed, J.M. Wainaina, G. Dominguez-Huerta, E. Pelletier, J. Guo, M. Mohssen, F.

Tian, A.A. Pratama, B. Bolduc, O. Zablocki, D. Cronin, L. Solden, E. Delage, A. Alberti, J.-M. Aury, Q. Carradec, C. da Silva, K. Labadie, J. Poulain, H.-J. Ruscheweyh, G. Salazar, E. Shatoff, Tara Oceans Coordinators, R. Bundschuh, K. Frederick, L.S. Kubatko, S. Chaffron, A.I. Culley, S. Sunagawa, J. Kuhn, P. Wincker, M.B. Sullivan, Cryptic and abundant marine viruses at the evolutionary origins of Earth's RNA virome. *Science* abm5847 *Accepted*. https://www.science.org/doi/10.1126/science.abm5847

20. A. Longhurst, S. Sathyendranath, T. Platt, C. Caverhill, An estimate of global primary production in the ocean from satellite radiometer data. *Journal of Plankton Research*. **17**,1245–1271 (1995). https://doi.org/10.1093/plankt/17.6.1245

21. A. C. Gregory, A. A. Zayed, N. Conceição-Neto, B. Temperton, B. Bolduc, A. Alberti, M. Ardyna, K. Arkhipova, M. Carmichael, C. Cruaud, C. Dimier, G. Domínguez-Huerta, J. Ferland, S. Kandels, Y. Liu, C. Marec, S. Pesant, M. Picheral, S. Pisarev, J. Poulain, J. É. Tremblay, D. Vik, S. G. Acinas, M. Babin, P. Bork, E. Boss, C. Bowler, G. Cochrane, C. de Vargas, M. Follows, G. Gorsky, N. Grimsley, L. Guidi, P. Hingamp, D. Iudicone, O. Jaillon, S. Kandels-Lewis, L. Karp-Boss, E. Karsenti, F. Not, H. Ogata, N. Poulton, J. Raes, C. Sardet, S. Speich, L. Stemmann, M. B. Sullivan, S. Sunagawa, P. Wincker, A. I. Culley, B. E. Dutilh, S. Roux, Marine DNA viral macro- and microdiversity from pole to pole. *Cell*. **177**, 1109–1123.e14 (2019). https://doi.org/10.1016/j.cell.2019.03.040

22. G. Salazar, L. Paoli, A. Alberti, J. Huerta-cepas, M. Cuenca, C. M. Field, L. P. Coelho, C. Cruaud, S. Engelen, A. C. Gregory, K. Labadie, C. Marec, E. Pelletier, Gene expression changes and community turnover differentially shape the global ocean metatranscriptome. *Cell*. **179**, 1068–1083.e21 (2019). https://dx.doi.org/10.1016%2Fj.cell.2019.10.014

23. K. Bandara, Ø. Varpe, L. Wijewardene, V. Tverberg, K. Eiane, Two hundred years of zooplankton vertical migration research. *Biol. Rev. Camb. Philos. Soc.* **96**, 1547–1589

(2021). https://doi.org/10.1111/brv.12715

24. G. Sommeria-Klein, R. Watteaux, D. Iudicone, C. Bowler, H. Morlon, Global drivers of eukaryotic plankton biogeography in the sunlit ocean. *Science*. **374**, 594–599 (2020). https://doi.org/10.1126/science.abb3717
25. S. Chaffron, E. Delage, M. Budinich, D. Vintache, N. Henry, C. Nef, M. Ardyna, A. A. Zayed, P. C. Junger, P. E. Galand, C. Lovejoy, A. E. Murray, H. Sarmento, Tara Oceans coordinators, S. G. Acinas, M. Babin, D. Iudicone, O. Jaillon, E. Karsenti, P. Wincker, L. Karp-Boss, M. B. Sullivan, C. Bowler, C. de Vargas, D. Eveillard, Environmental vulnerability of the global ocean epipelagic plankton community interactome. *Sci Adv*. **7**, eabg1921 (2021). https://doi.org/10.1126/sciadv.abg1921
26. M. Słowakiewicz, A. Borkowski, M. D. Syczewski, I. D. Perrota, F. Owczarek, A. Sikora, A. Detman, E. Perri, M. E. Tucker, Newly-discovered interactions between bacteriophages and the process of calcium carbonate precipitation. *Geochim. Cosmochim. Acta*. **292**, 482–498 (2021). https://doi.org/10.1016/j.gca.2020.10.012
27. M. Sadeghi, Y. Tomaru, T. Ahola, RNA Viruses in Aquatic Unicellular Eukaryotes. *Viruses*. **13**, 362 (2021). https://doi.org/10.3390/v13030362
28. K. F. Edwards, G. F. Steward, C. R. Schvarcz, Making sense of virus size and the tradeoffs shaping viral fitness. *Ecol. Lett.* **24**, 363-373 (2021). https://doi.org/10.1111/ele.13630
29. S. Sunagawa, L. P. Coelho, S. Chaffron, J. R. Kultima, K. Labadie, G. Salazar, B. Djahanschiri, G. Zeller, D. R. Mende, A. Alberti, F. M. Cornejo-Castillo, P. I. Costea, C. Cruaud, F. d'Ovidio, S. Engelen, I. Ferrera, J. M. Gasol, L. Guidi, F. Hildebrand, F. Kokoszka, C. Lepoivre, G. Lima-Mendez, J. Poulain, B. T. Poulos, M. Royo-Llonch, H. Sarmento, S. Vieira-Silva, C. Dimier, M. Picheral, S. Searson, S. Kandels-Lewis, C. Bowler, C. de Vargas, G. Gorsky, N. Grimsley, P. Hingamp, D. Iudicone, O. Jaillon, F.

Not, H. Ogata, S. Pesant, S. Speich, L. Stemmann, M. B. Sullivan, J. Weissenbach, P. Wincker, E. Karsenti, J. Raes, S. G. Acinas, P. Bork, Structure and function of the global ocean microbiome. *Science*. **348**, 1261359 (2015). https://doi.org/10.1126/science.1261359

30. M. J. Costello, C. Chaudhary, Marine biodiversity, biogeography, deep-sea gradients, and conservation. *Curr. Biol.* **27**, R511–R527 (2017). https://doi.org/10.1016/j.cub.2017.04.060
31. D. Righetti, M. Vogt, N. Gruber, A. Psomas, N. E. Zimmermann, Global pattern of phytoplankton diversity driven by temperature and environmental variability. *Sci Adv*. **5**, eaau6253 (2019). https://doi.org/10.1126/sciadv.aau6253
32. M. R. Willig, D. M. Kaufman, R. D. Stevens, Latitudinal gradients of biodiversity: pattern, process, scale, and synthesis. *Annual Review of Ecology, Evolution, and Systematics*. **34**, 273–309 (2003). https://doi.org/10.1146/annurev.ecolsys.34.012103.144032
33. F. M. Ibarbalz, N. Henry, M. C. Brandão, S. Martini, G. Busseni, H. Byrne, L. P. Coelho, H. Endo, J. M. Gasol, A. C. Gregory, F. Mahé, J. Rigonato, M. Royo-Llonch, G. Salazar, I. Sanz-Sáez, E. Scalco, D. Soviadan, A. A. Zayed, A. Zingone, K. Labadie, J. Ferland, C. Marec, S. Kandels, M. Picheral, C. Dimier, J. Poulain, S. Pisarev, M. Carmichael, S. Pesant, S. G. Acinas, M. Babin, P. Bork, E. Boss, C. Bowler, G. Cochrane, C. de Vargas, M. Follows, G. Gorsky, N. Grimsley, L. Guidi, P. Hingamp, D. Iudicone, O. Jaillon, L. Karp-Boss, E. Karsenti, F. Not, H. Ogata, N. Poulton, J. Raes, C. Sardet, S. Speich, L. Stemmann, M. B. Sullivan, S. Sunagawa, P. Wincker, E. Pelletier, L. Bopp, F. Lombard, L. Zinger, Global trends in marine plankton diversity across kingdoms of life. *Cell*. **179**, 1084–1097.e21 (2019). https://doi.org/10.1016/j.cell.2019.10.008
34. J. B. Emerson, B. C. Thomas, K. Andrade, K. B. Heidelberg, J. F. Banfield, New approaches indicate constant viral diversity despite shifts in assemblage structure in an Australian hypersaline lake. *Applied and Environmental Microbiology*. **79**, 6755–6764

(2013). https://doi.org/10.1128/AEM.01946-13

35. A. C. Gregory, O. Zablocki, A. A. Zayed, A. Howell, B. Bolduc, M. B. Sullivan, The Gut Virome Database reveals age-dependent patterns of virome diversity in the human gut. *Cell Host Microbe*. **28**, 724–740.e8 (2020). https://doi.org/10.1016/j.chom.2020.08.003

36. E. A. Gould, Methods for long-term virus preservation. *Molecular Biotechnology*. **13**, 57–66 (1999). https://doi.org/10.1385/mb:13:1:57

37. C. J. Carlson, C. M. Zipfel, R. Garnier, S. Bansal, Global estimates of mammalian viral diversity accounting for host sharing. *Nat Ecol Evol*. **3**, 1070–1075 (2019). https://doi.org/10.1038/s41559-019-0910-6

38. M. Breitbart, L. R. Thompson, C. A. Suttle, M. B. Sullivan, Exploring the vast diversity of marine viruses. *Oceanography*. **20**, 135–139 (2007). https://doi.org/10.5670/oceanog.2007.58

39. G. Meyers, T. Rümenapf, H.-J. Thiel, Ubiquitin in a togavirus. *Nature*. **341**, 491–491 (1989). https://doi.org/10.1038/341491a0

40. A. M. Burroughs, L. Aravind, RNA damage in biological conflicts and the diversity of responding RNA repair systems. *Nucleic Acids Res.* **44**, 8525–8555 (2016). https://doi.org/10.1093/nar/gkw722

41. A. R. Fehr, G. Jankevicius, I. Ahel, S. Perlman, Viral macrodomains: unique mediators of viral replication and pathogenesis. *Trends Microbiol.* **26**, 598–610 (2018). https://doi.org/10.1016/j.tim.2017.11.011

42. F. Vincent, U. Sheyn, Z. Porat, D. Schatz, A. Vardi, Visualizing active viral infection reveals diverse cell fates in synchronized algal bloom demise. *Proc. Natl. Acad. Sci. U. S. A.* **118**, e2021586118 (2021). https://doi.org/10.1073/pnas.2021586118

43. T. E. G. Biggs, J. Huisman, C. P. D. Brussaard, Viral lysis modifies seasonal

phytoplankton dynamics and carbon flow in the Southern Ocean. *ISME J.* **15**, 3615–3622 (2021). https://doi.org/10.1038/s41396-021-01033-6

44. H. Kaneko, R. Blanc-Mathieu, H. Endo, S. Chaffron, T. O. Delmont, M. Gaia, N. Henry, R. Hernández-Velázquez, C. H. Nguyen, H. Mamitsuka, P. Forterre, O. Jaillon, C. de Vargas, M. B. Sullivan, C. A. Suttle, L. Guidi, H. Ogata, Eukaryotic virus composition can predict the efficiency of carbon export in the global ocean. *iScience*. **24**, 102002 (2021). https://doi.org/10.1016/j.isci.2020.102002

45. A. A. Zayed, J. M. Wainaina, G. Dominguez-Huerta, Cryptic and abundant marine viruses at the evolutionary origins of Earth's RNA virome. CyVerse Data Commons (2021). doi.org/10.25739/qhr0-8j16

46. A. Alberti, J. Poulain, S. Engelen, K. Labadie, S. Romac, I. Ferrera, G. Albini, J.-M. Aury, C. Belser, A. Bertrand, C. Cruaud, C. Da Silva, C. Dossat, F. Gavory, S. Gas, J. Guy, M. Haquelle, E. Jacoby, O. Jaillon, A. Lemainque, E. Pelletier, G. Samson, M. Wessner, S. G. Acinas, M. Royo-Llonch, F. M. Cornejo-Castillo, R. Logares, B. Fernández-Gómez, C. Bowler, G. Cochrane, C. Amid, P. T. Hoopen, C. De Vargas, N. Grimsley, E. Desgranges, S. Kandels-Lewis, H. Ogata, N. Poulton, M. E. Sieracki, R. Stepanauskas, M. B. Sullivan, J. R. Brum, M. B. Duhaime, B. T. Poulos, B. L. Hurwitz, S. Pesant, E. Karsenti, P. Wincker, Viral to metazoan marine plankton nucleotide sequences from the Tara Oceans expedition. *Sci Data*. **4**, 170093 (2017). https://doi.org/10.1038/sdata.2017.93

47. L. Bobay, H. Ochman, Biological species in the viral world. *Proc. Natl. Acad. Sci. U. S. A*. **115**, 1–6 (2018). https://doi.org/10.1073/pnas.1717593115

48. Y. Boucher, C. J. Douady, R. T. Papke, D. A. Walsh, M. E. R. Boudreau, C. L. Nesbø, R. J. Case, W. F. Doolittle, Lateral gene transfer and the origins of prokaryotic groups. *Annu. Rev. Genet.* **37**, 283–328 (2003). https://doi.org/10.1146/annurev.genet.37.050503.084247

49. H. Cadillo-Quiroz, X. Didelot, N. L. Held, A. Herrera, A. Darling, M. L. Reno, D. J. Krause, R. J. Whitaker, Patterns of gene flow define species of thermophilic Archaea. *PLoS Biology*. **10**, e1001265 (2012). https://doi.org/10.1371/journal.pbio.1001265

50. Q. Carradec, E. Pelletier, C. Da Silva, A. Alberti, Y. Seeleuthner, R. Blanc-Mathieu, G. Lima-Mendez, F. Rocha, L. Tirichine, K. Labadie, A. Kirilovsky, A. Bertrand, S. Engelen, M. A. Madoui, R. Meheust, J. Poulain, S. Romac, D. J. Richter, G. Yoshikawa, C. Dimier, S. Kandels-Lewis, M. Picheral, S. Searson, C. Tara Oceans, O. Jaillon, J. M. Aury, E. Karsenti, M. B. Sullivan, S. Sunagawa, P. Bork, F. Not, P. Hingamp, J. Raes, L. Guidi, H. Ogata, C. de Vargas, D. Iudicone, C. Bowler, P. Wincker, A global ocean atlas of eukaryotic genes. *Nat Commun*. **9**, 373 (2018). https://doi.org/10.1038/s41467-017-02342-1

51. L. Deng, J. C. Ignacio-Espinoza, A. C. Gregory, B. T. Poulos, J. S. Weitz, P. Hugenholtz, M. B. Sullivan, Viral tagging reveals discrete populations in *Synechococcus* viral genome sequence space. *Nature*. **513**, 242–245 (2014). https://doi.org/10.1038/nature13459

52. P. Dixon, VEGAN, a package of R functions for community ecology. *Journal of Vegetation Science.* **14**, 927–930 (2003). https://doi.org/10.1111/j.1654-1103.2003.tb02228.x

53. S. Duffy, Why are RNA virus mutation rates so damn high? *PLOS Biology*. **16**, e3000003 (2018). https://doi.org/10.1371/journal.pbio.3000003

54. A. H. Fitzpatrick, A. Rupnik, H. O'Shea, F. Crispie, S. Keaveney, P. Cotter, High throughput sequencing for the detection and characterization of RNA viruses. *Frontiers in Microbiology*. **12**, 190 (2021). https://doi.org/10.3389/fmicb.2021.621719

55. A. C. Gregory, S. A. Solonenko, J. C. Ignacio-Espinoza, K. LaButti, A. Copeland, S. Sudek, A. Maitland, L. Chittick, F. Dos Santos, J. S. Weitz, A. Z. Worden, T. Woyke, M.

B. Sullivan, Genomic differentiation among wild cyanophages despite widespread horizontal gene transfer. *BMC Genomics*. **17**, 930 (2016). https://doi.org/10.1186/s12864-016-3286-x

56. L. Guidi, G. A. Jackson, L. Stemmann, J. C. Miquel, M. Picheral, G. Gorsky, Relationship between particle size distribution and flux in the mesopelagic zone. *Deep Sea Research Part I: Oceanographic Research Papers*. **55**, 1364–1374 (2008). https://doi.org/10.1016/j.dsr.2008.05.014

57. R. W. Hendrix, M. C. Smith, R. N. Burns, M. E. Ford, G. F. Hatfull, Evolutionary relationships among diverse bacteriophages and prophages: all the world's a phage. *Proceedings of the National Academy of Sciences of the United States of America*. **96**, 2192–2197 (1999). https://doi.org/10.1073/pnas.96.5.2192

58. B. J. Shapiro, J. Friedman, O. X. Cordero, S. P. Preheim, S. C. Timberlake, G. Szabo, M. F. Polz, E. J. Alm, Population genomics of early events in the ecological differentiation of bacteria. *Science*. **336**, 48–51 (2012). https://doi.org/10.1126/science.1218198

59. S. R. Krishnamurthy, D. Wang, Extensive conservation of prokaryotic ribosomal binding sites in known and novel picobirnaviruses. *Virology*. **516**, 108–114 (2018). https://doi.org/10.1016/j.virol.2018.01.006

60. C.-X. Li, W. Li, J. Zhou, B. Zhang, Y. Feng, C.-P. Xu, Y.-Y. Lu, E. C. Holmes, M. Shi, High resolution metagenomic characterization of complex infectomes in paediatric acute respiratory infection. *Sci Rep*. **10**, 3963 (2020). https://doi.org/10.1038/s41598-020-60992-6

61. T. N. Mavrich, G. F. Hatfull, Bacteriophage evolution differs by host, lifestyle and genome. *Nature Microbiology*. **2**, 17112 (2017). https://doi.org/10.1038/nmicrobiol.2017.112

62. S. Pesant, F. Not, M. Picheral, S. Kandels-Lewis, N. Le Bescot, G. Gorsky, D. Iudicone, E.

Karsenti, S. Speich, R. Troublé, C. Dimier, S. Searson, Open science resources for the discovery and analysis of Tara Oceans data. *Sci Data*. **2**, 150023 (2015). https://doi.org/10.1038/sdata.2015.23

63. M. Picheral, L. Guidi, L. Stemmann, D. M. Karl, G. Iddaoud, G. Gorsky, The Underwater Vision Profiler 5: An advanced instrument for high spatial resolution studies of particle size spectra and zooplankton. *Limnology and Oceanography: Methods*. **8**, 462–473 (2010). https://doi.org/10.4319/lom.2010.8.462

64. S. Roux, E. M. Adriaenssens, B. E. Dutilh, E. V. Koonin, A. M. Kropinski, M. Krupovic, J. H. Kuhn, R. Lavigne, J. R. Brister, A. Varsani, C. Amid, R. K. Aziz, S. R. Bordenstein, P. Bork, M. Breitbart, G. R. Cochrane, R. D. A., C. Desnues, M. B. Duhaime, J. B. Emerson, F. Enault, J. F. A., P. Hingamp, P. Hugenholtz, B. L. Hurwitz, N. N. Ivanova, Jessica, M. Labonté, K.-B. Lee, R. R. Malmstrom, Manuel Martinez-Garcia, I. Karsch, Mizrachi, H. Ogata, M.-A. P. , David Páez-Espino, C. Putonti, Thomas, Rattei, A. Reyes, F. Rodriguez-Valera, K. Rosario, L. Schriml, Frederik, Schulz, G. F. Steward, M. B. Sullivan, S. Sunagawa, C. A. Suttle, B. Temperton, S. G. Tringe, R. V. Thurber, N. S. Webster, K. L., Whiteson, S. W. Wilhelm, K. E. Wommack, T. Woyke, K. Wrighton, Pelin Yilmaz, T. Yoshida, M. J. Young, N. Yutin, L. Z. Allen, N. C., Kyrpides, E. A. Eloe-Fadrosh, Minimum Information about an Uncultivated Virus Genome (MIUViG): a community consensus on standards and best practices for describing genome sequences from uncultivated viruses. *Nature biotechnology*. **37**, 29–37 (2018). https://doi.org/10.1038/nbt.4306

65. M. Shaffer, M. A. Borton, B. B. McGivern, A. A. Zayed, S. L. 0003 3527 8101 La Rosa, L. M. Solden, P. Liu, A. B. Narrowe, J. Rodríguez-Ramos, B. Bolduc, M. C. Gazitúa, R. A. Daly, G. J. Smith, D. R. Vik, P. B. Pope, M. B. Sullivan, S. Roux, K. C. Wrighton, DRAM

for distilling microbial metabolism to automate the curation of microbiome function. *Nucleic Acids Research*. **48**, 8883–8900 (2020). https://doi.org/10.1093/nar/gkaa621

66. M. Wille, M. Shi, M. Klaassen, A. C. Hurt, E. C. Holmes, Virome heterogeneity and connectivity in waterfowl and shorebird communities. *ISME J*. **13**, 2603–2616 (2019). https://doi.org/10.1038/s41396-019-0458-0
67. Y. I. Wolf, S. Silas, Y. Wang, S. Wu, M. Bocek, D. Kazlauskas, M. Krupovic, A. Fire, V. V. Dolja, E. V. Koonin, Doubling of the known set of RNA viruses by metagenomic analysis of an aquatic virome. *Nature Microbiology*. **5**, 1262–1270 (2020). https://doi.org/10.1038/s41564-020-0755-4
68. A. I. Culley, A. S. Lang, C. A. Suttle, High diversity of unknown picorna-like viruses in the sea. *Nature*. **424**, 1054–1057 (2003). https://doi.org/10.1038/nature01886
69. A. I. Culley, A. S. Lang, C. A. Suttle, Metagenomic analysis of coastal RNA virus communities. *Science*. **312**, 1795–1798 (2006). https://doi.org/10.1126/science.1127404
70. A. I. Culley, A. S. Lang, C. A. Suttle, The complete genomes of three viruses assembled from shotgun libraries of marine RNA virus communities. *Virology Journal*. **4**, 69 (2007). https://doi.org/10.1186/1743-422X-4-69
71. A. I. Culley, G. F. Steward, New genera of RNA viruses in subtropical seawater, inferred from polymerase gene sequences. *Applied and Environmental Microbiology*. **73**, 5937-5944 (2007). https://doi.org/10.1128/AEM.01065-07
72. A. I. Culley, J. A. Mueller, M. Belcaid, E. M. Wood-Charlson, G. Poisson, G. F. Steward, The characterization of RNA viruses in tropical seawater using targeted PCR and metagenomics. *mBio*. **5**, e01210-14 (2014). https://doi.org/10.1128/mBio.01210-14
73. T. Lachnit, T. Thomas, P. Steinberg, Expanding our understanding of the seaweed holobiont: RNA viruses of the red alga *Delisea pulchra*. *Frontiers in Microbiology*. **6**, 1489

(2016). https://doi.org/10.3389/fmicb.2015.01489

74. S. Urayama, Y. Takaki, S. Nishi, Y. Yoshida-Takashima, S. Deguchi, K. Takai, T. Nunoura, Unveiling the RNA virosphere associated with marine microorganisms. *Molecular Ecology Resources*. **18**, 1444–1455 (2018). https://doi.org/10.1111/1755-0998.12936

75. M. Labbé, F. Raymond, A. Lévesque, M. Thaler, V. Mohit, M. Audet, J. Corbeil, A. Culley, Communities of phytoplankton viruses across the transition zone of the St. Lawrence Estuary. *Viruses*. **10**, 672 (2018). https://doi.org/10.3390/v10120672

76. B. C. Kolody, J. P. McCrow, L. Z. Allen, F. O. Aylward, K. M. Fontanez, A. Moustafa, M. Moniruzzaman, F. P. Chavez, C. A. Scholin, E. E. Allen, A. Z. Worden, E. F. Delong, A. E. Allen, Diel transcriptional response of a California Current plankton microbiome to light, low iron, and enduring viral infection. *ISME J*. **13**, 2817–2833 (2019). https://doi.org/10.1038/s41396-019-0472-2

77. M. Vlok, A. S. Lang, C. A. Suttle, Marine RNA virus quasispecies are distributed throughout the oceans. *mSphere.* **4**, e00157-19 (2019). https://doi.org/10.1128/mSphereDirect.00157-19

78. J. A. Gustavsen, C. A. Suttle, Role of phylogenetic structure in the dynamics of coastal viral assemblages. *Applied and Environmental Microbiology.* **87**, e02704-20 (2021). https://doi.org/10.1128/AEM.02704-20

79. V. Tai, J. E. Lawrence, A. S. Lang, A. M. Chan, A. I. Culley, C. A. Suttle, Characterization of HaRNAV, a single-stranded RNA virus causing lysis of *Heterosigma akashiwo* (Raphidophyceae). *Journal of Phycology*. **39**, 343–352 (2003). https://doi.org/10.1046/j.1529-8817.2003.01162.x

80. A. S. Lang, A. I. Culley, C. A. Suttle, Genome sequence and characterization of a virus

(HaRNAV) related to picorna-like viruses that infects the marine toxic bloom-forming alga *Heterosigma akashiwo*. *Virology*. **320**, 206–217 (2004). https://doi.org/10.1016/j.virol.2003.10.015

81. K. Nagasaki, Y. Tomaru, N. Katanozaka, Y. Shirai, K. Nishida, S. Itakura, M. Yamaguchi, Isolation and characterization of a novel single-stranded RNA virus infecting the bloom-forming diatom *Rhizosolenia setigera*. *Applied and Environmental Microbiology.* **70**, 704-711 (2004). https://doi.org/10.1128/AEM.70.2.704-711.2004
82. Y. Tomaru, N. Katanozaka, K. Nishida, Y. Shirai, K. Tarutani, M. Yamaguchi, K. Nagasaki, Isolation and characterization of two distinct types of HcRNAV, a single-stranded RNA virus infecting the bivalve-killing microalga *Heterocapsa circularisquama*. *Aquatic Microbial Ecology*. **34**, 207–218 (2004). https://doi.org/10.3354/ame034207
83. K. Nagasaki, Y. Tomaru, Y. Takao, K. Nishida, Y. Shirai, H. Suzuki, T. Nagumo, Previously unknown virus infects marine diatom. *Applied and Environmental Microbiology.* **71**, 3528-3535 (2005). https://doi.org/10.1128/AEM.71.7.3528-3535.2005
84. Y. Takao, K. Mise, K. Nagasaki, T. Okuno, D. 2006 Honda, Complete nucleotide sequence and genome organization of a single-stranded RNA virus infecting the marine fungoid protist *Schizochytrium sp*. *Journal of General Virology*. **87**, 723–733. https://doi.org/10.1099/vir.0.81204-0
85. Y. Shirai, Y. Tomaru, Y. Takao, H. Suzuki, T. Nagumo, K. Nagasaki, Isolation and characterization of a single-stranded RNA virus infecting the marine planktonic diatom *Chaetoceros tenuissimus* Meunier. *Applied and Environmental Microbiology.* **74**, 4022-4027 (2008). https://doi.org/10.1128/AEM.00509-08
86. Y. Tomaru, Y. Takao, H. Suzuki, T. Nagumo, K. Nagasaki, Isolation and characterization of a single-stranded RNA virus infecting the bloom-forming diatom *Chaetoceros socialis*.

Applied and Environmental Microbiology. **75**, 2375-2381 (2009). https://doi.org/10.1128/AEM.02580-08

87. Y. Tomaru, K. Toyoda, K. Kimura, N. Hata, M. Yoshida, K. Nagasaki, First evidence for the existence of pennate diatom viruses. *ISME J.* **6**, 1445–1448 (2012). https://doi.org/10.1038/ismej.2011.207

88. L. Arsenieff, N. Simon, F. Rigaut-Jalabert, F. Le Gall, S. Chaffron, E. Corre, E. Com, E. Bigeard, A.-C. Baudoux, First viruses infecting the marine diatom *Guinardia delicatula*. *Frontiers in Microbiology*. **9**, 3235 (2019). https//doi.org/10.3389/fmicb.2018.03235

89. K. Toyoda, K. Kimura, K. Osada, D. M. Williams, T. Adachi, K. Yamada, Y. Tomaru, Novel marine diatom ssRNA virus NitRevRNAV infecting *Nitzschia reversa*. *Plant Ecology and Evolution*. **152**, 178–187 (2019). https://doi.org/10.5091/plecevo.2019.1615

90. D. K. Steinberg, M. R. Landry, Zooplankton and the Ocean Carbon Cycle. *Annual Review of Marine Science*. **9**, 413–444 (2017). https://doi.org/10.1146/annurev-marine-010814-015924

91. D. M. Karl, N. R. Bates, S. Emerson, P. J. Harrison, C. Jeandel, O. Llinâs, K.-K. Liu, J.-C. Marty, A. F. Michaels, J. C. Miquel, S. Neuer, Y. Nojiri, C. S. Wong, in *Ocean Biogeochemistry: The Role of the Ocean Carbon Cycle in Global Change*, M. J. R. Fasham, Ed. (Springer, Berlin, Heidelberg) (2003). https://doi.org/10.1007/978-3-642-55844-3_11

92. M. A. Moran, E. B. Kujawinski, A. Stubbins, R. Fatland, L. I. Aluwihare, A. Buchan, B. C. Crump, P. C. Dorrestein, S. T. Dyhrman, N. J. Hess, B. Howe, K. Longnecker, P. M. Medeiros, J. Niggemann, I. Obernosterer, D. J. Repeta, J. R. Waldbauer, Deciphering ocean carbon in a changing world. *PNAS*. **113**, 3143–3151 (2016). https://doi.org/https://doi.org/10.1073/pnas.1514645113

93. J. A. Fuhrman, J. A. Cram, D. M. Needham, Marine microbial community dynamics and their ecological interpretation. *Nat Rev Microbiol*. **13**, 133–146 (2015). https://doi.org/10.1038/nrmicro3417

94. J. A. Cram, C.-E. T. Chow, R. Sachdeva, D. M. Needham, A. E. Parada, J. A. Steele, J. A. Fuhrman, Seasonal and interannual variability of the marine bacterioplankton community throughout the water column over ten years. *ISME J*. **9**, 563–580 (2015). https://doi.org/10.1038/ismej.2014.153

95. J. A. Fuhrman, Microbial community structure and its functional implications. *Nature*. **459**, 193–199 (2009). https://doi.org/10.1038/nature08058

96. M. Ardyna, C. J. Mundy, M. M. Mills, L. Oziel, P.-L. Grondin, L. Lacour, G. Verin, G. van Dijken, J. Ras, E. Alou-Font, M. Babin, M. Gosselin, J.-É. Tremblay, P. Raimbault, P. Assmy, M. Nicolaus, H. Claustre, K. R. Arrigo, Environmental drivers of under-ice phytoplankton bloom dynamics in the Arctic Ocean. *Elementa: Science of the Anthropocene*. **8**, 30 (2020). https://doi.org/10.1525/elementa.430

97. P. Ramond, R. Siano, S. Schmitt, C. de Vargas, L. Marié, L. Memery, M. Sourisseau, Phytoplankton taxonomic and functional diversity patterns across a coastal tidal front. *Sci Rep*. **11**, 2682 (2021). https://doi.org/10.1038/s41598-021-82071-0

98. A. R. Longhurst, in *Ecological Geography of the Sea (Second Edition)*, A. R. Longhurst, Ed. (Academic Press, Burlington (2007). https://doi.org/10.1016/B978-0-12-455521-1.X5000-1

99. C. A. Smoot, R. R. Hopcroft, Depth-stratified community structure of Beaufort Sea slope zooplankton and its relations to water masses. *Journal of Plankton Research*. **39**, 79–91 (2017). https://doi.org/10.1093/plankt/fbw087

100. J. Zorz, C. Willis, A. M. Comeau, M. G. I. Langille, C. L. Johnson, W. K. W. Li, J.

LaRoche, Drivers of regional bacterial community structure and diversity in the Northwest Atlantic Ocean. *Frontiers in Microbiology*. **10**, 281 (2019). https://doi.org/10.3389/fmicb.2019.00281

101. C. Mena, P. Reglero, M. Hidalgo, E. Sintes, R. Santiago, M. Martín, G. Moyà, R. Balbín, Phytoplankton community structure is driven by stratification in the oligotrophic Mediterranean Sea. *Frontiers in Microbiology*. **10**, 1698 (2019). https://doi.org/10.3389/fmicb.2019.01698

102. G. L. Fernandes, B. D. Shenoy, S. R. Damare, Diversity of bacterial community in the oxygen minimum zones of Arabian Sea and Bay of Bengal as deduced by Illumina sequencing. *Frontiers in Microbiology*. **10**, 3153 (2020). https://doi.org/10.3389/fmicb.2019.03153

103. Y. Cui, S.-J. Chun, S. H. Baek, M. Lee, Y. Kim, H.-G. Lee, S.-R. Ko, S. Hwang, C.-Y. Ahn, H.-M. Oh, The water depth-dependent co-occurrence patterns of marine bacteria in shallow and dynamic Southern Coast, Korea. *Sci Rep*. **9**, 9176 (2019). https://doi.org/10.1038/s41598-019-45512-5

104. B. H. Robison, K. R. Reisenbichler, R. E. Sherlock, Giant Larvacean houses: rapid carbon transport to the deep sea floor. *Science*. **308**, 1609–1611 (2005). https://doi.org/10.1126/science.1109104

105. A. Amend, G. Burgaud, M. Cunliffe, V. P. Edgcomb, C. L. Ettinger, M. H. Gutiérrez, J. Heitman, E. F. Y. Hom, G. Ianiri, A. C. Jones, M. Kagami, K. T. Picard, C. A. Quandt, S. Raghukumar, M. Riquelme, J. Stajich, J. Vargas-Muñiz, A. K. Walker, O. Yarden, A. S. Gladfelter, Fungi in the marine environment: open questions and unsolved problems. *mBio.* **10**, e01189-18 (2019). https://doi.org/10.1128/mBio.01189-18.

106. R. Anderson, S. Charvet, P. J. Hansen, Mixotrophy in Chlorophytes and Haptophytes—

effect of irradiance, macronutrient, micronutrient and vitamin limitation. *Frontiers in Microbiology*. **9**, 1704 (2018). https://doi.org/10.3389/fmicb.2018.01704

107. A. Mitra, K. J. Flynn, J. M. Burkholder, T. Berge, A. Calbet, J. A. Raven, E. Granéli, P. M. Glibert, P. J. Hansen, D. K. Stoecker, F. Thingstad, U. Tillmann, S. Våge, S. Wilken, M. V. Zubkov, The role of mixotrophic protists in the biological carbon pump. *Biogeosciences*. **11**, 995–1005 (2014). https://doi.org/10.5194/bg-11-995-2014

108. S. Sengupta, P. C. Gorain, R. Pal, Aspects and prospects of algal carbon capture and sequestration in ecosystems: a review. *Chemistry and Ecology*. **33**, 695–707 (2017). https://doi.org/10.1080/02757540.2017.1359262

109. S. L. C. Giering, R. Sanders, R. S. Lampitt, T. R. Anderson, C. Tamburini, M. Boutrif, M. V. Zubkov, C. M. Marsay, S. A. Henson, K. Saw, K. Cook, D. J. Mayor, Reconciliation of the carbon budget in the ocean's twilight zone. *Nature*. **507**, 480–483 (2014). https://doi.org/10.1038/nature13123

110. P. Tréguer, C. Bowler, B. Moriceau, S. Dutkiewicz, M. Gehlen, O. Aumont, L. Bittner, R. Dugdale, Z. Finkel, D. Iudicone, O. Jahn, L. Guidi, M. Lasbleiz, K. Leblanc, M. Levy, P. Pondaven, Influence of diatom diversity on the ocean biological carbon pump. *Nature Geosci*. **11**, 27–37 (2018). https://doi.org/10.1038/s41561-017-0028-x

111. K. E. Poff, A. O. Leu, J. M. Eppley, D. M. Karl, E. F. DeLong, Microbial dynamics of elevated carbon flux in the open ocean's abyss. *PNAS*. **118**, e2018269118 (2021). https://doi.org/10.1073/pnas.2018269118

112. T. Volk, M. I. Hoffert, in *The Carbon Cycle and Atmospheric CO2: Natural Variations Archean to Present* (American Geophysical Union (AGU) (1985). https://doi.org/10.1029/GM032p0099

113. E. Villar, G. K. Farrant, M. Follows, L. Garczarek, S. Speich, S. Audic, L. Bittner, B.

Blanke, J. R. Brum, C. Brunet, R. Casotti, A. Chase, J. R. Dolan, F. D'Ortenzio, J.-P. Gattuso, N. Grima, L. Guidi, C. N. Hill, O. Jahn, J.-L. Jamet, H. L. Goff, C. Lepoivre, S. Malviya, E. Pelletier, J.-B. Romagnan, S. Roux, S. Santini, E. Scalco, S. M. Schwenck, A. Tanaka, P. Testor, T. Vannier, F. Vincent, A. Zingone, C. Dimier, M. Picheral, S. Searson, S. Kandels-Lewis, T. O. Coordinators, S. G. Acinas, P. Bork, E. Boss, C. de Vargas, G. Gorsky, H. Ogata, S. Pesant, M. B. Sullivan, S. Sunagawa, P. Wincker, E. Karsenti, C. Bowler, F. Not, P. Hingamp, D. Iudicone, Environmental characteristics of Agulhas rings affect interocean plankton transport. *Science*. **348**, 1261447–1 (2015). https://doi.org/10.1126/science.1261447

114. J. R. Brum, J. C. Ignacio-espinoza, S. Roux, G. Doulcier, S. G. Acinas, A. Alberti, S. Chaffron, J. Cesar Ignacio-Espinoza, S. Roux, G. Doulcier, S. G. Acinas, A. Alberti, S. Chaffron, C. Cruaud, C. De Vargas, J. M. Gasol, G. Gorsky, A. C. Gregory, L. Guidi, P. Hingamp, D. Iudicone, F. Not, H. Ogata, S. Pesant, B. T. Poulos, S. M. Schwenck, S. Speich, C. Dimier, S. Kandels-Lewis, M. Picheral, S. Searson, P. Bork, C. Bowler, S. Sunagawa, P. Wincker, E. Karsenti, M. B. Sullivan, E. Boss, M. Follows, N. Grimsley, O. Jaillon, L. Karp-Boss, U. Krzic, J. Raes, E. G. Reynaud, C. Sardet, M. Sieracki, L. Stemmann, D. Velayoudon, J. Weissenbach, Patterns and ecological drivers of ocean viral communities. *Science*. **348**, 1261498 (2015). https://doi.org/10.1126/science.1261498

115. C. de Vargas, S. Audic, N. Henry, J. Decelle, F. Mahe, R. Logares, E. Lara, C. Berney, N. Le Bescot, I. Probert, M. Carmichael, J. Poulain, S. Romac, S. Colin, J.-M. Aury, L. Bittner, S. Chaffron, M. Dunthorn, S. Engelen, O. Flegontova, L. Guidi, A. Horak, O. Jaillon, G. Lima-Mendez, J. Luke, S. Malviya, R. Morard, M. Mulot, E. Scalco, R. Siano, F. Vincent, A. Zingone, C. Dimier, M. Picheral, S. Searson, S. Kandels-Lewis, S. G. Acinas, P. Bork, C. Bowler, G. Gorsky, N. Grimsley, P. Hingamp, D. Iudicone, F. Not, H.

Ogata, S. Pesant, J. Raes, M. E. Sieracki, S. Speich, L. Stemmann, S. Sunagawa, J. Weissenbach, P. Wincker, E. Karsenti, E. Boss, M. Follows, L. Karp-Boss, U. Krzic, E. G. Reynaud, C. Sardet, M. B. Sullivan, D. Velayoudon, Eukaryotic plankton diversity in the sunlit ocean. *Science*. **348**, 1261605–1261605 (2015). https://doi.org/10.1126/science.1261605

116. G. Lima-Mendez, K. Faust, N. Henry, J. Decelle, S. Colin, F. Carcillo, S. Chaffron, J. C. Ignacio-Espinosa, S. Roux, F. Vincent, L. Bittner, Y. Darzi, J. Wang, S. Audic, L. Berline, A. M. Cabello, L. Coppola, F. M. Cornejo-Castillo, F. d'Ovidio, L. DeMeester, I. Ferrera, M.-J. Garet-Delmas, L. Guidi, E. Lara, S. Pesant, M. Royo-Lonch, G. Salazar, P. Sánchez, M. Sebastian, C. Souffreau, C. Dimier, M. Picheral, S. Searson, S. Kandels-Lewis, T. O. Coordinators, G. Gorsky, F. Not, H. Ogata, S. Speich, J. Weissenbach, P. Wincker, G. Bontempi, S. G. Acinas, S. Sunagawa, P. Bork, M. B. Sullivan, C. Bowler, E. Karsenti, C. de Vargas, J. Raes, Determinants of community structure in the global plankton interactome. *Science.* **348**, 1262073 (2015). https://doi.org/10.1126/science.1262073

117. S. Roux, J. R. Brum, B. E. Dutilh, S. Sunagawa, M. B. Duhaime, A. Loy, B. T. Poulos, N. Solonenko, E. Lara, J. Poulain, S. Pesant, S. Kandels-Lewis, C. Dimier, M. Picheral, S. Searson, C. Cruaud, A. Alberti, C. M. Duarte, J. M. Gasol, D. Vaqué, P. Bork, S. G. Acinas, P. Wincker, M. B. Sullivan, D. Vaque, P. Bork, S. G. Acinas, P. Wincker, M. B. Sullivan, T. O. Coordinators, Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature*. **537**, 689–693 (2016). https://doi.org/10.1038/nature19366

118. S. G. Acinas, P. Sánchez, G. Salazar, F. M. Cornejo-Castillo, M. Sebastián, R. Logares, M. Royo-Llonch, L. Paoli, S. Sunagawa, P. Hingamp, H. Ogata, G. Lima-Mendez, S. Roux, J. M. González, J. M. Arrieta, I. S. Alam, A. Kamau, C. Bowler, J. Raes, S. Pesant, P. Bork,

S. Agustí, T. Gojobori, D. Vaqué, M. B. Sullivan, C. Pedrós-Alió, R. Massana, C. M. Duarte, J. M. Gasol, Deep ocean metagenomes provide insight into the metabolic architecture of bathypelagic microbial communities. *Commun Biol*. **4**, 1–15 (2021). https://doi.org/10.1038/s42003-021-02112-2

119. E. van den Born, M. V. Omelchenko, A. Bekkelund, V. Leihne, E. V. Koonin, V. V. Dolja, P. Ø. Falnes, Viral AlkB proteins repair RNA damage by oxidative demethylation. *Nucleic Acids Research*. **36**, 5451–5461 (2008). https://doi.org/10.1093/nar/gkn519

120. M. J. Roossinck, S. Sabanadzovic, R. Okada, R. A. Y. 2011 Valverde, The remarkable evolutionary history of endornaviruses. *Journal of General Virology*. **92**, 2674–2678. https://doi.org/10.1099/vir.0.034702-0

121. D. Linder-Basso, J. N. Dynek, B. I. Hillman, Genome analysis of Cryphonectria hypovirus 4, the most common hypovirus species in North America. *Virology*. **337**, 192–203 (2005). https://doi.org/10.1016/j.virol.2005.03.038

122. D. Khatchikian, M. Orlich, R. Rott, Increased viral pathogenicity after insertion of a 28S ribosomal RNA sequence into the haemagglutinin gene of an influenza virus. *Nature*. **340**, 156–157 (1989). https://doi.org/10.1038/340156a0

123. C. Liu, D. K. Lakshman, S. M. Tavantzis, Expression of a hypovirulence-causing double-stranded RNA is associated with up-regulation of quinic acid pathway and down-regulation of shikimic acid pathway in *Rhizoctonia solani*. *Curr Genet*. **42**, 284–291 (2003). https://doi.org/10.1007/s00294-002-0348-1

124. A. A. Agranovsky, V. P. Boyko, A. V. Karasev, E. V. Koonin, V. V. Dolja, Putative 65 kDa protein of beet yellows closterovirus is a homologue of HSP70 heat shock proteins. *Journal of Molecular Biology*. **217**, 603–610 (1991). https://doi.org/10.1016/0022-2836(91)90517-A

Acknowledgments: We thank Anya Crane (Integrated Research Facility at Fort Detrick, National Institute of Allergy and Infectious Diseases, National Institutes of Health) for critically editing the manuscript. *Tara* Oceans would not exist without the leadership of the *Tara* Expeditions Foundation and the continuous support of 23 institutes (expeditionary support is detailed in the Supplementary Text).

Funding: The virus-specific work presented here was supported in part through the following:

U.S. National Science Foundation (awards OCE#1829831, ABI#1759874, and DBI# 2022070)

The Gordon and Betty Moore Foundation (award #3790)

The Ohio Supercomputer and Ohio State University's Center of Microbiome Science

Ramon-Areces Foundation Postdoctoral Fellowship to GD-H

Laulima Government Solutions, LLC prime contract with the U.S. National Institute of Allergy and Infectious Diseases (NIAID)—Contract No. HHSN272201800013C.

France Génomique (ANR-10-INBS-09)

Author contributions:

G.D.-H., A.A.Z., and M.B.S. planned and supervised the work, interpreted the results, and wrote the manuscript with inputs from all authors. G.D.-H., A.A.Z., J.M.W., J.G., F.T., A.A.P., B.B., M.M., O.Z., E.P., E.D., and S.C. developed and/or implemented the

informatic analyses. A.A., J.-M.A., Q.C., C.d.S., K.L., J.P., A.A.Z., P.W., and Tara Oceans coordinators all contributed to expeditionary infrastructure needed for global ocean sampling, sample processing and/or previously published data resource development. C.B., D.E., E.K. and H.O. provided domain expertise on Global Ocean ecology. L.G., J.H.K., and A.C. provided domain expertise on carbon export, the ecological unit, and RNA virus ecology, respectively. All authors read and commented on the manuscript and approved it in its final form.

Competing interests: Authors declare that they have no competing interests.

Data and materials availability: The authors declare that all data reported herein are fully and freely available from the date of publication without restrictions, and that all of the analyses, publications, and ownership of data are free from legal entanglement or restriction by the various nations whose waters were sampled during the *Tara* Oceans expeditions. This article is contribution number XX of *Tara* Oceans.

Processed data are publicly available through iVirus (*45*), including all metatranscriptome assemblies, RNA virus contigs, and RNA vOTUs. Scripts used to generate figures are uploaded to the MAVERICKlab bitbucket page (https://bitbucket.org/MAVERICLab/global-rna-virus-ecology-2022/).

Supplementary Materials

Materials and Methods

Supplementary Text

Figs. S1 to S8

Tables S1 to S7

References (*46–124*)

Data S1 to S2

5

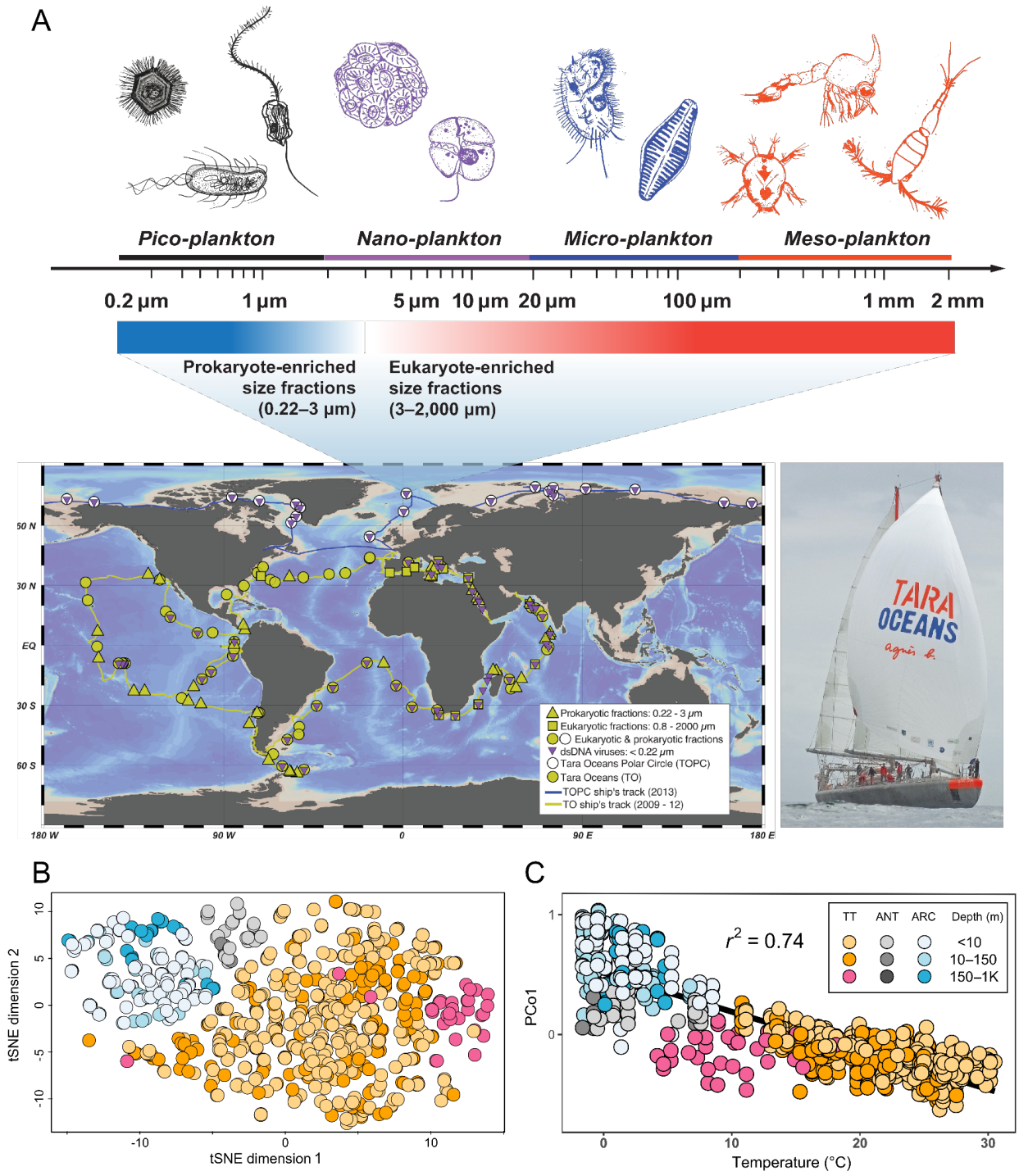

Fig. 1. The cross-domain Global Ocean plankton sampling and resultant RNA virus meta-communities identified from the metatranscriptomes. (**A**) The Global Ocean sampling map shows the cruise of the *Tara* Oceans and *Tara* Oceans Polar Circle expeditions and location of their stations, shown with green and white shapes, respectively. Down-pointing triangles indicate

stations from where dsDNA viromes were previously collected. Up-pointing triangles, squares, and circles show stations with samples of prokaryote-enriched size fractions, eukaryote-enriched size fractions, and both, respectively. The upper blowout panel shows a graded arrow that represents a logarithmic scale of the plankton organismal size fractions captured in this study. The four operational size fractions (piconanoplankton, nanoplankton, microplankton, and mesoplankton) are indicated by the top colored bars and are classified as "prokaryote-enriched" or "eukaryote-enriched" size fractions (highlighted by the bottom gradient-colored bars). Note that such categories, despite being enriched in a type of organism, do not exclude other types. Thus, prokaryote-enriched samples could contain giant viruses and picoeukaryotes, and eukaryotic holobionts of eukaryote-enriched samples could harbor prokaryotes or viruses either as symbionts or food. A picture of the research vessel *Tara* is included as well. (**B**) Statistical analysis (t-Distributed Stochastic Neighbor Embedding [t-SNE]) of a Bray-Curtis dissimilarity matrix calculated from all RNA virus sequence samples in this study regardless of size fraction or library preparation method. Dot colors follow the legend shown in panel C (also see fig. S4–5 for vOTU definition sensitivity analyses). (**C**) Regression analysis of the first coordinate of a principal coordinate analysis of the same Bray–Curtis dissimilarity matrix in panel A (also see fig. S2) and temperature showing that samples across all the size fractions were separated by their local temperatures with an r^2 of 0.74 (p-values =0). ANT, Antarctic; ARC, Arctic; TT, Temperate and Tropical.

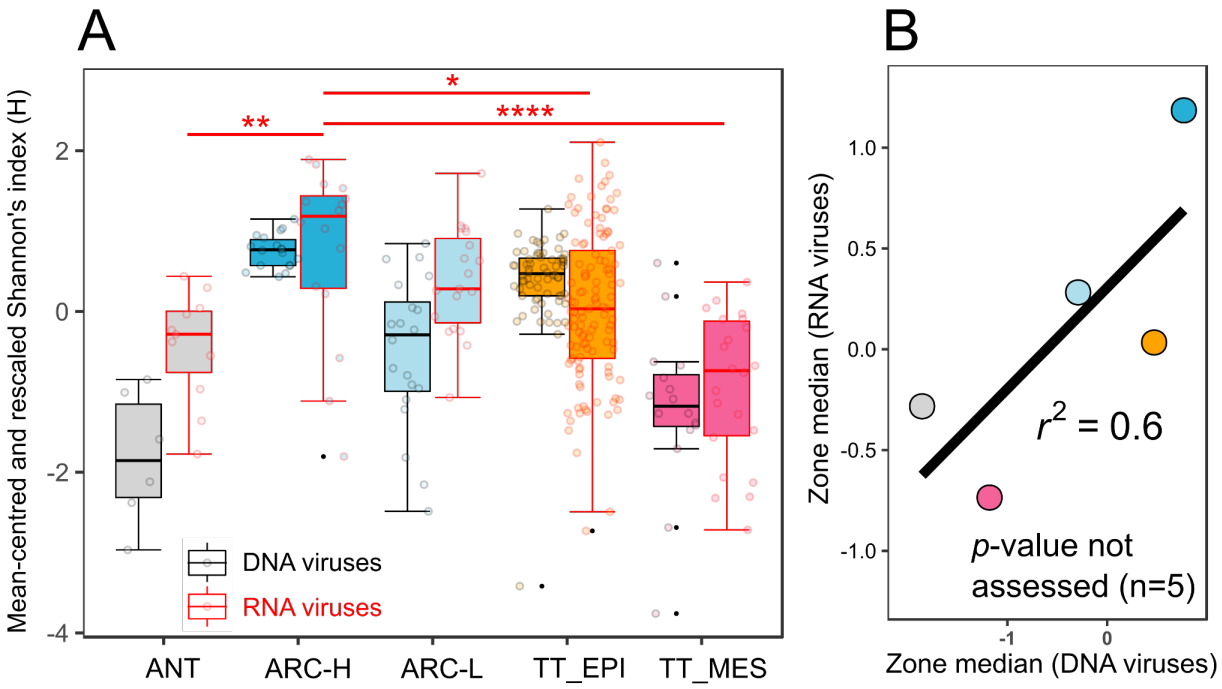

Fig. 2. RNA and DNA virus "species"-level diversity show large-scale congruence. Boxplot (**A**) and regression (**B**) analyses of RNA and DNA virus "species"-level diversity across their shared ecological zones. Shannon's *H* values were mean-centered and rescaled across the two virus nucleic acid types for visual comparisons. All boxplots show medians and quartiles. The medians of each boxplot were used for direct regression analysis. Statistical support (Tukey Honest Significant Differences method on an analysis of variance) is indicated in the figure as follows: ∗ adjusted p<0.05, ∗∗ adjusted p<0.01, and ∗∗∗∗ adjusted p<0.000001. Only RNA viruses from the prokaryotic fraction were used (see **fig. S3** for comparison with the eukaryotic fractions) as this fraction showed the smallest library preparation biases (**fig. S1**; see **Materials and Methods**). ANT, Antarctic; ARC-H, Arctic High-diversity; ARC-L, Arctic Low-diversity; TT_EPI, Temperate and Tropical Epipelagic; TT_MES, Temperate and Tropical Mesopelagic.

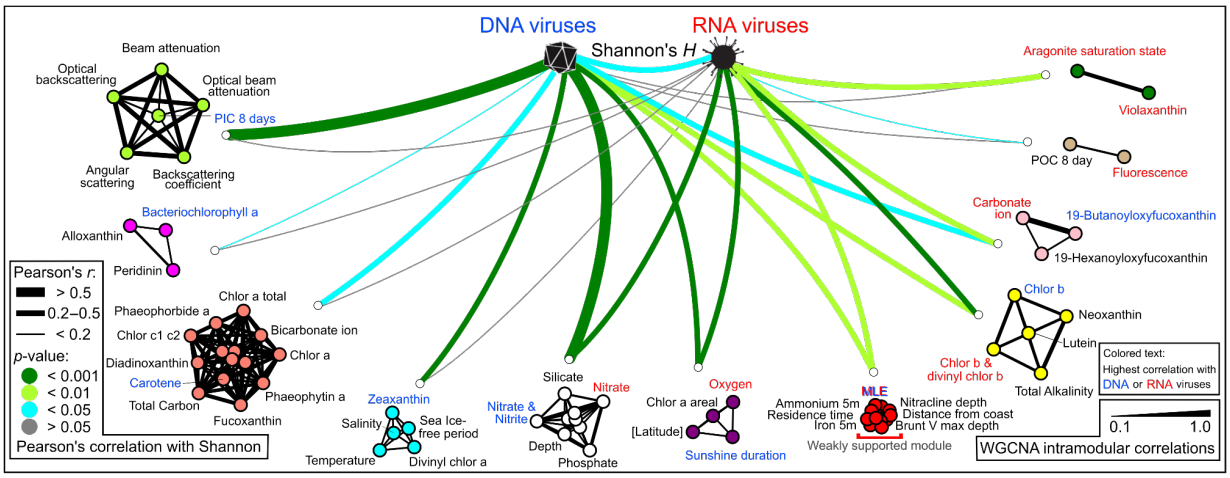

Fig. 3. "Species"-level diversity correlates of marine RNA viruses. Weighted Gene Correlation Network Analysis (WGCNA)-supported modules (to account for collinearity) of environmental variables (see **Methods**) showing the cofactors of RNA and DNA virus diversity. Modules are Pearson-correlated to the Shannon's *H* values of each virus group. Shown are only those relationships that were statistically supported by both Pearson's and Spearman's tests. Only RNA viruses from the prokaryotic fraction were used (see **Fig. 2** for explanation). Notably, aragonite and carbonates could be indicative of coccolithophores, whereas violaxanthin and the latitude-chlor a signal could be related to diatoms.

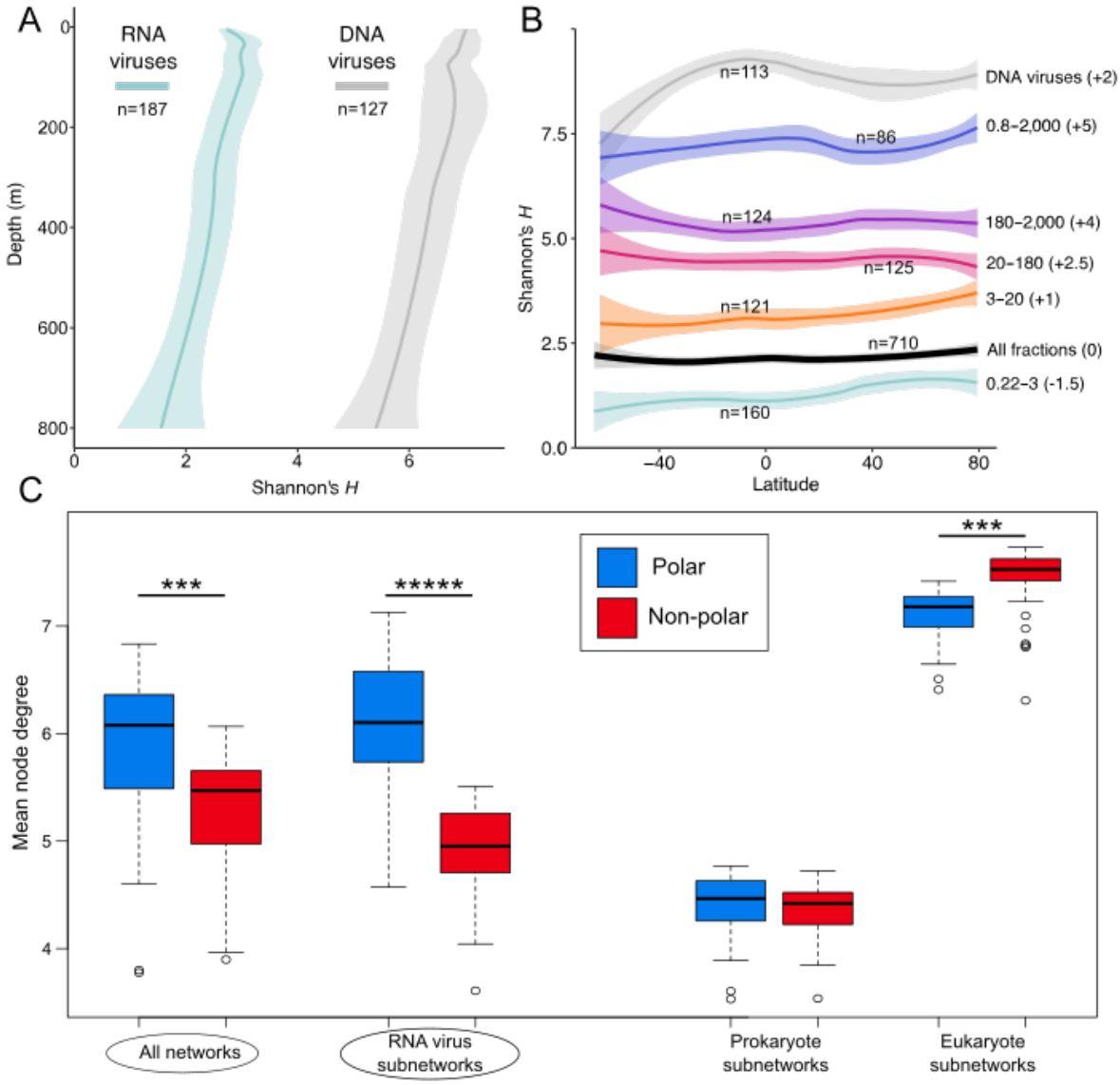

Fig. 4. RNA virus "species"-level diversity across depth and latitudinal gradients. (**A**) Locally estimated scatterplot smoothing (LOESS) smooth plots showing the depth distributions of "species"-level diversity for RNA and DNA viruses. Shown are only RNA viruses from the prokaryotic fraction due to the very limited number of deep ocean samples from the eukaryotic fraction (eukaryotic fraction results shown in fig. S3). (**B**) LOESS plots showing the latitudinal distributions of "species"-level diversity for DNA (grey) and RNA viruses (remaining colors). Plots are nudged along the y-axis (with a baseline offset as indicated in the parentheses on the right) for visibility. Size fraction and nudge value are indicated next to each plot, with the

collective estimate of Shannon's *H* values across all the size fractions of RNA viruses shown in (black). On all of the smoothing plots, the lines represent the LOESS best fit for the samples included (n), whereas the lighter band corresponds to the 95% confidence interval of the fit (also see fig. S4–5 for vOTU definition sensitivity analyses). (**C**) Global and organismal domain-specific co-occurrence networks connectivity (mean node degree) in polar vs. non-polar samples showing that the significantly higher connectivity in the polar waters (red ellipses) is driven solely by RNA viruses. All boxplots show medians and quartiles. Statistical significance was assayed by the Mann-Whitney U test and is documented in the figure as follows: ∗ adjusted $p<0.05$, ∗∗∗ adjusted $p<0.0001$, and ∗∗∗∗∗ adjusted $p<0.0000001$.

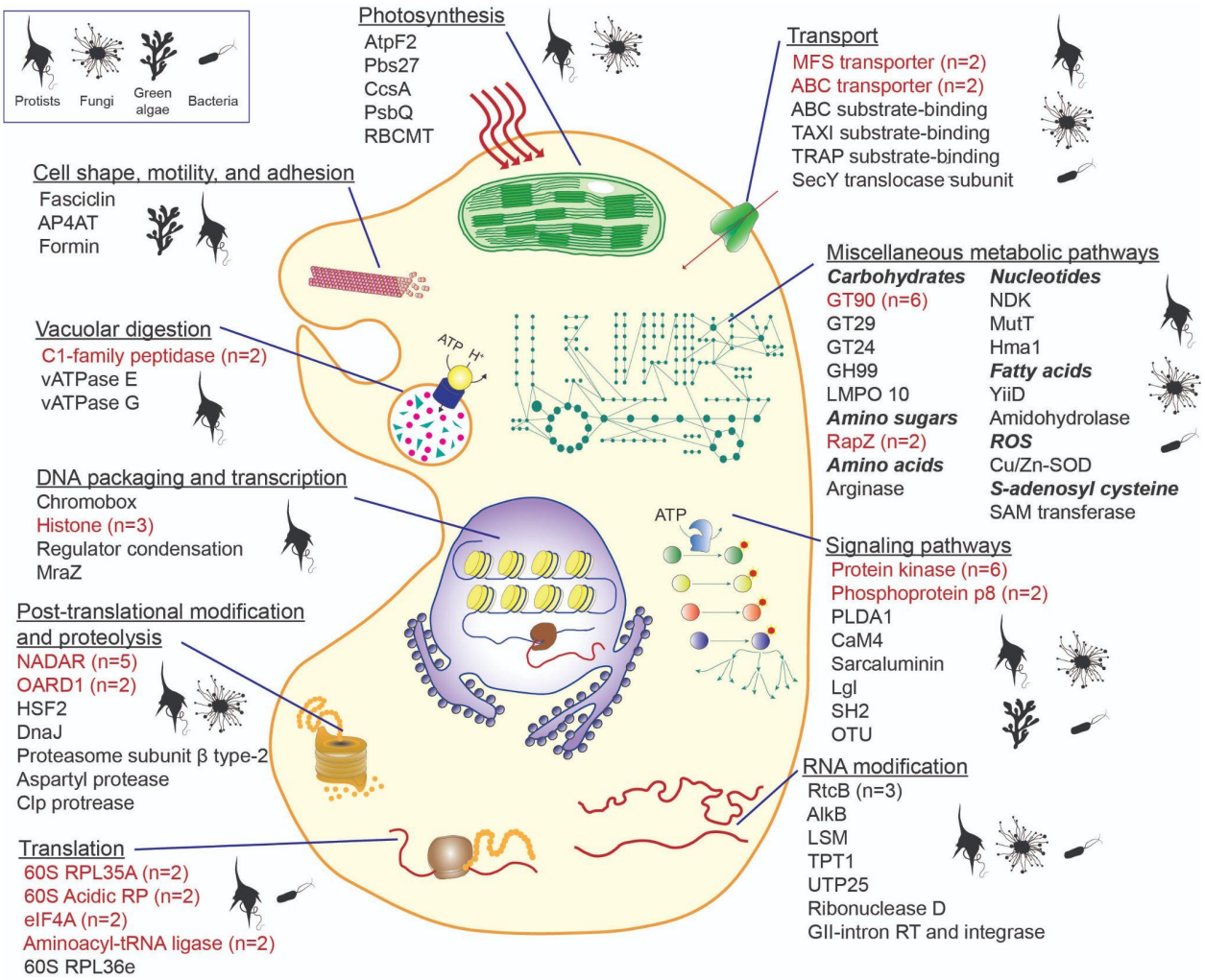

Fig. 5. Functional diversity of AMGs carried by marine RNA viruses. Schematic representation of the hypothesized roles played in manipulation of host metabolism by RNA virus AMGs, which are separated according to functional categories. Red text corresponds to proteins that were found encoded independently in several vOTUs with the number of vOTUs listed in parentheses. The putative hosts, inferred using available information for RNA viruses with established orthornaviran taxonomy, are indicated by organism silhouettes in each section. Inferred plants were interpreted as their closest relatives, chlorophytes (green algae), in the marine environment. Bacteria were inferred from picobirnavirids. Annotated proteins associated with multiple, disparate cellular processes, or whose function remains obscure, are not shown (see annotation details for corresponding vOTUs and virus contigs in **table S6**).

Science
AAAS

Supplementary Materials for

Diversity and ecological footprint of Global Ocean RNA viruses

Guillermo Dominguez-Huerta, Ahmed A. Zayed, James M. Wainaina, Jiarong Guo, Funing Tian, Akbar Adjie Pratama, Benjamin Bolduc, Mohamed Mohssen, Olivier Zablocki, Eric Pelletier, Erwan Delage, Adriana Alberti, Jean-Marc Aury, Quentin Carradec, Corinne da Silva, Karine Labadie, Julie Poulain, ***Tara* Oceans Coordinators**, Chris Bowler, Damien Eveillard, Lionel Guidi, Eric Karsenti, Jens H. Kuhn, Hiroyuki Ogata, Patrick Wincker, Alexander Culley, Samuel Chaffron, and Matthew B. Sullivan

Correspondence to: sullivan.948@osu.edu

This PDF file includes:

Materials and Methods
Supplementary Text
Figs. S1 to S8
Captions for Tables S1 to S7
Captions for Data S1 to S2

Other Supplementary Materials for this manuscript include the following:

Tables S1 to S7
Table S1. Ocean RNA virus studies.
Table S2. List of RNA and DNA samples, their metadata, and their unique identifiers.
Table S3. List of additional example literature (section A) and Global Ocean surveys (section B) studying the central ecological roles of marine plankton.
Table S4. Correlations between environmental variables and alpha diversity for RNA and DNA viruses.
Table S5. Host prediction results for the RNA vOTUs identified in this study.
Table S6. AMGs encoded by RNA viruses in this study and previous literature.
Table S7. RNA vOTU modules (subnetworks) predictive of ocean carbon flux.

Data S1 to S2
Data S1. Gene schemes of vOTU representatives carrying 19 AMG functional types observed either in short-read or long-read sequencing data.
Data S2. Endogenous virus elements detected in this study.

Materials and Methods

Sampling, purification of nucleic acids, library preparation, and short- and long-read sequencing

We used all available sequencing data (≈28 Tb) derived from 771 metatranscriptomes collected across 121 sites in the framework of the *Tara* Oceans (TO) and *Tara* Oceans Polar Circle (TOPC) expeditions (2009–2013). The 771 metatranscriptomes are derived from 187 prokaryote-enriched size fractions from the TO and TOPC expeditions (≈5.3 Tb of data), that were previously published (Salazar et al. 2019), and 441 eukaryote-enriched size fractions from the TO expedition (≈16.3 Tb), that were published in (Carradec et al. 2018). The remaining 143 metatranscriptomes were generated from eukaryote-enriched samples (≈6.3 Tb of data) collected during the TOPC expedition (Zayed et al. 2022; BioProjects PRJEB9738 and PRJEB9739). Technical details of sampling and size fractionation procedures for both *Tara* expeditions were previously published (Pesant et al. 2015). Specific operational sizes are described in **table S2**. A size scale of the plankton organisms captured in TO and TOPC campaigns is provided in **Fig. 1A**, dividing into prokaryote- and eukaryote-enriched size fractions.

Different optimized protocols for extraction and purification of RNA, for which a full description of details have been previously published (Alberti et al. 2017), were applied depending on the type of organismal fraction. Despite the recalcitrancy to lysis of certain taxonomic groups of protists due to complex cell envelopes (e.g., diatoms, dinoflagellates), the optimized protocols were tested against the Roscoff marine culture collection (https://roscoff-culture-collection.org/). After purification of nucleic acids, potentially contaminant DNA was digested by DNase (Alberti et al. 2017). Different protocols of rRNA depletion and library preparation were followed according to campaign, type of organismal size fraction, and total RNA available for input. The most significant features and steps of the different library building approaches, as well as their estimated bias introduced to capturing RNA virus molecules with or without poly(A)-tails, are described in **fig. S1**. Libraries were short-read sequenced with 101 base-length read chemistry in a paired-end flow cell on HiSeq2000 or HiSeq2500 sequencing machines (Illumina) (Alberti et al. 2017). A subset of 20 RNA samples utilized here for standard Illumina short-read sequencing were also used for long-read sequencing (please refer to Zayed et al. 2022 for technical details).

Our collection of organismal size fractions covers a broad spectrum of ranges, and hence, it allows to discriminate multiple organismal types from plankton (from prokaryotes to fish larvae) across the Global Ocean. However, it is worth noting that the different plankton organismal size fractions are still loosely associated to a certain number of organism types, due to known intrinsic spillover among fractions during sampling. For example, plankton biomass of the fractions labeled as "prokaryotic" are expected to be enriched in bacteria and archaea, but can also contain viruses (likely predominantly intracellular), giant viruses (intracellular or as free particles), and picoeukaryotes. Given that prokaryotic mRNA lacks poly(A)-tails, selection of poly(A)-tailed RNA [either by oligo(dT)-bead capture or oligo(dT)-priming] was not applied as an rRNA depletion step to such prokaryote-enriched samples, but only random priming for reverse transcription after degradation of rRNA (**fig. S1A**). Given this, we presume prokaryotic plankton fractions are likely to have most randomly captured the diversity of virus types in these samples. Random hexamer priming without poly(A)-tail selection is known to conserve the proportions of mock communities (Fitzpatrick et al. 2021), and has been used extensively in diverse high-throughput RNA virus studies (Wolf et al. 2020; Li et al. 2020; Starr et al. 2019; Wille et al. 2019). Hence, we do not see obvious sources of biases for the capture of RNAs from virus taxa that lack poly(A)-tails (**fig. S1B**).

In contrast, selection of poly(A)-tailed RNA is performed as an rRNA depletion step for eukaryotic plankton fractions, and this undoubtedly introduces a bias against virus RNAs that are not poly(A)-tailed (**fig. S1A**). Given this potential bias associated with library-building approaches, we explored the proportion of orthornaviran taxa of viruses expected to carry poly(A)-tails across prokaryote- and eukaryote-enriched samples (**fig. S1B**). This analysis showed that the two orders and seven families currently established for RNA viruses with poly(A)-tailed genomes and/or transcripts were enriched in poly(A)-selected, size-fractionated eukaryotic datasets, despite our inability to systematically disentangle the influence of fractionation size (prokaryote- and eukaryote-enriched samples) from the one derived from library building approaches. However, for the global-scale ecological gradients explored here, there appears to be a significantly stronger biological (e.g., diversity and community structure patterns) signal than experimental (e.g., size fractions and library building methods) signal (**Fig. 1–4; fig. S2–5; fig. S7**). These results suggest that the scope of the analyzed metatranscriptomic sequencing data (derived from either eukaryote- or prokaryote-enriched plankton fractions) is sufficient for our goal of studying the diversity and ecology of marine RNA viruses, in spite of the different library building approaches utilized across the 771 plankton metatranscriptomes.

Metatranscriptome assembly, virus identification, and taxonomic classification

Detailed description of the read assembly, virus identification, and taxonomic classification methods can be found in (Zayed et al, 2022). Briefly, reads were assembled *de novo* into contigs using MEGAHIT v1.1.3 with default parameter settings. Bioinformatic identification of contigs derived from RNA virus genomes or transcripts across the metatranscriptomic data was based on screening for domain protein sequences of virus RNA-directed RNA polymerases (RdRp) using hidden Markov models (HMMs). In addition, to confirm that the RNA virus contigs originated in active infections, and not from endogenous viruses, their sequences were searched against the available collection of co-sampled metagenomes (see "Functional annotation of RNA virus genomes" below). Taxonomic classification was conducted using both a benchmarked iterative clustering approach and phylogenetic analyses along with reference RNA virus RdRp sequences (full details in Zayed et al, 2022).

Establishment of genome-based virus operational taxonomic units (vOTU) and robustness of the ecological patterns to fragment length and vOTU definition.

A grand challenge in ecology and evolutionary biology is to identify uniform units of interactions (i.e., ecological units) for the studied organisms. Though relatively agreed upon for large organisms, identification of these units has long been controversial for prokaryotes as horizontal gene flow is considered likely to blur species boundaries (Boucher et al. 2003). However, as some prokaryotic lineages have been characterized by deep whole-genome sequencing and assessed using a gene-flow- and selection-based biological species definition, a new paradigm has emerged for prokaryotes, namely that gene flow is likely higher within members of a species than among members of distinct species and that gene flow itself is what creates species cohesion (Jesse Shapiro et al. 2012; Cadillo-Quiroz et al. 2012). For viruses, such gene flow is often considered to be much more frequent with "rampant mosaicism" presumed to result in virus genomic sequence space tending towards a continuum such that any "species-rank" structure that a researcher observes is a function of shallow and/or biased sampling (Hendrix et al. 1999). However, studying large swaths of dsDNA viruses, whether linked to a common host (Deng et al. 2014; Gregory et al. 2016) or not (Bobay and Ochman 2018; Gregory et al. 2019),

demonstrated that dsDNA virus sequence space was structured, which was thought due to the studied biomes (e.g., ocean biomes) not being affected by high levels of gene-flow (Mavrich and Hatfull 2017). Additionally, this sequence space structure has been attributed to population-constrained gene flow and selection (Gregory et al. 2016; Bobay and Ochman 2018), fitting a biological species definition.

Though RNA viruses are known to be subject to high rates of mutation and gene flow (Duffy 2018; Bobay and Ochman 2018), we wondered whether their sequence space might yet also reveal structure emergent from our Global Ocean datasets (extensively tested in Zayed et al., 2022) and hence establish working ecological units loosely connected to a "biological species definition" (*sensu* (Roux et al. 2019)). Briefly, we empirically evaluated whether "structure" emerges from all-versus-all comparisons of our 44,779 virus contigs using MUMmer v3.23 after excluding self matches. This was achieved by (i) computing the frequency of two values – the average nucleotide identity (ANI) and the alignment fraction of the shorter contig (AF) – across all contig pairs ≥1 kb, and (ii) searching for emergent clusters that can represent virus Operational Taxonomic Units (vOTUs). In our data, vOTU clustering thresholds that resulted in two different groups of contig pairs with high frequency were 90% ANI across 80% of the shorter sequence length, resulting in 5,504 vOTUs ≥1 kb. Such vOTUs represent working ecological units at an approximately species-rank taxonomy that are emergent from analyses suggested via research community consensus (Roux et al. 2019), but their formal taxonomic assessment would require examination for gene flow and selection from whole-genome population data as done for cyanophages and mycobacteriophages (Gregory et al. 2016; Bobay and Ochman 2018). Critically, although we found ocean RNA virus sequence space to be structured (Zayed et al., 2022), as had the community consensus effort (Roux et al. 2019), our empirical cutoffs were different (ANI, AF, and whole contig ANI (wcANI) = 90%, 80%, and 72%, respectively (Zayed et al., 2022) due to the community effort combining data from DNA and RNA viruses and ours focusing solely on RNA viruses.

We then applied this re-evaluated vOTU definition across our dataset to make ecological inferences (**Fig. 1–4; fig. S2–3; fig. S7**), while also conducting extensive sensitivity analyses to evaluate the impact of the cutoffs chosen on these ecological inferences (**fig. S4–5**). First, ecological patterns were robust to genome fragment lengths of ≥1 kb, ≥2 kb, or ≥3 kb (**fig. S4**). Second, ecological patterns were robust to ANI/AF clustering cutoffs across an ANI range of 60% to 100% (wcANI range of 48–80%; **fig. S5**), with the largest of these investigated wcANI values (≥80%) being the recommended value in the viral community consensus statement (Roux et al. 2019). These results indicate that, even though intra-vOTU (intra-population) analyses can be impacted by vOTU clustering cutoffs, the between-vOTU (between-population) diversity metrics investigated in our study are not sensitive to vOTU clustering cutoffs (at least below the very strict wcANI value chosen by the viral community in their consensus statement (Roux et al. 2019)). Thus, the ecological inferences in our study here are robust against methodological differences between our study and previous RNA virus studies.

Calculations of vOTU relative abundances and diversity metrics

The calculation of vOTU relative abundances is extensively described in (Zayed et al., 2022). Briefly, trimmed reads from each library and the virus contigs were first further trimmed off their polyA and polyT stretches to avoid inflated abundances and better estimate horizontal coverage for polyA-tailed viruses, respectively. PolyA/T-trimmed reads were mapped against all polyA/T-trimmed contigs using Bowtie2 v2.4.1 using the very sensitive, local, and non-deterministic

settings with the additional increase of sensitivity by reducing the word size to 16. The final abundances of the vOTUs were calculated by summing the adjusted abundances (by the number of mapped reads) of the contigs belonging to these vOTUs.

The final relative abundances calculated above were used to estimate the different diversity metrics. Estimation of α- (using the Shannon's index method) and β- (using the Bray-Curtis dissimilarity method) diversity statistics was performed by using the "vegan" package (Dixon 2003) in R (functions 'diversity ("shannon")' and 'vegdist ("bray")', respectively), with the abundance table being log-transformed before calculation of the dissimilarity matrix (function 'decostand("log")'). The dissimilarity matrix was non-linearly visualized in two dimensions (**Fig. 1B**) using the t-distributed Stochastic Neighbor Embedding (t-SNE) method using package "Rtsne" in R, with a perplexity value of 150 and maximum number of iterations of 1,000 (seed = 5). We conducted metric multidimensional scaling for correlation of the community composition to environmental variables. The dissimilarity matrix underwent dimensionality reduction (**fig. S2**) in a Principal Coordinate Analysis (function 'capscale' of the "vegan" package with no constraints applied). Hierarchical clustering (function pvclust; method.dist = "cor" and method.hclust = "average") was conducted on randomly subsampled sets (using the 'sample' function without replacement) from the prokaryotic fraction after calculating Bray-Curtis dissimilarity matrices with 1,000 bootstrap iterations and reporting the approximately unbiased (AU) bootstrap values. Only the prokaryotic fraction was used in this analysis due to the limited number of deep ocean samples in the other fractions (**fig. S2D**). The subsampling was done down to the number of deep ocean samples (if applicable) in an attempt to achieve balance in the representation from the different ecological zones upon conducting the hierarchical cluster. The heatmaps were generated using the heatmap3 package. Locally estimated scatterplot smoothing (LOESS) plots of Shannon's *H* values across latitude and depth (**Fig. 4A–B**) were generated using the 'geom_smooth' function of the package ggplot2 in R. Samples in which RNA virus contigs did not recruit sufficient reads to confidently call the presence of vOTUs (078_DCM, 128_DCM, 158_SRF_R1, and 158_SRF_R2; all from the 180–2,000-µm size fraction) were excluded from all ecological analyses. The three non-matching samples (of 118 shared samples) among the RNA and DNA viruses in the ecological zone analysis (**Fig. 1B**) were all obtained from subpolar station 155, where Atlantic and Arctic fronts meet (**Fig. 1A**).

Inference of potential diversity drivers

Regression analysis of the first coordinate of the Principal Coordinate Analysis (PCoA1) and *in situ* temperature measurements (as well as of the median Shannon's *H* values among DNA and RNA viruses) was conducted using the 'lm' function and visualized using the 'geom_smooth' of package "ggplot2" in R. Environmental variables were correlated with both PCoA1 and Shannon's *H* following both the Pearson's and Spearman's correlation methods using function 'cor("pairwise.complete.obs")' after removing Shannon's *H* outliers (e.g., samples 19_SRF and 205_SUR from the prokaryotic fraction) based on the boxplot analysis shown in **Fig. 2** for the α-diversity correlations. Correlation coefficients and *p*-values for both Pearson's and Spearman's tests are listed in **table S4**. All boxplot analyses were performed using functions 'geom_boxplot' and 'stat_boxplot(geom="errorbar")' of "ggplot2", plotting medians and standard quartiles, and statistically comparing the groups using an analysis of variance (function 'aov') and calculating the corrected *p*-values using the Tukey Honest Significant Differences method (function 'TukeyHSD').

To reduce collinearity in the large number of metadata measurements examined here, we used the Weighted Gene Correlation Network Analysis (WGCNA) method (package "WGCNA" in R) to construct modules of co-varying environmental variables as shown in **Fig. 3**. The environmental measurements were first standardized using the function 'decostand("standardize")' in the package "vegan". The standardized values were used to build a scale-free topology network (power transformation on the edges was picked to be 12) and the modules were defined using the 'blockwiseModules' function of "WGCNA" (networkType = "signed", TOMType = "signed", minModuleSize = 1, reassignThreshold = 0, deepSplit = 3, mergeCutHeight = 0.1, pamRespectsDendro = FALSE). The resultant modules were correlated to Shannon's *H* of both DNA and RNA viruses and the modules that had statistically significant correlations were picked for Pearson and Spearman analyses shown in **Fig. 3**.

Inference of hosts

Association of vOTUs and putative hosts were assessed based on three independent strategies (**table S5**), based on the following types of information: (i) available host information for viruses of established orthornaviran taxa, (ii) co-occurrence networks inferred from relative abundances (modified from (Kaneko et al. 2020)), and (iii) protein sequence similarity of RdRPs to EVEs (modified from (Shi et al. 2016)). Technical details of the three host-inference strategies were described in (Zayed et al. 2022), where only the cellular domain of hosts (prokaryotes or eukaryotes) was reported. Here we reported deeper taxonomies for the putative hosts, given their ecological value, albeit with different taxonomic resolution depending on the nature of the data. Specifically, very general taxonomic ranks were reported when only using inferences from established virus taxonomy, ranks lower than order from EVE signatures, and phyla from abundance-based co-occurrences. Given the marine context of this study and the lack of established orthornaviran taxa linked to microbial eukaryotes, inferred "plants" and "streptophytes" may be interpreted as chlorophytes (green algae), their closest relatives in the oceans. Bacteria inferred in the taxonomy-based strategy were derived from families linked to RNA phages by either empirical (leviviricete families and *Cystoviridae*) or *in silico* evidence (*Picobirnaviridae*) (Krishnamurthy and Wang 2018).

Functional annotation of RNA virus genomes

Prior inference of utilized genetic codes (**table S5**), the gene content and protein domains of vOTUs were explored by functional annotation of proteins predicted from their representative nucleotide sequences against several databases, for which technical details are described in (Zayed et al., 2022).

Briely, protein sequences encoded by vOTUs were considered of "cellular" origin (shared with cellular organisms), and hence auxiliary metabolic genes (AMGs) (see annotations in **table S6** and gene schemes in **Data S1**), if they fulfilled the following criteria: (i) the blastp topmatch in the NCBI nr database was both bit score >50 and cellular, and/or (ii) the Pfam hit with >95% probability score in hhblits was cellular, and did not include virus representatives in the corresponding species distribution of Pfam, or was not linked to any established virus HMM profile. To avoid overlapping annotation, separation of virus and "cellular" protein signal along the contig was manually confirmed for all cases. Due to possible high error rate typically associated with long-read sequencing, protein sequences encoded by RNA virus long reads were considered "cellular" only if the blastp topmatch was a cellular homolog protein. To confirm and refine the

functional annotation of such AMGs, additional annotation was performed against multiple databases (KEGG, MEROPS, VOG, and CAZY) by using DRAM (Shaffer et al. 2020).

To discard the possibility that the observed AMG sequences were derived from EVEs contained in potential, contaminating DNA (and hence not carried by actively replicating RNA viruses), all 44,779 virus contig sequences were searched against 52,283 curated *Tara* Oceans Single-Cell and Metagenome Assembled Genomes ("SMAGs", https://www.genoscope.cns.fr/tara/) comprising 10 million nucleotide sequences, using blastn with at least 95% nucleotide identity and 95% alignment fraction of the virus contigs whose length was shorter than the matching SMAG sequence. The only 14 matching virus contigs were considered derived from EVEs harbored in SMAGs (**Data S2**).

Carbon export analyses

Carbon exports values (carbon flux at 150 m; see **table S2**) were estimated based on particle size distributions and concentrations measured with the Underwater Vision Profiler (UVP (Picheral et al. 2010)) and according to previously validated methods (Guidi et al. 2008, 2016). For each biological sample, the associated carbon export value corresponds to the average flux of carbon between 125 and 175 m and within 100 km and two days of the sampling site.

Carbon export values were used in a WGCNA analysis as described above to infer the vOTUs that potentially play a role in this important process. Briefly, the abundance table of the vOTUs was subsetted per size fraction and each daughter table was independently processed through a three-step analysis: (i) building a WGCNA network and defining the modules (subnetworks) that correlate with carbon export, (ii) performing a cross-validated partial least squares regression analysis for the resultant modules individually to build a predictive model for carbon export based on vOTU abundances, and (iii) conducting a variable selection process to highlight potentially most important vOTUs for carbon export within each subnetwork. Notably, all of the vOTUs (n=5,504) were included in building the WGCNA network, with the final modules/subnetworks containing 45, 25, 1156, and 17 vOTUs for subnetworks 1, 2, 3, and 4 shown in **fig. S7**, respectively. Additionally, we sought to represent short/unassembled contigs within each vOTU in the calculated abundances (i) by using the most sensitive set of settings during the read mapping step and (ii) by allowing the read mapping cutoffs to encompass our definition of the vOTU boundaries (see "**Calculation of vOTU relative abundances**" above). Hence, these two steps maximized capturing the relative abundances of the contigs within the same vOTU even if they were not assembled or included in our dataset (as a result of our < 1 kb length cutoff or our RdRP domain completeness cutoff).

For the WGCNA analysis, the per-size-fraction tables were first Hellinger-transformed 'decostand("hellinger")' and the modules were defined using the following parameters in the function 'blockwiseModules': corType="bicor", maxBlockSize = 10,000, networkType = "signed", TOMType = "signed", minModuleSize = 15, reassignThreshold = 0, deepSplit = 3, mergeCutHeight = 0.1, pamRespectsDendro = FALSE, replaceMissingAdjacencies = TRUE. The highly correlated modules were then regressed in a partial least-squares analysis (function 'plsr' of package "pls" in R with the Leave-One-Out cross validation and Orthogonal Scores methods). The modules that passed the cross validation step were then selected for variable extraction using the Variable Importance in Projection method as implemented in VIP.R (http://mevik.net/work/software/VIP.R). vOTUs with both high VIP score and high Pearson's

correlation coefficient to carbon export are shown in **fig. S7**) and the hosts that they potentially infect were determined as described below.

Supplementary Text

The *Tara* Oceans Coordinators and Affiliations

Silvia G. Acinas[1,‡], Marcel Babin[2,‡], Peer Bork[3,4,5,‡], Emmanuel Boss[6,‡], Chris Bowler[7,‡], Guy Cochrane[8,‡], Colomban de Vargas[9,‡], Gabriel Gorsky[10,‡], Lionel Guidi[10,11,‡], Nigel Grimsley[12,13,‡], Pascal Hingamp[14,‡], Daniele Iudicone[15,‡], Olivier Jaillon[16,17,18,‡], Stefanie Kandels-Lewis[3,19,‡], Lee Karp-Boss[6,‡], Eric Karsenti[7,19,‡], Fabrice Not[20,‡], Hiroyuki Ogata[21,‡], Nicole Poulton[22,‡], Stéphane Pesant[23,24,‡], Christian Sardet[10,25,‡], Sabrina Speich[26,27,‡], Lars Stemmann[10,‡], Matthew B. Sullivan[28,29,‡], Shinichi Sunagawa[30,‡], and Patrick Wincker[16,17,18,‡].

[1]Department of Marine Biology and Oceanography, Institut de Ciències del Mar (CSIC), Barcelona, Catalonia, Spain.
[2]Département de biologie, Québec Océan and Takuvik Joint International Laboratory (UMI3376), 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, 13125 Berlin, Germany.
[5]Department of Bioinformatics, Biocenter, University of Würzburg, 97074 Würzburg, Germany.
[6]School of Marine Sciences, University of Maine, Orono, Maine 04469, USA.
[7]Ecole Normale Supérieure, PSL Research University, Institut de Biologie de l'Ecole Normale Supérieure (IBENS), CNRS UMR 8197, INSERM U1024, 46 rue d'Ulm, F-75005 Paris, France.
[8]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Welcome Trust Genome Campus, Hinxton, Cambridge, UK.
[9]CNRS, UMR 7144, EPEP & Sorbonne Universités, UPMC Université Paris 06, Station Biologique de Roscoff, 29680 Roscoff, France.
[10]Sorbonne Universités, UPMC Université Paris 06, CNRS, Laboratoire d'oceanographie de Villefranche (LOV), Observatoire Océanologique, 06230 Villefranche-sur-Mer, France.
[11]Department of Oceanography, University of Hawaii, Honolulu, Hawaii 96822, USA.
[12]CNRS, UMR 7232, BIOM, Avenue du Fontaulé, 66650 Banyuls-sur-Mer, France.
[13]Sorbonne Universités Paris 06, OOB UPMC, Avenue du Fontaulé, 66650 Banyuls-sur-Mer, France.
[14]Aix Marseille Univ, Université de Toulon, CNRS, IRD, MIO, Marseille, France.
[15]Stazione Zoologica Anton Dohrn, Villa Comunale, 80121 Naples, Italy.
[16]CEA - Institut de Génomique, Genoscope, 2 rue Gaston Crémieux, Evry France.
[17]CNRS, UMR 8030, 2 rue Gaston Crémieux, Evry France.
[18]Université d'Evry, UMR 8030, CP5706, Evry France.
[19]Directors' Research European Molecular Biology Laboratory Meyerhofstr. 1 69117 Heidelberg Germany.
[20]CNRS, UMR 7144, Sorbonne Universités, UPMC Université Paris 06, Station Biologique de Roscoff, 29680 Roscoff, France.

[21]Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto, 611-001, Japan.
[22]Bigelow Laboratory for Ocean Sciences, East Boothbay, ME, 04544, USA.
[23]MARUM, Center for Marine Environmental Sciences, University of Bremen, Bremen, Germany.
[24]PANGAEA, Data Publisher for Earth and Environmental Science, University of Bremen, Bremen, Germany.
[25]CNRS, UMR 7009 Biodev, Observatoire Océanologique, F-06230 Villefranche-sur-mer, France.
[26]Laboratoire de Physique des Océans, UBO-IUEM, Place Copernic, 29820 Plouzané, France.
[27]Department of Geosciences, Laboratoire de Météorologie Dynamique (LMD), Ecole Normale Supérieure, 24 rue Lhomond, 75231 Paris Cedex 05, France.
[28]Department of Microbiology, The Ohio State University, Columbus, OH 43214, USA.
[29]Department of Civil, Environmental and Geodetic Engineering, The Ohio State University, Columbus OH 43214 USA.
[30]Department of Biology, Institute of Microbiology and Swiss Institute of Bioinformatics, ETH Zurich, Vladimir-Prelog-Weg 4, 8093 Zurich, Switzerland.
[‡]Tara Oceans Consortium.

Extended list of acknowledgements and funding

Tara Oceans (which includes both the *Tara* Oceans and *Tara* Oceans Polar Circle expeditions) would not exist without the leadership of the *Tara* Expeditions Foundation and the continuous support of 23 institutes (http://oceans.taraexpeditions.org). We further thank the commitment of the following sponsors: CNRS (in particular Groupement de Recherche GDR3280 and the Research Federation for the study of Global Ocean Systems Ecology and Evolution, FR2022/*Tara* Oceans-GOSEE), European Molecular Biology Laboratory (EMBL), Genoscope/CEA, the French Ministry of Research, the French Government's 'Investissements d'Avenir' programmes OCEANOMICS (ANR-11-BTBR-0008), FRANCE GENOMIQUE (ANR-10-INBS-09-08), MEMO LIFE (ANR-10-LABX-54), and PSL* Research University (ANR-11-IDEX-0001-02), GENCI grants (t2011076389, t2012076389, t2013036389, t2014036389, t2015036389, and t2016036389) for HPC computation, Swiss National Science Foundation (SNF - 205321_184955), Gordon and Betty Moore Foundation (award #3790), U.S. National Science Foundation (awards OCE#1829831, ABI#1759874, and DBI# 2022070), Ohio State University Center of Microbiome Science's support to M.B.S., the Ohio Supercomputer for computational support, and a Ramon-Areces Foundation Postdoctoral Fellowship to G.D-H. Funding for the collection and processing of the *Tara* data set was provided by NASA Ocean Biology and Biogeochemistry program under grants NNX11AQ14G, NNX09AU43G, NNX13AE58G and NNX15AC08G to the University of Maine and Canada Excellence Research Chair on *Remote sensing of Canada's new Arctic frontier* Canada foundation for innovation.

We also thank the support and commitment of agnès B. and Etienne Bourgois, the Prince Albert II de Monaco Foundation, the Veolia Foundation, Région Bretagne, Lorient Agglomeration, Serge Ferrari, Worldcourier, and KAUST. The global sampling effort was enabled by countless scientists and crews who sampled aboard the *Tara* from 2009–2013. We thank MERCATOR-CORIOLIS and ACRI-ST for providing daily satellite data during the expeditions. We are also grateful to the countries who graciously granted sampling permissions.

J.H.K. performed this work as an employee of Tunnell Government Services (TGS), a subcontractor of Laulima Government Solutions, LLC, under Contract No. HHSN272201800013C. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Department of Health and Human Services, or of the institutions and companies affiliated with the authors.

Figures

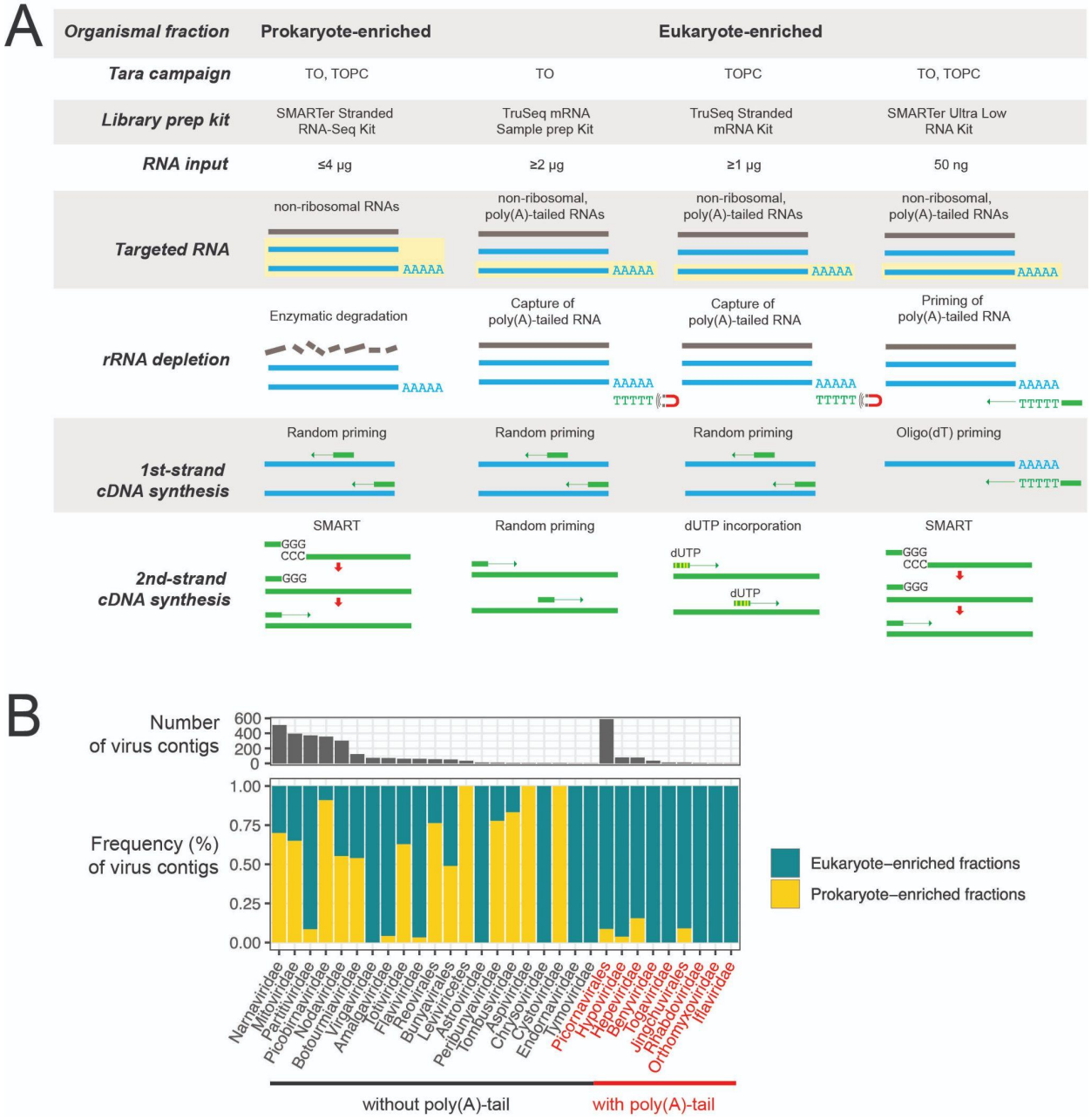

Fig. S1. Plankton organismal sizes considered in this study and their sequencing libraries construction.

(A) Tabular schematics summarizing key features (left) of library construction applied (described with detailed in Zayed et al., 2022) to RNA derived from different size fractions (prokaryote- or eukaryote-enriched) collected during Tara Arctic (TOPC) and non-Arctic (TO) campaigns. Molecules of rRNA, virus RNA, and cDNA are represented by grey, blue, and green bars, respectively. Targeted RNA molecule types by the different library building approaches are highlighted within a yellow rectangle. Depletion of rRNA was achieved either by enzymatic degradation or selection of poly(A)+ RNA. After synthesis of 1st-strand cDNA by random

priming or oligo(dT) priming, the second strand was generated following stranded (SMART template-switching technology or dUTP incorporation) or non-stranded (random priming) methods. **(B)** Stacked bar plot showing the proportions of virus contigs derived from prokaryote- (yellow) and eukaryote-enriched (blue) fractions that were observed across RNA virus classes, orders, and families of viruses with genomes and/or transcripts known to lack (black text) or possess (red text) poly(A)-tails. The stacked bar plot indicates that different approaches applied to build the RNA library before sequencing could affect the captured RNA virus diversity as the RNA virus taxa with poly(A)-tailed genomes and/or transcripts were enriched in poly(A)-selected, size-fractionated eukaryotic datasets. The taxonomic classification utilized in the histogram was taken from the RdRP phylogenies built in (Zayed et al., 2022).

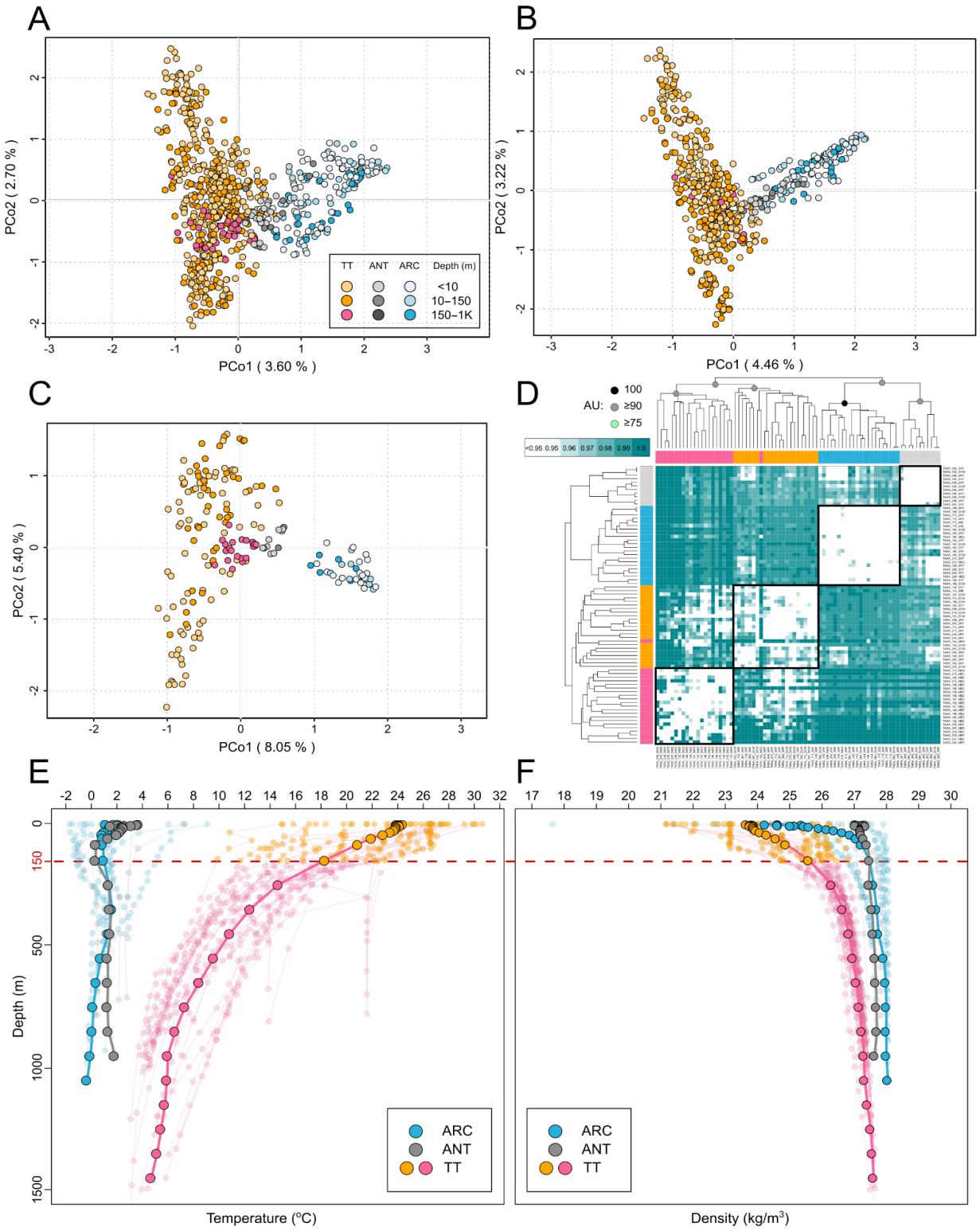

Fig. S2. Meta-community structure evaluation of the Global Ocean RNA viruses.
(A) Principal Coordinate Analysis of a Bray-Curtis dissimilarity matrix of all the samples in this study across the prokaryotic and eukaryotic size fractions showing the relative relationships of

the ecological zones. The eukaryotic **(B)** and prokaryotic **(C)** subsets of the PCoA analysis are also individually shown to highlight the independence of the ecological zones from the size fraction and library preparation methods. **(D)** Correlation-based hierarchical clustering of a Bray-Curtis dissimilarity matrix calculated from a randomly subsampled set from the prokaryotic fraction (the only fraction with >10 deep ocean samples) down to the number of samples from the deep ocean to roughly obtain balanced groups (see **Methods**). The hierarchical clustering analysis structured the viromes into four distinct global ecological zones (cyan; Arctic, grey; Antarctica, orange; Temperate and Tropical Epipelagic; pink; Temperate and Tropical Mesopelagic) with an approximately unbiased (AU) bootstrap value ≥90. **(E–F)** Thermoclines **(E)** and pycnoclines **(F)** at the shared stations between DNA and RNA viruses. Stratification-based decoupling of epipelagic and mesopelagic communities at 150 m in temperate and tropical regions can be explained by the strong thermoclines and pycnoclines. Polar regions, on the other hand, only show shallow density gradients in the uppermost, wind-mixed layers due to freshwater inputs impacing salinity and mixing in the water column. Subarctic and Mediterranean Sea stations were excluded from this analysis due to their non-uniform temperature and salinity measurements, respectively. Color scheme is the same as in (D). Faint dots represent the actual measurements made during the *Tara* Oceans expeditions (randomly subsampled to 1/300 for visibility) and faint lines connect the samples from the same station. Solid dots and lines represent the average of the individual measurements at the same depth. Samples in which RNA viruses did not recruit enough reads to confidently call the presence of vOTUs (078_DCM, 128_DCM, 158_SRF_R1, and 158_SRF_R2; all from the 180–2,000 µm size fraction) were excluded from all the ecological analyses. ANT, Antarctic; ARC, Arctic; TT, temperate and tropical.

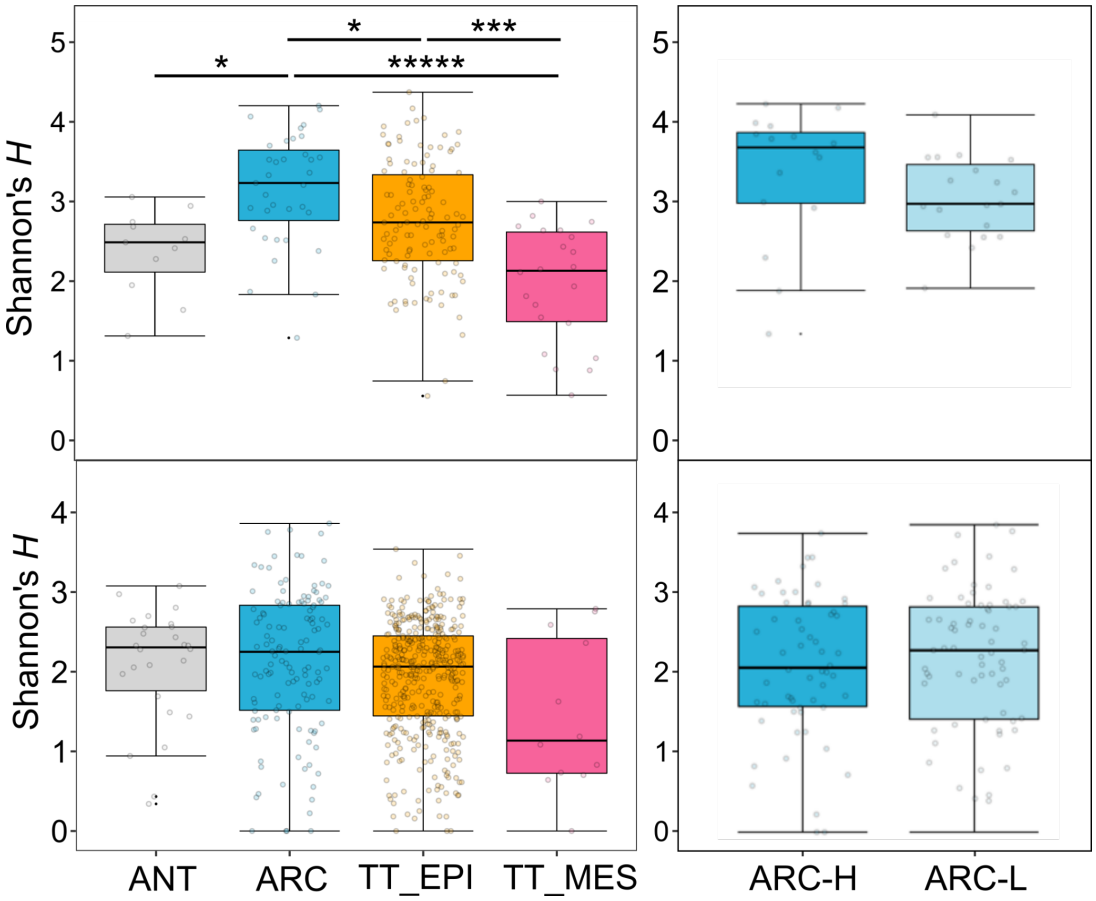

Fig. S3. Regional macrodiversity of RNA viruses.

Boxplots analyses of RNA virus macrodiversity across four ecological zones (left) and within the Arctic Ocean (right) for the prokaryotic (top) and eukaryotic (bottom) fractions. All boxplots show medians and quartiles. Statistical significance as assayed by the Tukey Honest Significant Differences method on an analysis of variance is documented in the figure as follows: ∗ adjusted $p<0.05$, ∗∗∗ adjusted $p<0.0001$, and ∗∗∗∗∗ adjusted $p<0.0000001$. ANT, Antarctic; ARC, Arctic; TT-EPI, Temperate and Tropical Epipelagic; TT-MES, Temperate and Tropical Mesopelagic.

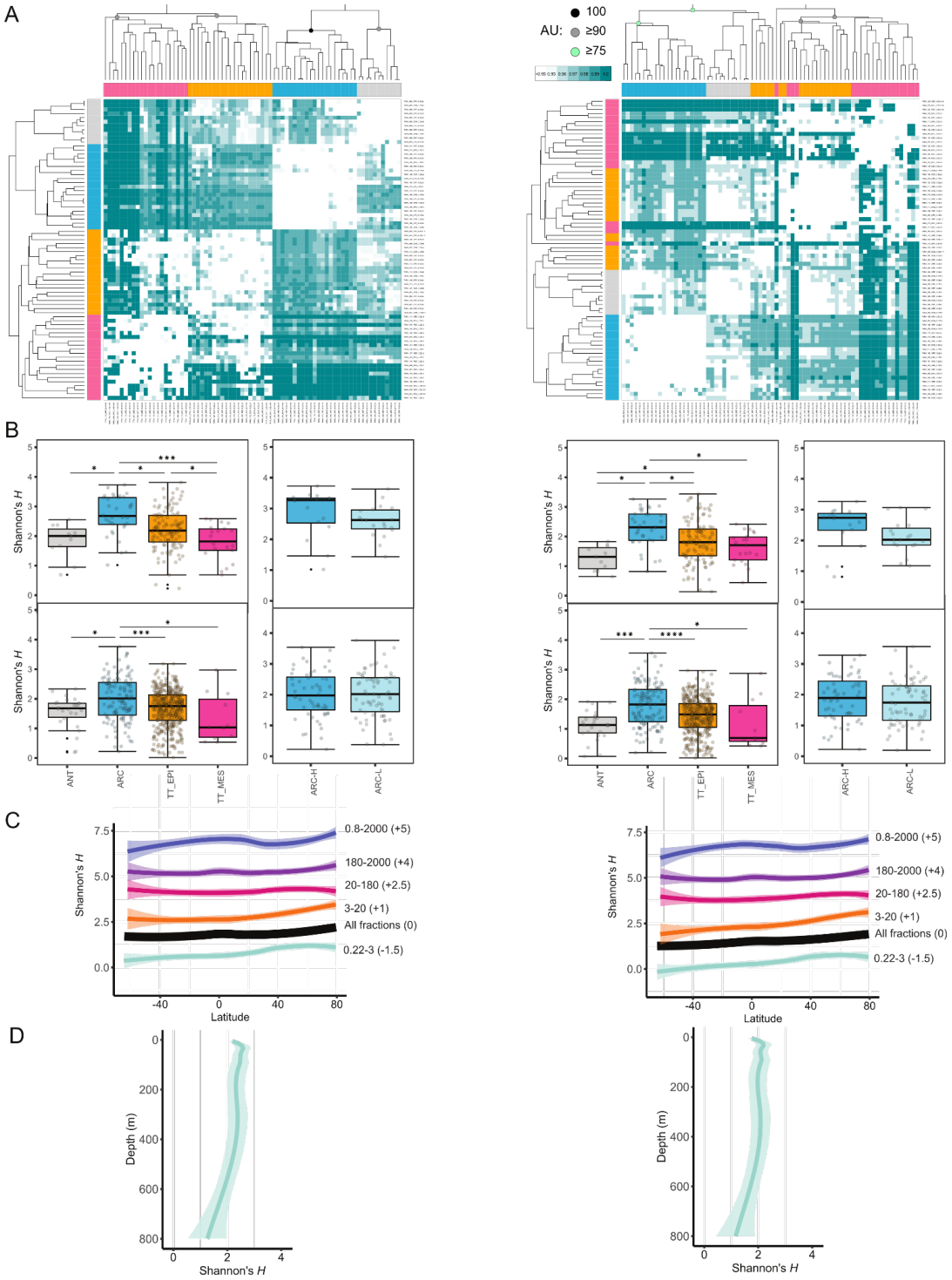

Fig. S4. Sensitivity analyses for the robustness of the ecological inferences under different genome fragment cutoffs (≥2 kb; left, ≥3 kb; right).

(A) same caption as **fig. S2D**. Notice the deterioration in the bootstrap values for the ≥3-kb contigs as a result of reduced statistical power (see more sensitivity analyses on contig length in Zayed et al., 2022). **(B)** same caption as **fig. S3A**. **(C–D)** same captions as **Fig. 4A–B**, without DNA viruses, respectively.

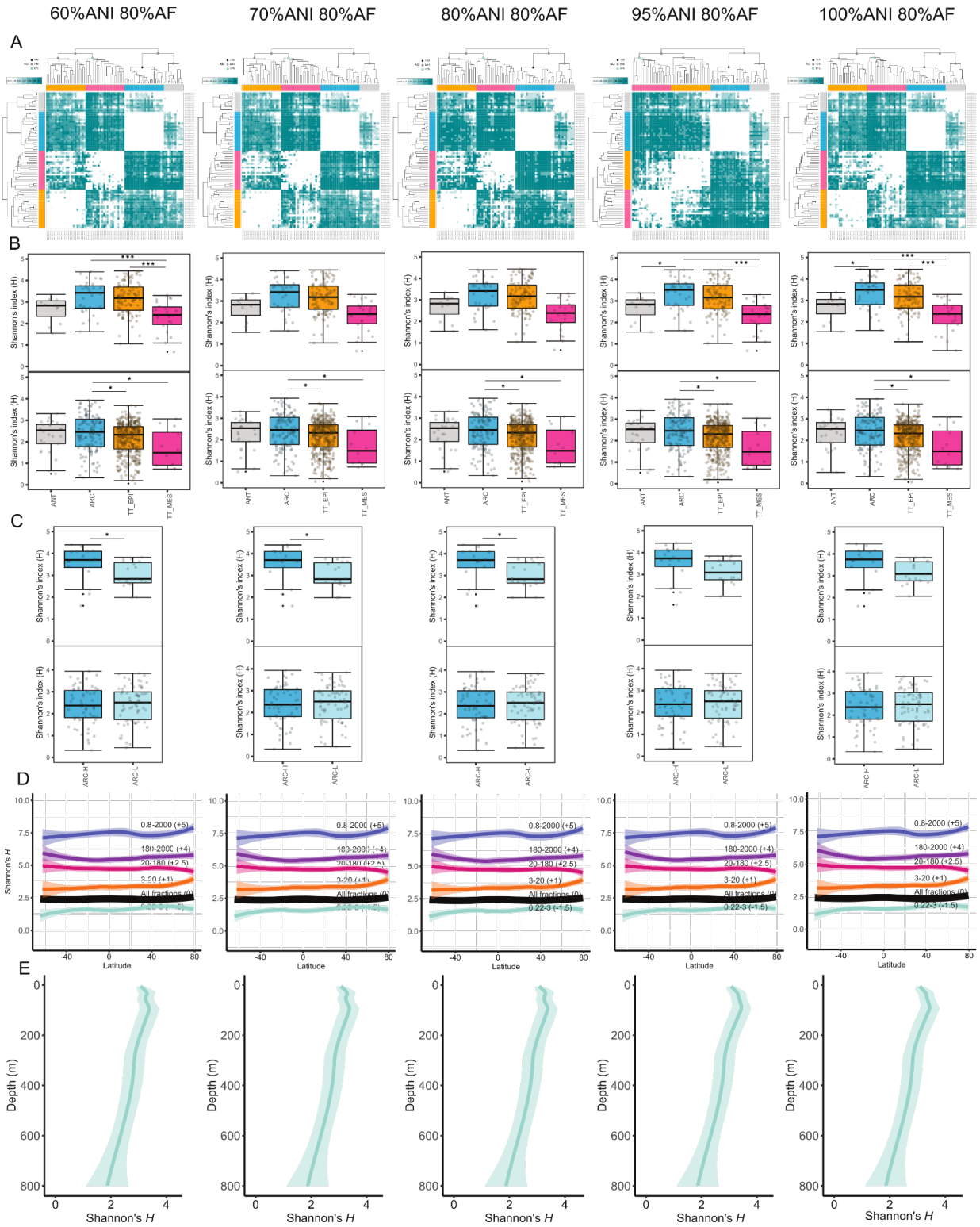

Fig. S5. Sensitivity analyses for the robustness of the ecological inferences under different vOTU definitions (shown at the top of each column).

(A) same caption as **fig. S2D**. **(B–C)** same captions as **fig. S3A**. **(D–E)** same captions as **Fig. 4A–B**, without the DNA viruses, respectively.

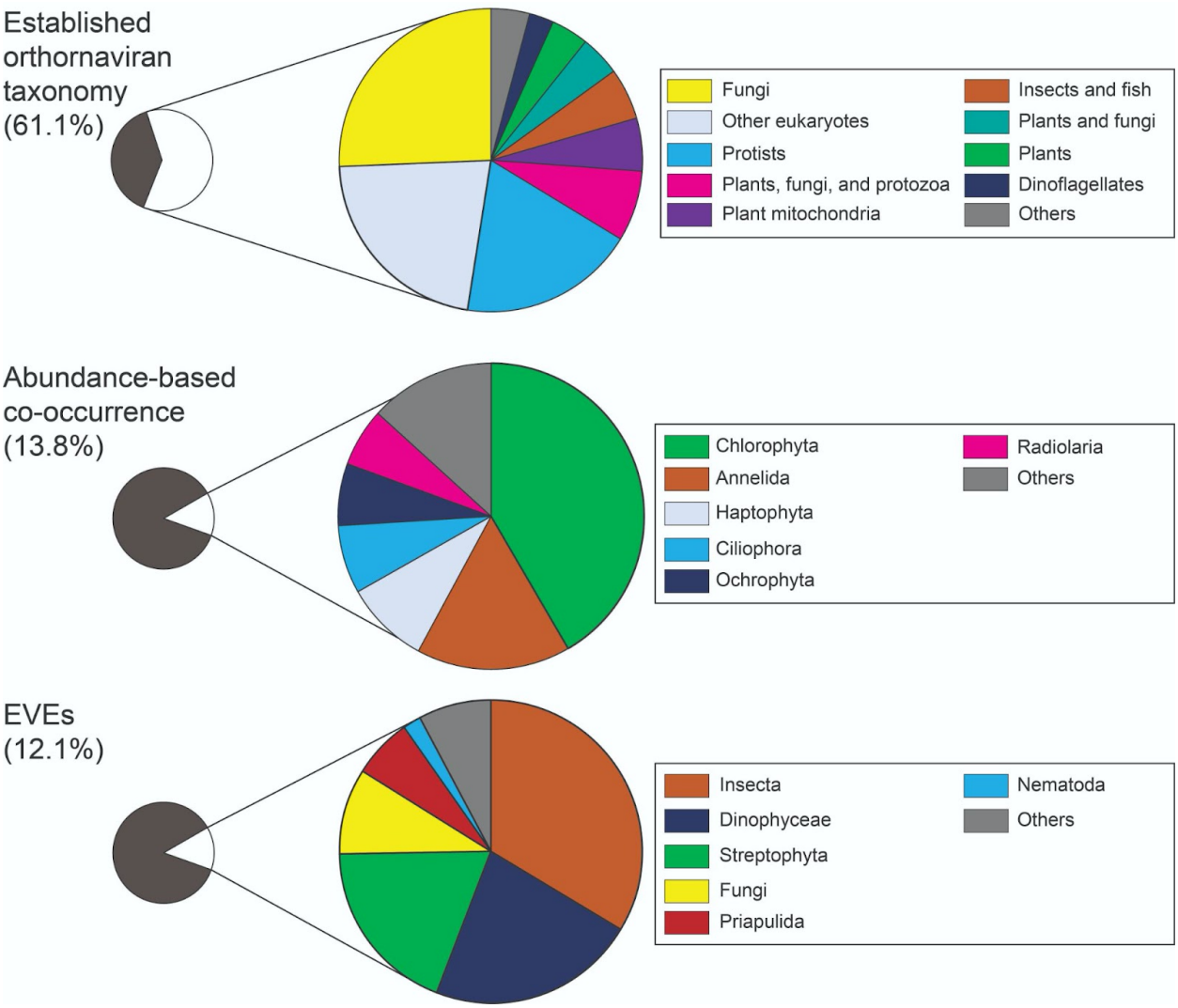

Fig. S6. Inferred hosts for marine RNA viruses.

Left pie charts and parentheses show the percentage of vOTUs for which hosts could or could not be predicted following three different approaches, and right pie charts show the percentage of vOTUs assigned to the most frequent host groups. Virus-host predictions were conducted using three different approaches (see **Materials and Methods**): previous host information following established orthornaviran taxonomy (61.1% of the vOTUs), abundance-based co-occurrence (13.8%), and sequence similarity with endogenous virus elements (EVEs, 12.1%). Note that, following the established virus taxonomy, some host ranges can be very broad (e.g., "Eukaryotes" or "Plants, fungi, and protozoa"). Note that "plants" and "streptophyta" may be interpreted as chlorophytes (green algae), their closest relatives in marine environments.

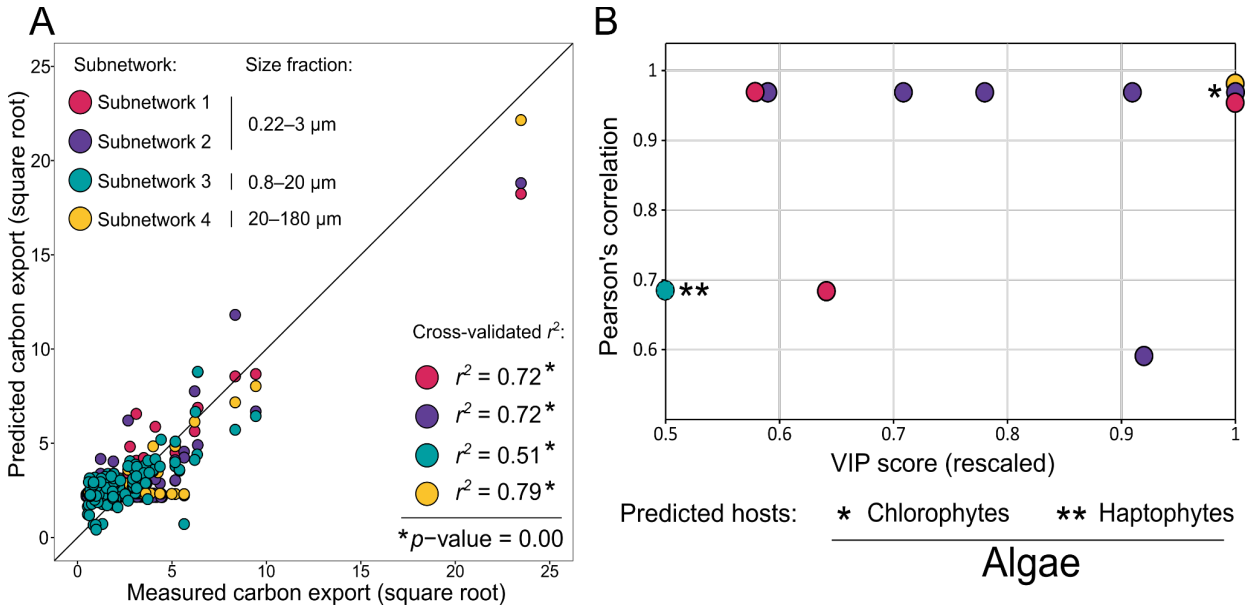

Fig. S7. Marine RNA viruses strongly predict carbon export from pole to pole.

(**A**) Four RNA virus subnetworks (across three different size fractions) from a Weighted Gene Correlation Network Analysis (WGCNA) predict carbon export (y-axis) with the correct magnitude (black line indicates the 1:1 line). Cross-validated partial least square regression (R square values; see Methods) was used to independently predict the values on the y-axis from RNA virus abundances within each of the four subnetworks. (**B**) Scatter plot showing the relationship of the Variable Importance in Projection (VIP) score (relative importance of the vOTU in the subnetwork for carbon export; see Methods) and standard Pearson's coefficient of the vOTU and carbon export. Hosts were only inferred for two vOTUs (both being algae based on the co-occurrence approach). The vOTU with two asterisks fell below 0.5 on the x-axis but had a very high non-scaled VIP score of 9 and hence was included. The x-axis was rescaled for visualization since the VIP score can vary by orders of magnitude.

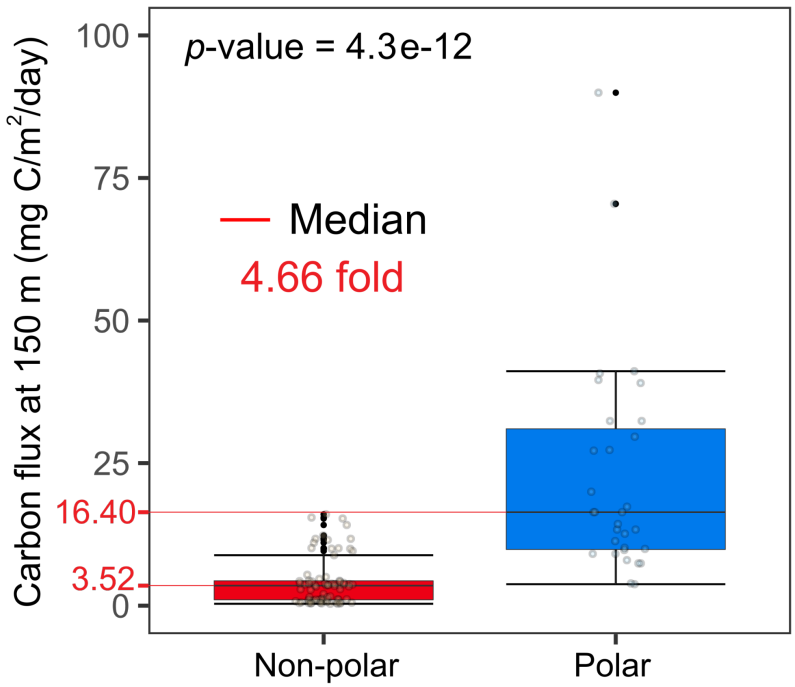

Fig. S8. Carbon export comparison between polar and non-polar waters.

Boxplot analysis (function 'geom_boxplot' of ggplot2 in R) showing the difference in carbon flux (at 150 m) between polar and non-polar waters. The *p*-value was calculated from a Wilcoxon Rank Sum test (function 'wilcox.test' in R) on the medians (red lines).

Tables (provided as a separate file)

Table S1.
Ocean RNA virus studies.

Table S2.
List of RNA and DNA samples, their metadata, and their unique identifiers.

Table S3.
List of additional example literature (section A) and Global Ocean surveys (section B) studying the central ecological roles of marine plankton.

Table S4.
Correlations between environmental variables and alpha diversity for RNA and DNA viruses.

Table S5.
Host prediction results for the RNA vOTUs identified in this study.

Table S6.
AMGs encoded by RNA viruses in this study and previous literature.

Table S7.
RNA vOTU modules (subnetworks) predictive of ocean carbon flux.

Data

Data S1. (separate file)

Gene schemes of vOTU representatives carrying 19 AMG functional types observed either in short-read or long-read sequencing data. RNA virus contig gene schemes for 19 AMG functional types encoded by RNA viruses observed either in short-read or long-read sequencing data. Gene schemes represent the positions of ORFs (white boxes) and functional domains of the virus RdRP (red) and the AMG sequences (blue and green). ORFs were predicted using Prodigal after estimation of the genetic code. Length and ID of the corresponding contigs are indicated. Gene schemes are accompanied by a brief description of the RdRP-based virus taxonomy and possible functions of AMGs. Virus taxonomy and predicted hosts are provided between parentheses in the header. For convenience, host inferences provided are based only on information for viruses of established orthornaviran taxa whenever available.

Data S2. (separate file)

Endogenous virus elements detected in this study. Global alignments of endogenous RNA virus elements (EVEs; derived from metatranscriptomes and labeled in black) with corresponding MAGs (derived from metagenomes, labeled in blue). EVEs were detected by searching against the co-sampled MAGs using highly strict blastn thresholds (see **Materials and Methods**). Protein sequence similarity of RdRPs of eleven of the fourteen EVEs detected suggest they are double-stranded RNA (dsRNA) viruses (likely "partiti-like" and "toti-like" viruses). Their most closely related viruses are indicated in the colored rectangle to the left of the contig ID. We could only annotate two of the seven MAGs (2 and 3) by using blastx (data not shown), which were identified as the chloroplast genomes of the diatoms *Haslea silbo* and *Skeletonema pseudocostatum*, according to their best hits in the NCBI non-redundant database (QUS63753.1 and QGR23538, respectively). End-to-end sequence alignments were performed using Geneious 10.0.9. The nucleotide position is indicated along the top consensus sequence. The nucleotide identity for each given position is represented below the top consensus sequence in a graph panel: green signifies complete identity, yellow is used for less than complete identity, and red refers to very low identity.