MARINE VIROME

Diversity and ecological footprint of Global Ocean RNA viruses

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], Quentin Carradec[5,6], Corinne da Silva[5], 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], Matthew B. Sullivan[1,2,3,4,14]*

DNA viruses are increasingly recognized as influencing marine microbes and microbe-mediated biogeochemical cycling. However, little is known about global marine RNA virus diversity, ecology, and ecosystem roles. In this study, 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 cosampled 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 that RNA viruses cause reprogramming of diverse host metabolisms, including photosynthesis and carbon cycling, and that RNA virus abundances predict ocean carbon export.

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 because 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 is 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 that constitute 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, and materials and methods) to identify and quantify RNA viruses (*19*). This effort led to the identification of 44,779 RNA virus contigs that were dereplicated to 5504 "species"-level virus operational taxonomic units (vOTUs), for which we established taxonomy, evolutionary origins, and biogeography. In this work, 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*), preclustered at 90% average nucleotide identity across 80% of the shorter sequence length and 1-kb minimum contig length (materials and methods), and their relative abundances, estimated by means of metatranscriptomic read mapping (materials and methods), to investigate marine RNA virus ecology globally.

By using a statistical method that nonlinearly deconvolutes high-dimensional data into two-dimensional space (Fig. 1B; *t*-distributed stochastic neighbor embedding, fig. S2, A to 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. This classification into only four ecological zones contrasts with the 56 biogeochemical provinces that are 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

[1]Department of Microbiology, The Ohio State University, Columbus, OH 43210, USA. [2]EMERGE Biology Integration Institute, The Ohio State University, Columbus, OH 43210, USA. [3]Center of Microbiome Science, The Ohio State University, Columbus, OH 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, 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, 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, OH 43210, USA.
*Corresponding author. Email: sullivan.948@osu.edu
†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 Materials.

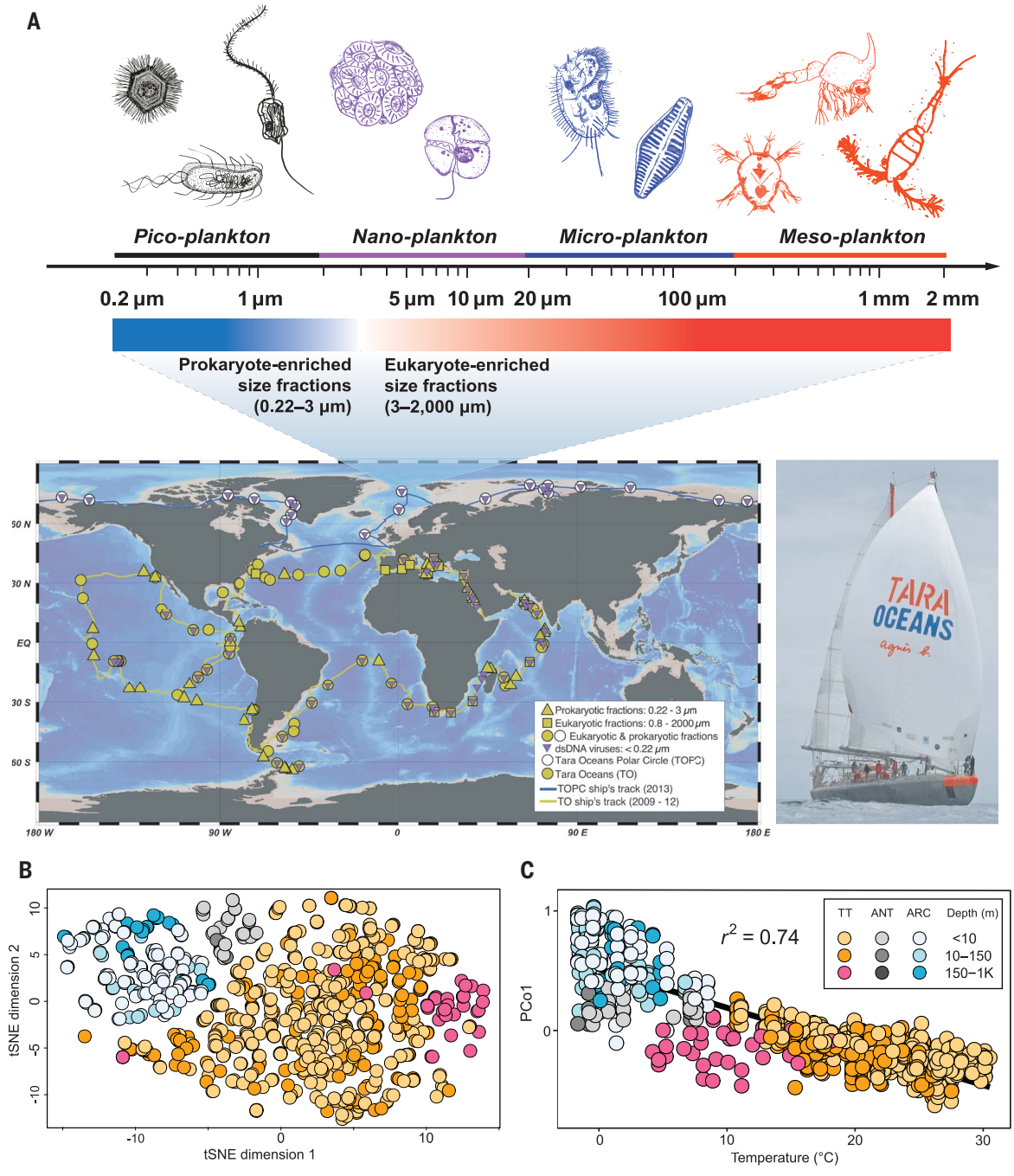

Fig. 1. The cross-domain Global Ocean plankton sampling and resultant RNA virus meta-communities identified from the metatranscriptomes. (**A**) Global Ocean sampling map shows the cruise of the *Tara* Oceans and *Tara* Oceans Polar Circle expeditions and the location of their stations, which are shown 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). 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 that was 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 (C) (also see figs. S4 and S5 for vOTU definition sensitivity analyses). (**C**) Regression analysis of the first coordinate of a principal coordinate analysis (PCo1) of the same Bray-Curtis dissimilarity matrix in (A) (also see fig. S2) and temperature, which shows 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.

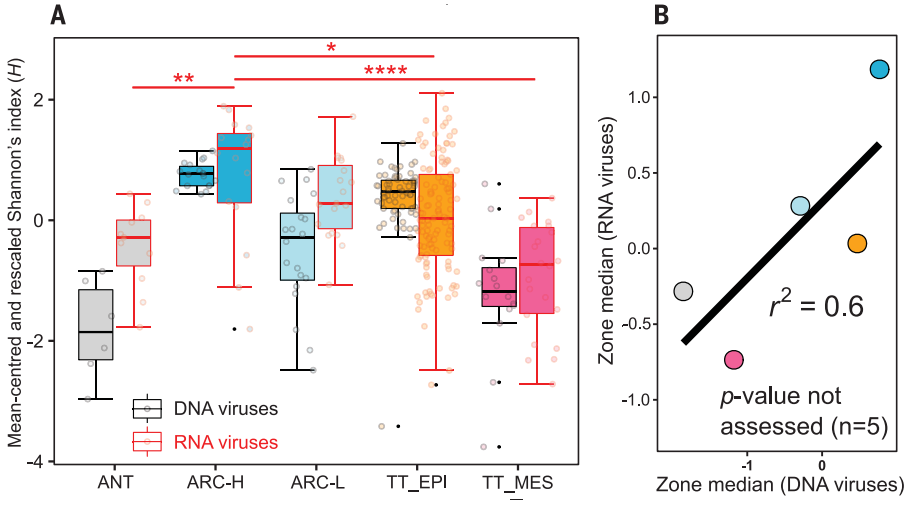

Fig. 2. RNA and DNA virus "species"-level diversity show large-scale congruence. (**A** and **B**) Boxplot (A) and regression (B) analyses of RNA and DNA virus "species"-level diversity across their shared ecological zones. Shannon's *H* values were mean-centered and rescaled across the two virus nucleic 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 and 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.

that were inferred for prokaryotic dsDNA viruses (materials and methods; the fifth Bathypelagic zone that was inferred from dsDNA virus analyses was not sampled here) (*21*) 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 to A14), 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 nonpolar regions (fig. S2, E and 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 ($r \approx 0.25$ 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 and materials and methods).

We found, first, that 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 covaries with dsDNA virus diversity (light green module in Fig. 3). More recently, the relationship between dsDNA viruses and PIC has been posited to be abiotic on the basis of 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), which indicates, 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 that are 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, as also deriving 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 "RNA virus 'species'-level diversity along ecological gradients"). 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 physicochemical tolerances, or ecological gradients, of RNA viruses are not understood. Organismal diversity typically decreases with

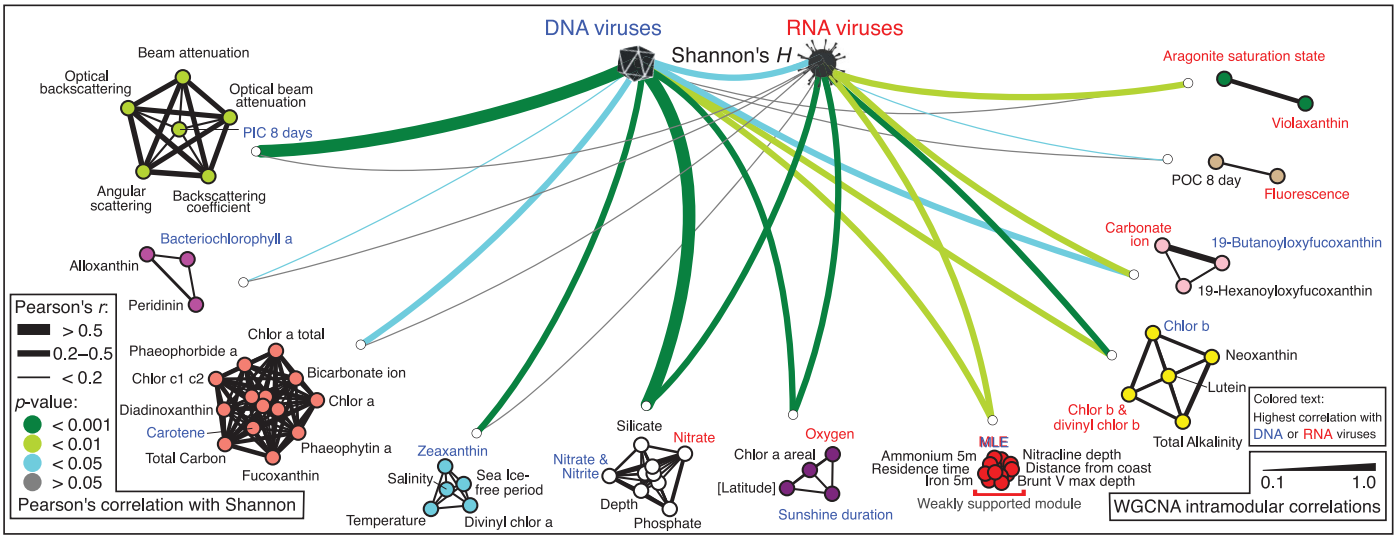

Fig. 3. "Species"-level diversity correlates of marine RNA viruses. Weighted gene correlation network analysis (WGCNA)–supported modules (to account for collinearity) of environmental variables (materials and methods) showing the cofactors of RNA and DNA virus diversity. Modules are Pearson correlated 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. MLE, maximum Lyapunov exponent; POC, particulate organic carbon.

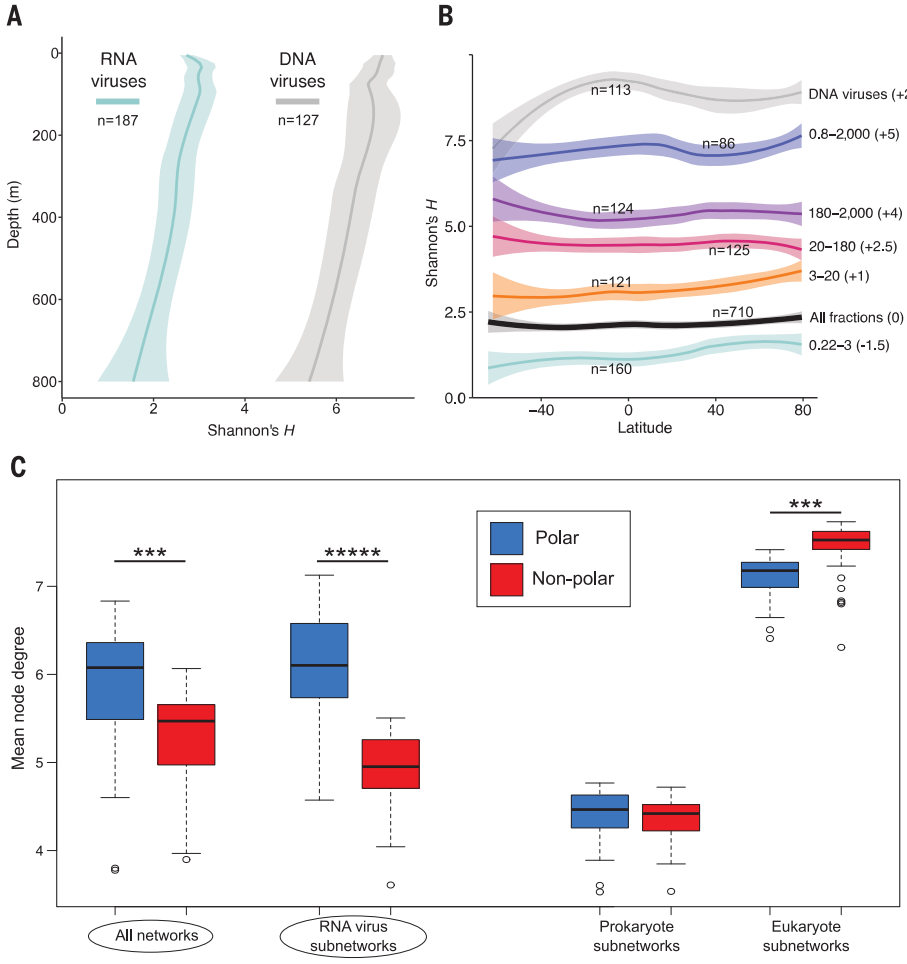

Fig. 4. RNA virus "species"-level diversity across depth and latitudinal gradients. (**A**) Locally estimated scatterplot smoothing (LOESS) smooth plots showing the depth distributions of "species"-level diversity for RNA and DNA viruses. Shown are only RNA viruses from the prokaryotic fraction because of the very limited number of deep ocean samples from the eukaryotic fraction (eukaryotic fraction results are shown in fig. S3). (**B**) LOESS plots showing the latitudinal distributions of "species"-level diversity for DNA (gray) 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 figs. S4 and S5 for vOTU definition sensitivity analyses). (**C**) Global and organismal domain-specific co-occurrence networks connectivity (mean node degree) in polar versus nonpolar samples showing that the significantly higher connectivity in the polar waters (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 $P < 0.0000001$.

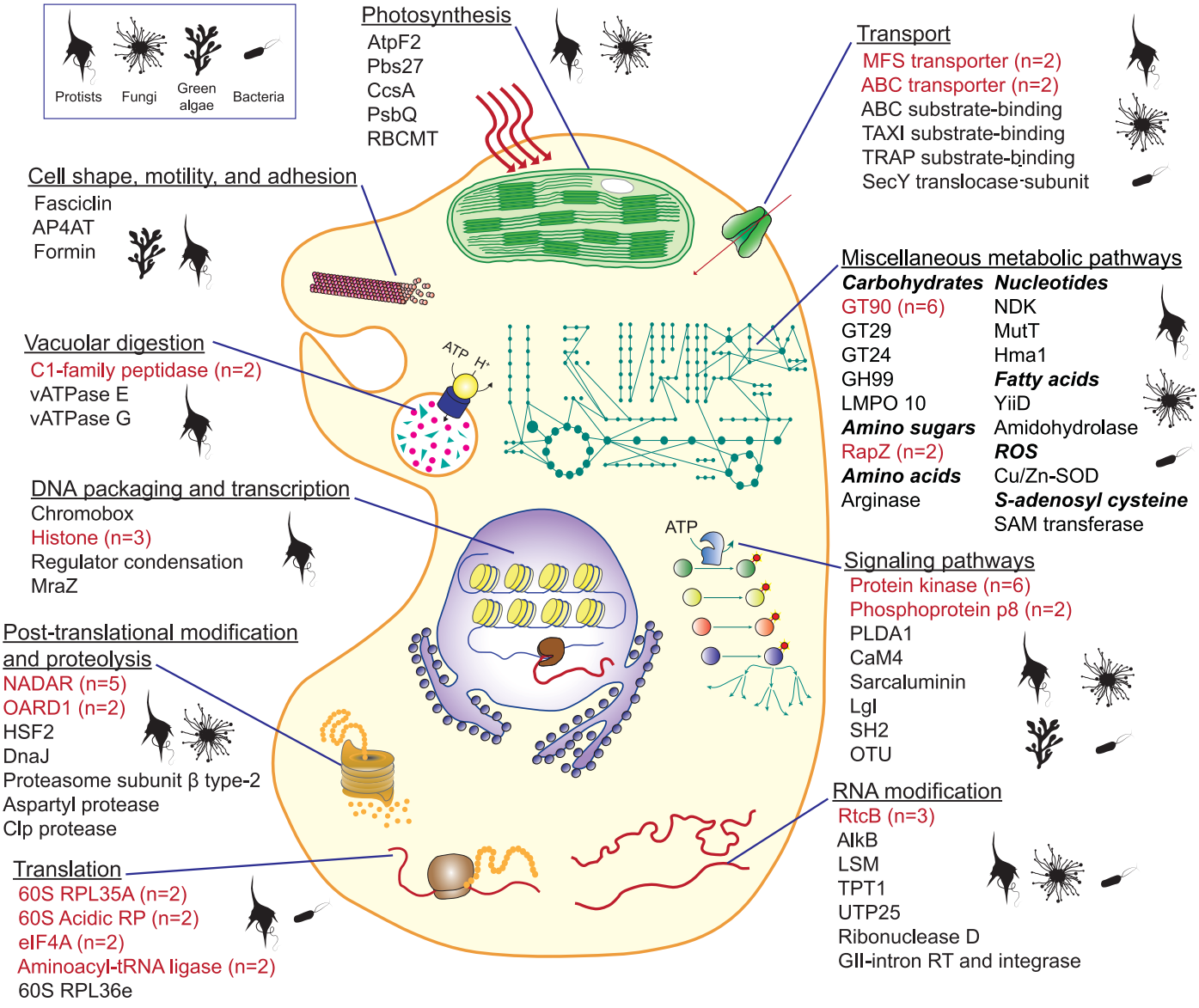

Fig. 5. Functional diversity of AMGs carried by marine RNA viruses. Schematic representation of the hypothesized roles played in manipulation of host metabolism by RNA virus AMGs, which are separated according to functional categories. Red text corresponds to proteins that were found encoded independently in several vOTUs with the number of vOTUs listed in parentheses. The putative hosts, which were inferred by 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). ABC, ATP-binding cassette; TRAP, tripartite ATP-independent periplasmic.

depth (*30*), as does dsDNA virus diversity (*21*), and we found that 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 and fig. S3; also see figs. S4 and S5 for other sensitivity analyses), which is reminiscent of the deviation seen for dsDNA viruses (*21*). This disconnect of virus and host diversity also has a precedent among nonviruses [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 that integrated RNA viruses, prokaryotes, and eukaryotes (materials and methods) to predict hosts for these RNA viruses [sensu (*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 nonpolar samples. This comparison showed significantly more connections in polar samples than nonpolar 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 nonlinear 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. 1, B and 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, as 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 of these hosts, 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 scales (table S1) that indicate that RNA viruses are central ecological players in the oceans. These findings also indicate that, although prokaryotic cells outnumber eukaryotic organisms in the oceans, few RNA viruses (only 3.4% of the vOTUs) infect bacteria, a result that is consistent with previous marine virome and virus isolate reports (*7*).

Second, ecosystem impact might be inferred from the "cellular" protein sequences that 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 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: (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 AMGs were present in any of the 14 virus contigs that were putatively derived from EVEs (data S2 and materials and methods). Mechanistically, we presume such AMGs were acquired by RNA virus genomes through copy-choice recombination with cellular RNAs, as was 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 that 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 and data S1). The most common functional type of AMG (15.8%) was involved in RNA modifications (RtcB, AlkB, and RNA 2′-phosphotransferase) and posttranslational modifications (NADAR and OARD1), which may reflect the common need of viruses to evade host antiviral responses through the repair of virus RNAs and proteins (*40*, *41*). Given that viruses must reprogram cells toward virus progeny production and that RNA viruses have relatively short genomes, it was not surprising to see that protein kinases were abundant (14.8%), as they would allow broad reprogramming capability through limited genetic capacity. The frequency of AMGs suggested that a suite of other processes are affected 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 ultralimiting nutrient conditions of seawater.

Recent experimental work has emerged to assess how DNA viruses affect 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 (materials and methods) (*10*) to evaluate in silico whether RNA viruses might affect 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 and table S7). Specifically, from 5504 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 modeling and leave-one-out cross-validation techniques (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 and 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 to A19).

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 (e.g.,

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: (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 (figs. S4 and S5); (ii) our analyses were composed purely of RNA viruses because of 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 *et al*., *Nature* **588**, 95–100 (2020).
2. P. G. Falkowski, T. Fenchel, E. F. Delong, *Science* **320**, 1034–1039 (2008).
3. C. B. Field, M. J. Behrenfeld, J. T. Randerson, P. Falkowski, *Science* **281**, 237–240 (1998).
4. P. W. Boyd, H. Claustre, M. Levy, D. A. Siegel, T. Weber, *Nature* **568**, 327–335 (2019).
5. A. Z. Worden *et al*., *Science* **347**, 1257594 (2015).
6. C. A. Suttle, *Nat. Rev. Microbiol.* **5**, 801–812 (2007).
7. A. Culley, *Virus Res.* **244**, 84–89 (2018).
8. G. F. Steward *et al*., *ISME J.* **7**, 672–679 (2013).
9. J. A. Miranda, A. I. Culley, C. R. Schvarcz, G. F. Steward, *Environ. Microbiol.* **18**, 3714–3727 (2016).
10. L. Guidi *et al*., *Nature* **532**, 465–470 (2016).
11. M. Moniruzzaman *et al*., *Nat. Commun.* **8**, 16054 (2017).
12. Y. Tomaru *et al*., *Environ. Microbiol.* **9**, 1376–1383 (2007).
13. L. Zeigler Allen *et al*., *mSystems* **2**, e00125–e16 (2017).
14. J. A. Gustavsen, D. M. Winget, X. Tian, C. A. Suttle, *Front. Microbiol.* **5**, 703 (2014).
15. E. P. Starr, E. E. Nuccio, J. Pett-Ridge, J. F. Banfield, M. K. Firestone, *Proc. Natl. Acad. Sci. U.S.A.* **116**, 25900–25908 (2019).
16. M. Shi *et al*., *Nature* **540**, 539–543 (2016).
17. Y. I. Wolf *et al*., *mBio* **9**, e02329–e18 (2018).
18. C.-X. Li *et al*., *eLife* **4**, e05378 (2015).
19. A. A. Zayed *et al*., *Science* **376**, 156–162 (2022).
20. A. Longhurst, S. Sathyendranath, T. Platt, C. Caverhill, *J. Plankton Res.* **17**, 1245–1271 (1995).
21. A. C. Gregory *et al*., *Cell* **177**, 1109–1123.e14 (2019).
22. G. Salazar *et al*., *Cell* **179**, 1068–1083.e21 (2019).
23. K. Bandara, Ø. Varpe, L. Wijewardene, V. Tverberg, K. Eiane, *Biol. Rev. Camb. Philos. Soc.* **96**, 1547–1589 (2021).
24. G. Sommeria-Klein *et al*., *Science* **374**, 594–599 (2021).
25. S. Chaffron *et al*., *Sci. Adv.* **7**, eabg1921 (2021).
26. M. Słowakiewicz *et al*., *Geochim. Cosmochim. Acta* **292**, 482–498 (2021).
27. M. Sadeghi, Y. Tomaru, T. Ahola, *Viruses* **13**, 362 (2021).
28. K. F. Edwards, G. F. Steward, C. R. Schvarcz, *Ecol. Lett.* **24**, 363–373 (2021).
29. S. Sunagawa *et al*., *Science* **348**, 1261359 (2015).
30. M. J. Costello, C. Chaudhary, *Curr. Biol.* **27**, R511–R527 (2017).
31. D. Righetti, M. Vogt, N. Gruber, A. Psomas, N. E. Zimmermann, *Sci. Adv.* **5**, eaau6253 (2019).
32. M. R. Willig, D. M. Kaufman, R. D. Stevens, *Annu. Rev. Ecol. Evol. Syst.* **34**, 273–309 (2003).
33. F. M. Ibarbalz *et al*., *Cell* **179**, 1084–1097.e21 (2019).
34. J. B. Emerson, B. C. Thomas, K. Andrade, K. B. Heidelberg, J. F. Banfield, *Appl. Environ. Microbiol.* **79**, 6755–6764 (2013).
35. A. C. Gregory *et al*., *Cell Host Microbe* **28**, 724–740.e8 (2020).
36. E. A. Gould, *Mol. Biotechnol.* **13**, 57–66 (1999).
37. C. J. Carlson, C. M. Zipfel, R. Garnier, S. Bansal, *Nat. Ecol. Evol.* **3**, 1070–1075 (2019).
38. M. Breitbart, L. R. Thompson, C. A. Suttle, M. B. Sullivan, *Oceanography (Wash. D.C.)* **20**, 135–139 (2007).
39. G. Meyers, T. Rümenapf, H.-J. Thiel, *Nature* **341**, 491–491 (1989).
40. A. M. Burroughs, L. Aravind, *Nucleic Acids Res.* **44**, 8525–8555 (2016).
41. A. R. Fehr, G. Jankevicius, I. Ahel, S. Perlman, *Trends Microbiol.* **26**, 598–610 (2018).
42. F. Vincent, U. Sheyn, Z. Porat, D. Schatz, A. Vardi, *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2021586118 (2021).
43. T. E. G. Biggs, J. Huisman, C. P. D. Brussaard, *ISME J.* **15**, 3615–3622 (2021).
44. H. Kaneko *et al*., *iScience* **24**, 102002 (2020).
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).

ACKNOWLEDGMENTS

We thank A. 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* Oceans Foundation and its sponsors, 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: US 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 to M.B.S.; Ramon-Areces Foundation Postdoctoral Fellowship to G.D.-H.; and France Génomique (ANR-10-INBS-09) to P.W. This work was supported in part through Laulima Government Solutions, LLC, prime contract with the US National Institute of Allergy and Infectious Diseases (NIAID) under contract no. HHSN272201800013C. 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. **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:** The authors declare that they have no competing interests. **Data and materials availability:** The authors declare that all data reported here 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 no. 133 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

science.org/doi/10.1126/science.abn6358
Materials and Methods
Supplementary Text
Figs. S1 to S8
Tables S1 to S7
References (*46*–*124*)
MDAR Reproducibility Checklist
Data S1 to S2

View/request a protocol for this paper from *Bio-protocol*.

Submitted 20 December 2021; accepted 4 May 2022
10.1126/science.abn6358