



Supplementary Materials for

Cryptic and abundant marine viruses at the evolutionary origins of Earth's RNA virome

Ahmed A. Zayed *et al.*

Corresponding author: Matthew B. Sullivan, sullivan.948@osu.edu

Science **376**, 156 (2022)
DOI: 10.1126/science.abm5847

The PDF file includes:

Materials and Methods
Supplementary Text
Figs. S1 to S11
References

Other Supplementary Material for this manuscript includes the following:

Tables S1 to S12
Data S1 to S4
MDAR Reproducibility Checklist

Materials and Methods

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

The 771 ocean metatranscriptomes (≈ 28 terabases [Tb] of data) used in this study were collected during the *Tara Oceans* (TO) and *Tara Oceans Polar Circle* (TOPC) expeditions (2009–2013) from 121 sampling sites across all major oceanic provinces (**table S4**). Of these 771 metatranscriptomes, 187 prokaryotic-fraction (see definition under **fig. S1**) metatranscriptomes from the TO and TOPC expeditions (≈ 5.3 Tb of data) were previously published (Salazar et al., 2019), and 441 eukaryotic-fraction (see definition under **fig. S1**) metatranscriptomes from the TO expedition (≈ 16.3 Tb of data) were previously published (Carradec et al., 2018). The remaining 143 metatranscriptomes were newly sequenced here (≈ 6.3 Tb of data), and all represent eukaryotic-fraction metatranscriptomes from the TOPC expedition (BioProjects PRJEB9738 and PRJEB9739). A detailed description of ocean sampling strategies and protocols for the TO and TOPC campaigns was previously published (Pesant et al., 2015).

Optimized protocols for extraction and purification of nucleic acids were previously described (Alberti et al., 2017). Briefly, DNA and RNA from eukaryotic fractions were purified using the NucleoSpin RNA extraction kit (Macherey-Nagel, Düren, Germany) combined with DNA Elution buffer kit (Macherey-Nagel), whereas different protocol modifications of RNeasy Mini Kit (Qiagen) or Nucleospin RNA kit were used for the prokaryotic fractions. Although we acknowledge that some protistan taxonomic groups (such as, diatoms and dinoflagellates) could be particularly recalcitrant to lysis during extraction of nucleic acids (and hence some biases are expected), the protocols were originally optimized against the Roscoff marine culture collection (<https://roscoff-culture-collection.org/>), and there is evidence of extraction of nucleic acids of such recalcitrant protists using the methods applied here (Carradec et al., 2018).

As described previously (Alberti et al., 2017), post-extraction nucleic acids were treated with DNase to remove contaminant DNA, followed by library preparation. For a maximum of 4 μg of total RNA extracted from the prokaryote-enriched fractions, ribosomal RNA (rRNA) was depleted using the Ribo-Zero Magnetic Kit (Bacteria) (Epicentre Biotechnologies), and cDNA synthesis was performed using the SMARTer Stranded RNA-Seq Kit (Clontech, Mountain View, CA, USA), which enables retention of strand information for each RNA molecule. For eukaryote-enriched samples, cDNA synthesis was performed with different kits according to the quantity of RNA and/or sample processing timing but always depleting the rRNA by selection of poly(A)-tailed RNA molecules. First, for satisfactory total RNA amounts from eukaryote-enriched fractions derived from non-Arctic samples ($\geq 2 \mu\text{g}$), poly(A)-tailed RNA was captured with oligo(dT) beads, chemically fragmented, and converted into single-stranded cDNA using random hexamer priming. Using the TruSeq mRNA Sample preparation kit (Illumina, San Diego, CA, USA), second-strand cDNA was generated by random priming, leading therefore to non-stranded libraries. Second, satisfactory total RNA amounts from eukaryote-enriched fractions of Arctic samples ($\geq 1 \mu\text{g}$) were processed with a new version of the kit, the TruSeq Stranded mRNA kit, that allows strand specificity by incorporating dUTP during the second-strand cDNA synthesis. Finally, libraries for unsatisfactory total RNA amounts (between 50 ng and 2 μg in TO, and between 50 ng and 1 μg in TOPC) of eukaryote-enriched samples were built using the kit SMARTer Ultra Low RNA kit, in which poly(A)-tailed RNA was selected and reverse-transcribed by using oligo(dT) primers; the second-strand cDNA is synthesized using the SMART template switching technology. All DNA and RNA libraries were profiled using a 2100 Bioanalyzer System (Agilent Technologies) and

qPCR (MxPro, Agilent Technologies), and then sequenced with 101 base-length read chemistry in a paired-end flow cell on HiSeq2000 or HiSeq2500 sequencing machines (Illumina).

Metatranscriptome assembly, long-read processing, and estimation of genome completeness

The bioinformatic workflow is represented schematically in (**fig. S2**). Fastq raw reads from previously published metatranscriptomes were downloaded from the NCBI Sequence Read Archive (SRA) database (<https://www.ncbi.nlm.nih.gov/sra>), separated into forward and reverse reads, and trimmed for read quality using Trimmomatic v0.36 (Bolger et al., 2014) using default parameters. No filtering of low complexity reads or host reads was performed before assemblies. Reads were assembled *de novo* into contigs using MEGAHIT v1.1.3 (Li et al., 2015) with default parameter settings. MEGAHIT was chosen as an assembler as it has been successfully used for the assembly of metatranscriptomes (Nowinski et al., 2019; Crump et al., 2018; Chung et al., 2020) and RNA viruses (Li et al., 2020; Islam et al., 2021; Yuan et al., 2020) with comparable assembly performance to metaSPAdes (Islam et al., 2021; van der Walt et al., 2017), but with more efficient memory and time use than metaSPAdes. Genes were predicted from contigs using Prodigal v2.6.3 (Hyatt et al., 2010) using the default translation table with the “-p meta” option enabled.

A subset of 20 RNA samples utilized here for standard Illumina short-read sequencing were also used for long-read sequencing. Briefly, RNA was reverse-transcribed using SMARTer RNA library prep kit and ligated to adapters with the SQK-LSK109 kit before sequencing with MinION (Oxford Nanopore Technologies). Base-calling was performed using guppy 4.0.14 high accuracy. First, virus contigs were searched for closely related unassembled long-read sequences (90% nucleotide identity across 80% aligned fraction) by using blastn. The captured long reads were corrected by using standard Illumina short-read sequencing data with Pilon version 1.23 (Walker et al., 2014). Briefly, this entailed mapping the short-reads to the individual long-reads with bwa (version 0.7.17), and the resulting bam file served as input for Pilon. Pilon uses short-read mapping information to identify errors in the long-reads, and corrects for insertions, deletions and individual mismatches along the long reads that may occur. After correction of the long-read sequences, virus contigs were searched again for close relatives (as described above) and alignments for which the long read was shorter than the virus contig were excluded.

To estimate genome completeness for the viral contigs recovered in our study, CheckV (Nayfach et al., 2020) was used with its default parameters.

RNA virus identification

From the resultant metatranscriptomics assemblies, virus contigs were identified based on homology searches using profile hidden Markov model approaches (HMMs) of virus RNA-directed RNA polymerase (RdRp) domains (doi:10.5281/zenodo.5731488). Previous studies have also used HMMs for RNA virus identification (e.g., Callanan et al. 2021; Starr et al. 2019). We did not screen for riboviruses of kingdom *Pararnavirae*, whose hallmark gene is reverse transcriptase, and which is easily confounded with those from endogenous retroviruses and retro-elements. Realm *Ribozyviria* was also excluded because it only encompasses a few dozen viruses related to hepatitis D virus 1, none of which encode RdRps. To increase detection of divergent RdRp domain sequences (Te Velhuis, 2014), profile HMMs were generated and updated over ten iterations by recruiting newly detected sequences from our study (**fig. S2F**) as described previously (Callanan et al., 2020; Sunagawa et al., 2009). Alignments of RdRp domain sequences from the 57 RNA virus lineages defined before (approximately between family and genus ranks) (Wolf et

al., 2018), five recently established families (*Cretegaviridae*, *Gresnaviridae*, *Nanghoshaviridae*, *Nanhypoviridae*, *Olifoviridae*), two proposed families [“*Fusagraviridae*” (Wang et al., 2016), “*Megatotiviridae*” (Arjona-Lopez et al., 2018)] and a novel taxonomic group [“quenyaviruses” (Obbard et al., 2020)], were used to generate the initial profile HMMs. For every iteration, previously known RdRp protein sequences from NCBI GenBank (release 233) and various other sources such as recent RNA virus metatranscriptomic studies from invertebrates (Shi et al., 2016; Li et al., 2015), vertebrates (Shi et al., 2018), marine sponge (Urayama et al., 2020), marine plankton (Urayama et al., 2018; Moniruzzaman et al., 2017), the marine microbial eukaryote transcriptome sequencing project (MMETSP) (Keeling et al., 2014), grassland soil (Starr et al., 2019), and RNA phages (Callanan et al., 2020; Krishnamurthy et al., 2016), along with all the protein sequences predicted from the metatranscriptome samples, were searched against the profile HMMs. The portion aligning to the HMM was trimmed by using HMMsearch (HMMER 3.1) (Eddy, 1998) with the flag -A for the hits with a bit score ≥ 30 . Only sequences longer than 70% of the average length of the best-matching profile HMM were recruited and clustered to generate new HMMs by using the vFam pipeline (Skewes-Cox et al., 2014) with default parameters. The original HMM length was kept after each iteration to calculate the length fraction of the footprint. In total, we identified 48,450 putative virus RdRp hits that were subjected to further investigation (see next section).

Evaluation of authenticity and completeness of putative virus RdRps

To evaluate the authenticity of the 48,450 putative virus RdRp hits, we identified false-positive virus RdRp hits by a competitive HMM search approach and further manual inspection of the HMM-protein alignments for low-scoring, true-positive virus RdRp hits as follows. The 48,450 putative matches were assessed to be *bona fide* virus RdRp sequences using competitive searches with HMMsearch of 84 RdRp and 16,268 PfamA (v33.0) non-RdRp profile HMMs (Wolf et al., 2018). Hits longer than 100 amino acids with a best match to an RdRp HMM and with a bitscore ≥ 30 were kept as true positives for proteins containing the virus RdRp domain. Lower-scoring hits were manually inspected for presence of the seven canonical RdRp domain motifs. In total, 44,779 contigs encoding putative virus RdRps were detected by these identification and curation processes. Notably, none of the false positive virus RdRp hits of at least 100 amino acids long were to the profile of eukaryotic RdRps involved in RNA interference (pfam:PF05183). This complete lack of false positive hits to eukaryotic RdRps indicate that virus RdRps are quite divergent from such eukaryotic RdRps, corroborating the notion that eukaryotic and virus RdRps do not share a common ancestor (Burroughs et al., 2014; Zong et al., 2009). Among the 86 false-positive virus RdRp hits that were identified as cellular protein domains using Pfam HMM profiles, we found 63 reverse transcriptases (RVT_1), 17 P-loop ATPase proteins (ATP_bind_2), four delta carbonic anhydrase (CA_like), a cellulase (Cellulase), and a C-terminal of rhodanese (Rhodanese_C). We applied two additional iterations of the same search-and-updated pipeline with all RdRp profiles available from a coastal RNA virome recently generated (Wolf et al., 2020), in order to extend our power of detection of highly divergent RdRp domain sequences across the Global Ocean metatranscriptomes. After an equivalent process of removal of false positives, we found 49 additional contigs encoding putative virus RdRps, bringing the total to 44,828 contigs. The corresponding RdRp protein sequences for these 49 additional contigs are named using the label “Additional_Tara”, and their clustering revealed the orthornaviran class 13 (**Data S3**).

To estimate RdRp domain completeness, protein sequences translated from putative virus contigs were generated by using transeq (EMBOSS version 6.6.0.0) (Rice et al., 2000) with six

possible frames and the standard translational code to resolve difficulties associated with alternative genetic code usage, non-canonical translation events, and divergent RdRp domain sequences poorly aligning to the profile HMMs. Translated sequences were searched once more against the 84 RdRp profile HMMs, and those with a bit score ≥ 30 and with aligning regions of $\geq 90\%$ of the average length of the best-matching HMMs were considered proteins containing “complete” virus RdRp domain sequences. Sequences containing more than one hit in the same or different frames (presumably non-canonical translation cases) were resolved manually by joining the aligning regions into the same protein sequence.

RdRp-based taxonomic annotation of RNA viruses

To globally assess the orthornaviran taxonomy, we established and evaluated a network-based analytic against phylogenetic methods. Briefly, all available orthornaviran RdRp sequences were collected and compared against those from our study. In total, there were 209,588 RdRp domain amino-acid sequences (111,742 “complete” and 97,846 “partial”) that were derived from our study (n=44,828, of which 6,686 were “complete”), from GenBank release 233 (n=160,167, of which 101,819 were “complete”), and from recently published coastal ocean viromes (Wolf et al., 2018) (n=4,593, of which 3,255 were “complete”). These sequences were first pre-clustered at 50% amino-acid identity using Uclust v10.0.240 (Edgar, 2010) picking the centroid sequence based on length (usearch --cluster_fast -id 0.50 -sort length). Centroids of the resultant 13,109 clusters (n=7,335, 4,440, and 2,236 from our study, GenBank release 233, and the coastal ocean virome, respectively) were then extracted and filtered for domain completeness, including only those considered “complete” (n=6,238). Pairwise comparisons of these “complete” centroid sequences were then conducted using blastp v2.10.0+ (Altschul et al., 1990) after reducing the gap penalty (-gapopen 9 -gapextend 1 -word_size 3 -threshold 10) to extend the length of the alignment of each pair. E-values for each pair were extracted and negative-log₁₀-transformed in MCL v14-137 (Enright et al., 2002) (--stream-mirror --stream-neg-log10 -stream-tf 'ceil(200)'). Transformed e-values were used in an MCL network for iterative clustering, changing the granularity parameter at each iteration (range 0.1–8). All the cluster sets from MCL (**table S7**) were individually compared to the previously established phylogeny-based taxonomy in GenBank (release 233) (**table S7**) using the ‘adj.rand.index’ function of the package “pdfCluster” in R and against the 2020 taxonomy release of the International Committee on Taxonomy of Viruses (ICTV) (Master Species List [MSL] #36; <https://talk.ictvonline.org/taxonomy/>). Network-based cluster sets that gave rise to the highest agreement with the phylogeny-based taxonomy at the phylum and class ranks were picked (as described in **Fig. 1** and **fig. S3**) and the taxonomic delineation was extended from the reference sequences within each cluster. The domain sequences were manually inspected to ensure that predicted novel RNA virus phyla and classes derived from divergent RdRps are not false positives. Specifically, the seven canonical motifs (A–G, though motif E is missing in some *bona fide* RNA viruses) Te Velthuis et al., 2014) of virus RdRps were screened by searching for conserved regions in the consensus sequence of global alignments, and motif identity was confirmed based on HHPred homology searches and available literature.

To generate phylogenetic trees for each resultant network-derived major cluster, sequences from each of these clusters were aligned separately using the E-INS-i strategy over 1,000 iterations in MAFFT v7.017 (Katoh and Standley, 2013). Aligned sequences were subsequently trimmed using Trimal (Capella-Gutierrez et al., 2009) with sites having more than 20% gaps removed and the alignments manually inspected. Prior to phylogenetic analysis, sequences were screened for possible recombination events using 3Seq (Boni et al., 2007), with a recombinant event determined

by a Bonferroni-corrected p -value cutoff of 0.05. Of the 6,238 RdRp centroids that were considered “complete”, only 78 were considered as recombinant sequences and were excluded from phylogenetic analyses. Phylogenetic relationships of sequences within a cluster were first assigned the appropriate evolutionary model using ModelFinder (Kalyaanamoorthy et al., 2017). Then, a subsequent Maximum Likelihood phylogenetic tree was generated using bootstrap support generated for 1,000 iterations in IQ-TREE (Nguyen et al., 2015).

Family rank clades were conservatively assigned by both evaluating all the cluster sets from MCL and the class-specific phylogenetic trees. The different cluster sets were iteratively evaluated to be exclusively composed of reference sequences representing the same virus family (accepting the rare cases of singletons from different taxa) and the putative taxonomic assignment for new sequences was extended from the reference sequences within each cluster to the novel sequences. For phylogenetic trees, taxonomic assignment of sequences was based on the placement of these sequences in the tree, requiring them to fall within a clade to be assigned to the same taxon or to form a sister clade with the tentative name to be identical to the established clade with a “-like” suffix. The most specific taxonomy assignment (e.g., without the “-like” suffix or higher resolution taxonomic assignment) of the two methods was picked as the final putative classification for the novel sequences.

RdRp-based global phylogenetic tree

To generate the global phylum-level phylogenetic tree (**Fig. 3A**), we used an approach that combined consensus [used for highly divergent sequences (Grandi et al., 2020; Grandi et al., 2018; Vargiu et al., 2016; Chen et al., 2015; Fernandez-Caso et al., 2019; Alipour et al., 2013; Zhang and Firestein, 2002)] and individual sequences in the alignment. Each consensus sequence was generated by first aligning individual sequences per megataxon, then obtaining the consensus sequence of the alignment using Geneious v8.1.9 (<https://www.geneious.com>). The number of ambiguous residues (i.e., ‘X’s) within each consensus sequence was then determined and each consensus sequence composed of >20% ambiguous sites was replaced by the individual sequences within the megataxon to preserve the quality of the alignment (Wiens, 2006). Almost all of the new megataxa had >20% ambiguous sites and hence, for consistency, they were all represented by their individual sequences. Subsequent alignment, trimming and phylogenetic inferences were as described above (see “**RdRp-based taxonomic annotation of RNA viruses**”), with the only modification being using the -gappyout option during trimming. The approximate global tree (**fig. S4A**) was visualized from the complete set of previously published 4,617 virus RdRps (Wolf et al., 2018) in iTOL v3 (Letunic and Bork, 2016), collapsing clades into families and orders based on the overwhelming dominance of the family/order-specific lineages within these clades.

3D structure network analysis

To examine the 3D structural similarity between the RdRp domains from the new and previously established megataxa, we first predicted the 3D structures for the new megataxa from their representative primary amino-acid sequences (the longest sequence with no ambiguous residues (i.e., no ‘X’s in the primary sequence) per megataxon) using Phyre2 (Kelley et al., 2015) in the “Normal” mode. The predicted structures were combined with reference (experimentally resolved) RdRp 3D structures from Protein Data Bank for pairwise comparisons (accession numbers are shown in **Fig. 3B**). We also included the reference (experimentally resolved) reverse transcriptase 3D structures of non-LTR retrotransposons and group II introns. Next, pairwise 3D

alignments were performed on the combined dataset (selecting only the RdRp amino-acid chains in the multi-domain reference 3D structures) using Matras (v1.2) (Kawabata, 2003). For each pair, the protein superfamily reliability score was extracted and used to build the 3D structure network using the “Edge-weighted Spring Embedded” method for visualization in cytoscape (Shannon et al., 2003). The 3D structures were visualized using the PyMOL software (<http://www.pymol.org/pymol>).

Suggested names for the novel RNA virus phyla

The largest RNA virus phylum (220 near-complete RdRp domains) described in this work was named “*Taraviricota*” after the Sanskrit word “तारारि [Tārā]”, meaning (i) “a female deity (a female Buddha)” that can take many forms (since the RdRp of “taraviricots” resembles both RdRps and RTs as seen in **Fig. 3B**), (ii) “star” or “planet”, which fits the high abundance of “taraviricots” in the ocean (**Fig. 4**) and terrestrial systems (Obbard et al., 2020), and (iii) “the deity who helps men cross to the other shore”, which fits the wide distribution of “taraviricots” throughout the Global Ocean (**Fig. 4**); and the suffix for phyla, “-viricota”. The root “tara” also refers to the “Tara Oceans expeditions” during which these viruses were discovered and found to be, on average, the most abundant viruses in temperate and tropical waters (**Fig. 4**).

The consensus amino-acid sequence generated from the alignment of the 37 near-complete RdRp domains derived from viruses of “*Pomiviricota*” shared a protein identity of 38% with viruses infecting phytopathogenic fungi (order Erysiphales) causing powdery mildew (Erysiphe necator associated bipartite virus 1 and Podosphaera virus A). Hence, “*Pomiviricota*” is a portmanteau of “powdery mildew viruses” and the suffix for phyla, “-viricota”.

The consensus amino-acid sequence generated from the alignment of the 36 near-complete RdRp domains derived from viruses of “*Arctiviricota*” had no matches in the NCBI non-redundant database. Given the lack of similarity with references and the fact that all sequences (with only one exception) were captured in the Arctic Ocean, we suggest the name “*Arctiviricota*” (a portmanteau of “Arctic Ocean viruses” and the suffix for phyla, “-viricota”) for this potential novel RNA virus phylum. These viruses were also among the most abundant in the Arctic Ocean (**Fig. 4**).

The consensus amino-acid sequence generated from the alignment of the ten near-complete RdRp domains derived from viruses of “*Paraxenoviricota*” shared no protein identity with any subject of the NCBI non-redundant database. Given the lack of similarity with references and no other biological or geographic feature that could be associated with viruses of this group, we suggest the name “*Paraxenoviricota*” (from the Greek “παράξενος [paráxenos]”, meaning “strange” and the suffix for phyla, “-viricota”) for this potential novel RNA virus phylum.

The consensus amino-acid sequence generated from the alignment of the two near-complete RdRp domains derived from viruses of “*Wamoviricota*” had no matches in the NCBI non-redundant database. Given that the only known virus that was binned with these two novel RdRp sequences during MCL clustering was *Phytophthora infestans* RNA virus 2 (*Phytophthora infestans* is an oomycete or water mold), we suggest the name “*Wamoviricota*” (a portmanteau of “water mold viruses” and the suffix for phyla, “-viricota”) for this potential novel RNA virus phylum.

Organization of the RdRp domain sequences in novel megataxa, and assessment of their potential chimeric origin

We sought to further evaluate the authenticity of the novel megatxa viruses by (i) evaluating the organization of their divergent RdRp domain sequences, and (ii) examining their potential origin from chimeric assemblies.

First, we evaluated representative RdRp domain protein sequences of viruses of novel megataxa (phyla and classes) for seven known motifs (Velthuis et al., 2014) expected in RdRp domains (**Data S1–2**). The results suggest that these RdRp domain sequences appear authentic. Specifically, of the normally ordered “G-F-A-B-C-D-E” motifs (Velthuis et al., 2014), sequences assigned to novel phyla contained all seven motifs (one phylum) or six of the seven motifs (“F-A-B-C-D-E”), whereas sequences assigned to novel classes contained a range of motifs from seven (five classes), six (five classes), five (two classes), or four (one class) (**table S10**). Motif A of “*Taraviricota*” is DxxxxE instead of the canonical DxxxxD (**Data S1**). We also found an unusual motif C (IDD) in sequences representing novel negarnaviricot class 67 instead of the canonical (G/S)DD, and a motif order permutation (“C-A-B” instead of the canonical “A-B-C”) in sequences representing novel kitrinoviricot class 42, similar to what is found in viruses of families *Birnaviridae* and *Permutotetraviridae* (associated with phylum *Pisuviricota*; Gorbalenya et al., 2002; Zeddiam et al., 2010), and in specific lineages of the aquatic “Yangshan assemblage” virome (associated with phylum *Kitrinoviricota*; Wolf et al., 2020).

Second, we assessed whether the RNA virus contigs could derive from chimeric assemblies by returning to the prokaryotic or eukaryotic size fraction RNA samples and generating new long-read nanopore sequencing data from complementary DNA (cDNA), derived from 20 samples (see “**Metatranscriptome assembly, long-read processing, and estimation of genome completeness**” above). Although these samples were not RNA-virus-targeted, the long-read data captured a large number of the RNA viruses (n=3,234) we had identified across the dataset. In all cases, the long-read nanopore data confirmed the contigs derived from short-read assemblies (90% nucleotide identity across 80% of the aligned region; **table S5**). This confirmation by long-read data suggests that at least these 3,234 short-read RNA virus contigs are authentic. Additionally, 33,163 of the 44,779 short-read assembled virus contigs were found in more than one assembly (90% nucleotide identity across 80% of the aligned region), which further outrules chimeric contigs (**table S5**) and implies that any possible chimeras in the remaining short-read data would be extremely rare. The 3,234 long-read contigs include some belonging to the novel phyla “*Taraviricota*” (n=32), “*Pomiviricota*” (n=13), and “*Arctiviricota*” (n=5), and include 42 contigs from the novel classes (referred to here with numbers) 38 (n=2), 42 (n=3), 43 (n=14), 48 (n=5), 66 (n=9), and 67 (n=9), with the remaining novel megataxa represented by multiple contigs assembled independently from different samples (**table S5; table S7; fig. S3**).

Establishment of genome-based virus operational taxonomic units and their resemblance to known species

Following the recent consensus on using whole-contig (or -genome) average nucleotide identity (wcANI) for the classification of DNA and RNA viruses at the “species” rank (Roux et al., 2019) and designating them as virus operational taxonomic units (vOTUs), we sought to determine the clustering thresholds that maximize the distance of these vOTUs to represent sequence discrete ecological units for RNA viruses. Briefly, this approach seeks to empirically evaluate whether ‘structure’ emerges from all-versus-all comparisons of sequence similarity

between virus genome pairs, whereby any emergent units represent vOTUs. Pragmatically, vOTUs are approximately species-rank clusters that await whole-genome population genetics analyses to formally evaluate gene flow and selection in the resultant vOTUs. To that end, we conducted pairwise comparisons of the 44,779 virus contigs using MUMmer v3.23 (Delcher et al., 2003), tabulating the average nucleotide identity (ANI) and alignment fraction of the shorter contig (AF) for each pair (excluding self matches). The frequency of the two values across all the contig pairs ≥ 1 kb (the minimum length used to estimate wcANI for fragmented genomes [Roux et al., 2019]) was then computed and visualized in **fig. S9A**. vOTU clustering thresholds were selected to include two different groups of contig pairs with high frequency that mirrored those obtained from complete genomes (Roux et al., 2019), representing sequences with more genetic exchange within their vOTU than with other vOTUs (i.e., resembling a “biological species definition”). Hence, a cutoff of 90% ANI across 80% of the shorter sequence length was used in our study, which resulted in 17,369 total vOTUs, of which 5,504 were ≥ 1 kb. Notably, our analyses suggest needed revision of the empirical cutoffs from those in the prior work for ANI, AF, and wcANI as follows.

First, the consensus statement combined DNA and RNA viruses, which can skew the empirical global cutoffs towards those appropriate for DNA viruses. Indeed, the 95% ANI used in the consensus statement agrees with our previously determined cut-off for marine dsDNA viruses (Deng et al., 2014; Gregory et al., 2016; Gregory et al., 2019). However, we sought to separately evaluate the faster-evolving RNA viruses, which revealed that a more permissive 90% ANI is more appropriate for RNA viruses (i.e., more inclusive of the different genome groups shown in the global similarity analysis; **fig. S9**). *Second*, the consensus statement was understandably limited to reference genomes, which are prone to sampling bias and do not necessarily represent population-level sampling. In contrast, our Global Ocean dataset better evaluates naturally occurring diversity, at least for abundant RNA viruses. *Third*, the consensus statement assessed the impact of genome fragment length on the inferred vOTUs by randomly shearing whole genomes for simulation analyses, with the expectation that the larger the fragment size, the lower the risk to count the same vOTU multiple times upon estimating diversity (this relationship was benchmarked in an earlier study; Roux et al., 2017). In our work, by requiring that each contig carried the RdRp domain, we avoided counting the same vOTU multiple times and removed these issues from our diversity estimations. Hence, a large length cut-off, such as the ≥ 10 kb cut-off recommended for the dsDNA vOTUs (Roux et al., 2017), is not necessarily suitable for RNA viruses. In fact, using a cutoff of ≥ 10 kb or ≥ 5 kb would have removed almost all ($\approx 97.5\%$) or close to half ($\approx 45\%$) of the high- and medium-quality genomes in our dataset, respectively (**fig. S9D**) and missed entire, complete-genome RNA virus megataxa (Wolf et al., 2018). *Fourth*, given that natural samples often require the use of incomplete genome data (or ignoring large swaths of virus genome sequence space), we conducted sensitivity analyses to assess how genome fragment lengths might impact these cutoffs. In these analyses, we evaluated genome fragment lengths of ≥ 2 kb and ≥ 3 kb (anything longer was underpowered due to data sparsity as discussed above), which demonstrated that the cutoffs were robust to changes in fragment lengths (**fig. S9**).

In summary, we re-evaluated RNA virus sequence space for a ‘universal’ cut-off that is suitable across different genome fragment lengths (and hence including genomic information beyond the RdRp domain; **fig. S9**) to define an approximate “species-rank” ecological unit. These units were called vOTUs according to the recommendations of the community consensus statement (Roux et al., 2019) to reflect the lack of whole genome-based population genetics analyses behind their definition. Pragmatically, even though the consensus statement (Roux et al., 2019) criteria (95% ANI and 85% AF; i.e., wcANI $\geq 80\%$) may seem stricter than those used in

our study (90% ANI and 80% AF; i.e., $wcANI \geq 72\%$), we in fact have shown here that our re-analysis of these cutoffs was more constrained by other biological and ecological information that were not available at the time of development of the consensus statement.

Finally, we compared our $wcANI$ -based vOTUs to vOTUs generated based on RdRp domain sequence similarity (the method that is classically used for such purpose, for instance in reference (Gustavsen et al., 2014)). We independently examined the frequency distribution of pairwise whole RdRp domain amino-acid identities (wdAAI) within each of the three major datasets compared in this study [Global Ocean dataset, GenBank release 233, and the coastal ocean viromes (Wolf et al., 2020)]. Only “complete” RdRp domains were used in this analysis and self matches were excluded. Pairwise comparisons were conducted using Usearch v10.0.240 (Edgar, 2010), requiring a global alignment and a minimum of 50% identity (`-usearch_global -id 0.5 -maxaccepts 300 -self`). The frequency histogram was displayed using the function ‘`gghistogram`’ of the package “`ggpubr`” of R. The sequences from our study provided a balanced representation of the RdRp domain sequence space whereas GenBank and the coastal virome datasets were biased towards (i.e., overrepresented) low and high taxonomic ranks, respectively (**fig. S9E**). Next, the range of wdAAIs in the histogram that provided a ‘trough’ or ‘break’ in sequence space (and hence can be likely used as a cutoff to delineate vOTUs) was examined and individually tested for percent agreement with the $wcANI$ method. Uclust was used to establish RdRp-based clusters at 75%, 80%, 85%, 87%, 90%, 92%, and 95% wdAAI (`--cluster_fast -sort length`) and each cluster set was individually compared to the $wcANI$ cluster set (at different contig lengths) using the “`adj.rand.index`” method described above. Cutoffs in the range of 85–92% wdAAI consistently gave high agreement (>90%) with our $wcANI$ clusters for virus contigs ≥ 1 kb (with the best wdAAI value being 87% as displayed in **fig. S9E**). To determine the novelty of our vOTUs at the “species” rank, the vOTU representatives (the longest sequence in each cluster) were searched against viral RefSeq v.203 using `blastn` and the matches with $\geq 90\%$ nucleic acid identity and $\geq 80\%$ alignment fraction (for the vOTU sequence) were considered to represent known “species”. The results from this analysis are shown in **table S5**.

Calculation of vOTU relative abundances

To calculate vOTU relative abundances in each sample, trimmed reads from each library were first further trimmed off their polyA and polyT stretches (`trimpolya=3 minlength=30`), to avoid inflated abundances for polyA-tailed viruses, using `bbduk` v38.51 (<https://jgi.doe.gov/data-and-tools/bbtools/>). This process was done three more times at a small window size (20 bases per iteration) to avoid aggressive trimming, removing from the right and left of each read only when 10+ consecutive (As/Ts) were found with a hamming distance of 2 (`literal=AAAAAAAAAA,TTTTTTTTTT hdist=2 k=10 minlength=30 ktrim=r restrictright=20; ktrim=l restrictleft=20`) and a final (`trimpolya=3 minlength=30`) run. The virus contigs also went through the same treatment above (without the `minlength=30` flag) to better estimate the horizontal coverage after read mapping.

PolyA/T-trimmed reads were mapped against all polyA/T-trimmed contigs using Bowtie2 v2.4.1 (Langmead and Salzberg, 2012) using the very sensitive, local, and non-deterministic settings and with additional increase of sensitivity by reducing the word size to 16 (`--local -D 20 -R 3 -N 0 -L 16 -i S,1,0.50 -I 0 -X 2000 --non-deterministic`), extracting only aligned reads. The vertical and horizontal coverages of the contigs were calculated independently. For the vertical coverage (i.e., for abundance estimation), reads that mapped at $\geq 90\%$ ID over $\geq 75\%$ of the read

length were extracted using CoverM v0.2.0-alpha6 (<https://github.com/wwood/CoverM>), calculating the trimmed mean (tmean) for each contig. For horizontal coverage (i.e., how many positions across each contig covered by reads), CoverM was used with the same parameters as above, but in the (-m covered_fraction) mode, on a parallel Bowtie2 run with the -a flag turned on, thereby enabling reads to map multiple times to the different members of the same vOTU. Only contigs with 30% or ≥ 1 kb length horizontally covered by reads from both Bowtie2 runs were kept. Tmean values (relative abundances) were adjusted by the number of mapped reads (as filtered by CoverM) to enable for sample-to-sample comparison. Only adjusted abundances of the ≥ 1 -kb contigs were kept, and final abundances of the vOTUs were calculated by summing the adjusted abundances of the ≥ 1 kb contigs belonging to these vOTUs.

Functional annotation of RNA virus genomes

Sequences of all vOTU representatives were first translated using all six frames and standard translation code with transeq (EMBOSS version 6.6.0.0) (Eddy, 1998). RdRp domains were identified by hmmsearch (HMMER, version 3.3) against RdRp HMMs (Wolf et al., 2020) and finding the best match. Sequences with stop codons within the RdRp domain-encoding parts were further checked for usage of alternative translation codes using all alternative codes available in transeq (excluding 0 and 1). Similarly, amino-acid sequences were searched against reference RdRp HMMs. RdRp domains were identified by highest scoring match to reference HMMs, and the frame and codon producing the longest RdRp domain were chosen.

For RNA virus genome domain annotation, open reading frames (ORFs) were identified by using Prodigal (version 2.6.3) with the genetic code identified in the previous step. One iteration of hhblits (-n 1 -e 0.001) [HHsuite version 3.3.0 (Steinegger et al., 2019)] using UniRef30_2020_03 database (Mirdita et al., 2017) was used to generate profiles for amino-acid sequences. Generated profiles were searched against Pfam and reference profiles from a previous study (Wolf et al., 2020) used to identify domains in the amino-acid sequences, and hits with $>95\%$ probability score were used for domain annotation. To increase the number of annotated domains, nucleotide sequences were also searched against the NCBI nr database with DIAMOND (Buchfink and Xie, 2015) blastx v2.0.4.142. Hits with a bitscore >50 were used for annotation.

Shine-Dalgarno sequence identification

Due to the fragmented nature of RNA virus genomes assembled from metatranscriptomes, we combined contigs in each RdRp cluster (50% protein sequence identity, roughly between the family and genus ranks) for Shine-Dalgarno (SD) identification. To remove duplicate or highly similar sequences, we extracted the regions 48 nucleotides upstream of the start codon, and clustered them at 50% identity with vsearch v2.15.1 (Rognes et al., 2016) (--cluster-fast --iddef 0 --id 0.5), using cluster representatives (centroids) for SD identification. ORFs lacking sequence data 48 nucleotides upstream of the start codon were discarded. To identify an SD sequence in a cluster representative, an anti-Shine-Dalgarno motif (ASD, in this case 3'-UUCCUCCA-5') was matched up to each position of the translation initiation region (defined as 0–15 nucleotides from the 5' end of the ASD sequence to the first nucleotide of the start codon). An SD sequence was identified when at any of these positions the ASD motif was determined to bind at -5 kcal/mol or stronger as determined by free_scan.pl from the free2bind suite of software (Starmer et al., 2005). The fraction of genes with an SD sequence for each cluster was calculated. As a control, we also calculated the number of "mock" SDs (MSDs) found during performance of the same analysis over

a window with 25–40 nucleotides from the 5' end of the ASD motif to the first nucleotide of the start codon (well outside the translation initiation region). To calculate the *p*-value of SD signal significance for each RdRp cluster we performed a binomial test, in which successes were the SD count, trials were the number of genes, and rates were given by MSD count / number of genes (MSD counts of 0 were set to 1). We limited testing to RdRp clusters with at least 10 genes. *P*-values were Bonferroni-corrected.

Inference of virus-host interactions

Virus-host associations were assessed based on three approaches that could be used for RNA viruses: (i) abundance-based co-occurrence (Kaneko et al., 2020), (ii) RdRp protein sequence similarity to endogenous virus elements (EVEs) (Shi et al., 2016), and (iii) RdRp protein sequence similarity to known RNA viruses. For the first approach, a global network of putative direct associations was built from abundance-based co-occurrence patterns (among virus OTUs and cellular hallmark-gene amplicons) using FlashWeave (Tackmann et al., 2019), run with (heterogeneous = false) and (sensitive = true) settings, both positive and negative weight values, and a *Q*-value <0.01 (default) were kept. The sensitivity analysis was performed using different thresholds for both negative and positive edge weights. Given that the number of negative associations was very low and that a virus needs its host to replicate, only positive associations with hosts were kept. A conservative threshold of edge weight ≥ 0.4 was used to assess virus-host inferences (Kaneko et al., 2020). TIM (<https://github.com/RomainBlancMathieu/TIM>) was used to distil the most significant connections from the co-occurrence analysis. TIM assumes that (i) evolutionarily related viruses infect the evolutionarily related host, and (ii) in the co-occurrence network, the number of connections between the presumed virus-host should be enriched compared to those of a non-host (not by chance). Results from TIM were filtered using the corrected *p*-value (*Q*) <0.05, and a single host was assigned to each vOTU based on the strongest correlation.

For the second approach, all nucleotide sequences from cellular organisms available in NCBI GenBank release 243 were used as a nucleotide database. To avoid including exogenous RNA virus genomes in the database, we excluded sequences shorter than 45 kb since the longest RNA virus genome reported so far is 41.1 kb (Saber et al., 2018). To assess the evolutionary relationship of OTUs of exogenous viruses to EVEs, the near-complete RdRp protein sequences were searched against the nucleotide database by using tblastn algorithm. As previously described (Shi et al., 2016), the thresholds were set to 100 amino acids for alignment length and 1×10^{-20} for *e*-value.

For the third approach, the known RNA virus taxa (phyla, classes, and families) that were assigned to the vOTUs after clustering the RdRp domain protein sequence similarity network (see **RdRp-based taxonomy annotation of RNA viruses**) were used to retrieve previous information on putative hosts. Since taxonomy-based host assignment cannot be done for novel RNA virus phyla and classes, we only used the co-occurrence and EVEs approaches to predict their hosts.

Determination of the genomic strandedness for novel orthornaviran phyla

To determine the strandedness for the new megataxa discovered here, we adopted a previously described read-mapping approach (Obbard et al., 2020) that inferred the RNA virus genome type from the observation that positive-sense single-stranded RNA (+ssRNA) viruses would be heavily biased towards being covered by forward reads in metatranscriptomes. Double-stranded RNA (dsRNA) and negative-sense single stranded RNA (-ssRNA) viruses, on the other

hand, would be slightly biased towards being covered by forward and reverse metatranscriptomic reads, respectively. In our study, Samtools v1.10 was used with the `f` flag to extract the reads that mapped into the forward (`-f 99` and `-f 147`) and reverse (`-f 83` and `-f 163`) directions from the bam files created above (from the Bowtie2 run with the `-a` flag turned on to enable reads to map multiple times to the different members of the same vOTU; see “**Calculation of vOTU relative abundances**”). Only contigs with non-adjusted vertical coverage $\geq 10X$ (calculated by the `tmean` method of CoverM) and with horizontal coverage $>70\%$ were kept for downstream analyses. Next, CoverM was used to calculate the number of reads mapping in the forward and reverse directions separately (`--min-read-percent-identity .90 --min-read-aligned-percent .75 -m count`). The number of reads mapping in each direction were independently summed for all the contigs per vOTU (taking into account the alignment of the contig relative to the vOTU representative), and the final strandedness for each vOTU was corrected for the vOTU orientation (from the assembly step) by taking into account the frame translation of the annotatable genes on the contig. The ratio of the reads mapping to the positive strand to those mapping to the negative strand of the vOTU in each sample were calculated and \log_2 transformed for visualization.

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,‡}.

¹Department of Marine Biology and Oceanography, Institut de Ciències del Mar (CSIC), Barcelona, Catalonia, Spain.

²Département de biologie, Québec Océan and Takuvik Joint International Laboratory (UMI3376), Université Laval (Canada) - CNRS (France), Université Laval, Québec, QC, G1V 0A6, Canada.

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

⁴Max Delbrück Centre for Molecular Medicine, 13125 Berlin, Germany.

⁵Department of Bioinformatics, Biocenter, University of Würzburg, 97074 Würzburg, Germany.

⁶School of Marine Sciences, University of Maine, Orono, Maine 04469, USA.

⁷Ecole 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.

⁸European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge, UK.

⁹CNRS, UMR 7144, EPEP & Sorbonne Universités, UPMC Université Paris 06, Station Biologique de Roscoff, 29680 Roscoff, France.

¹⁰Sorbonne Universités, UPMC Université Paris 06, CNRS, Laboratoire d'oceanographie de Villefranche (LOV), Observatoire Océanologique, 06230 Villefranche-sur-Mer, France.

¹¹Department of Oceanography, University of Hawaii, Honolulu, Hawaii 96822, USA.

¹²CNRS, UMR 7232, BIOM, Avenue du Fontaulé, 66650 Banyuls-sur-Mer, France.

¹³Sorbonne Universités Paris 06, OOB UPMC, Avenue du Fontaulé, 66650 Banyuls-sur-Mer, France.

¹⁴Aix Marseille Univ, Université de Toulon, CNRS, IRD, MIO, Marseille, France.

¹⁵Stazione Zoologica Anton Dohrn, Villa Comunale, 80121 Naples, Italy.

¹⁶CEA - Institut de Génomique, Genoscope, 2 rue Gaston Crémieux, Evry France.

¹⁷CNRS, UMR 8030, 2 rue Gaston Crémieux, Evry France.

¹⁸Université d'Evry, UMR 8030, CP5706, Evry France.

¹⁹Directors' Research European Molecular Biology Laboratory Meyerhofstr. 1 69117 Heidelberg Germany.

²⁰CNRS, UMR 7144, Sorbonne Universités, UPMC Université Paris 06, Station Biologique de Roscoff, 29680 Roscoff, France.

²¹Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto, 611-001, Japan.

²²Bigelow Laboratory for Ocean Sciences, East Boothbay, ME, 04544, USA.

²³MARUM, Center for Marine Environmental Sciences, University of Bremen, Bremen, Germany.

²⁴PANGAEA, Data Publisher for Earth and Environmental Science, University of Bremen, Bremen, Germany.

²⁵CNRS, UMR 7009 Biodev, Observatoire Océanologique, F-06230 Villefranche-sur-mer, France.

²⁶Laboratoire de Physique des Océans, UBO-IUEM, Place Copernic, 29820 Plouzané, France.

²⁷Department of Geosciences, Laboratoire de Météorologie Dynamique (LMD), Ecole Normale Supérieure, 24 rue Lhomond, 75231 Paris Cedex 05, France.

²⁸Department of Microbiology, The Ohio State University, Columbus, OH 43214, USA.

²⁹Department of Civil, Environmental and Geodetic Engineering, The Ohio State University, Columbus OH 43214 USA.

³⁰Department of Biology, Institute of Microbiology and Swiss Institute of Bioinformatics, ETH Zurich, Vladimir-Prelog-Weg 4, 8093 Zurich, Switzerland.

‡*Tara* Oceans Consortium.

Detailed *Tara* Oceans expeditionary support and additional 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 - Global Ocean System Ecology and Evolution), 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 Agglomération, 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

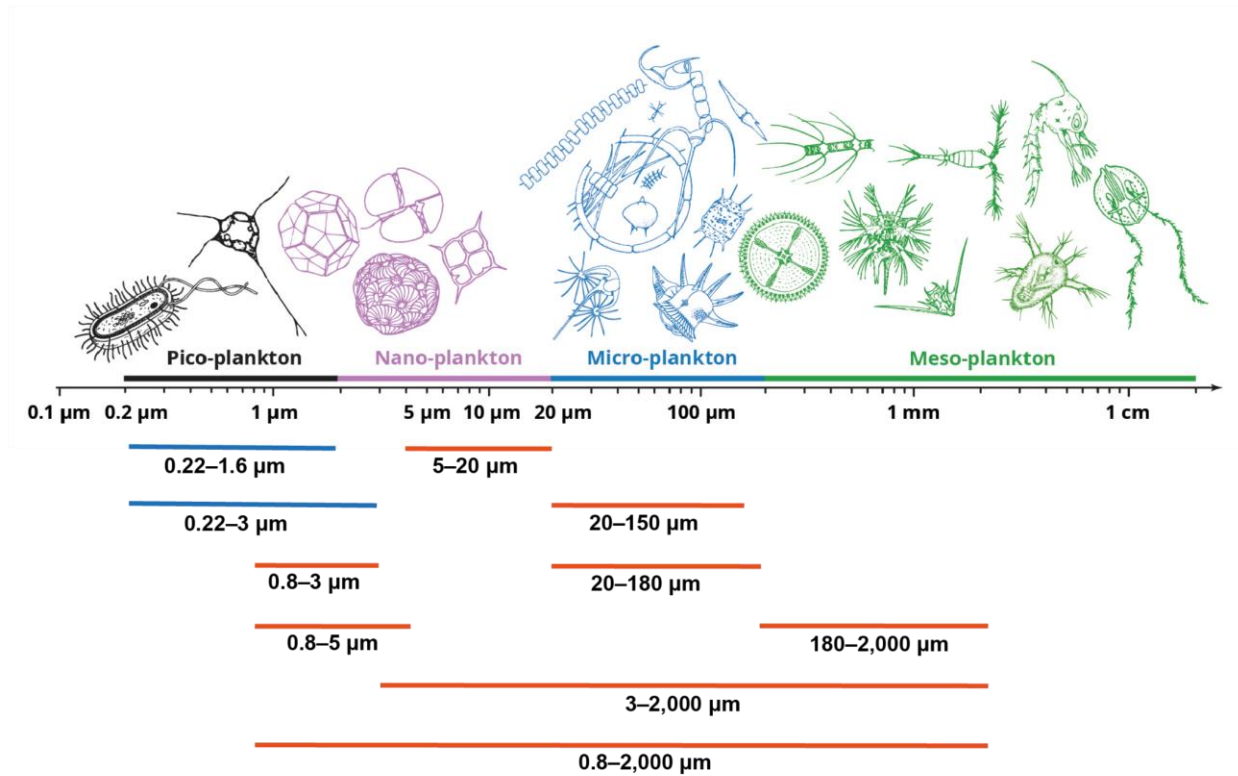


Fig. S1. Plankton organismal sizes considered in this study.

The graded top bar represents the logarithmic scale of the organismal (unicellular or multicellular) sizes in length units, from viruses via prokaryotes and protists to metazoans from seawater plankton. The colored bars on the top indicate operational size-fractions of plankton: pico-plankton (0.2–2 μm), nano-plankton (2–20 μm), micro-plankton (20–200 μm), and meso-plankton (200–2,000 μm). The blue (prokaryote-enriched) and orange (eukaryote-enriched) bars indicate the operational organismal size fractions utilized for viral metatranscriptomics in this work. We define “eukaryotic” or “eukaryote-enriched” fractions (orange bars) as those samples enriched for eukaryotes by filtration during sampling, though these fractions also contain prokaryotes and viruses that might be part of the eukaryotic holobiont either as a symbiont or as food. Similarly, we refer to “prokaryotic” or “prokaryote-enriched” fractions (blue bars) as those enriched for bacteria and archaea, but where smaller unicellular eukaryotes (e.g., picoeukaryotes) and viruses are also routinely recovered.

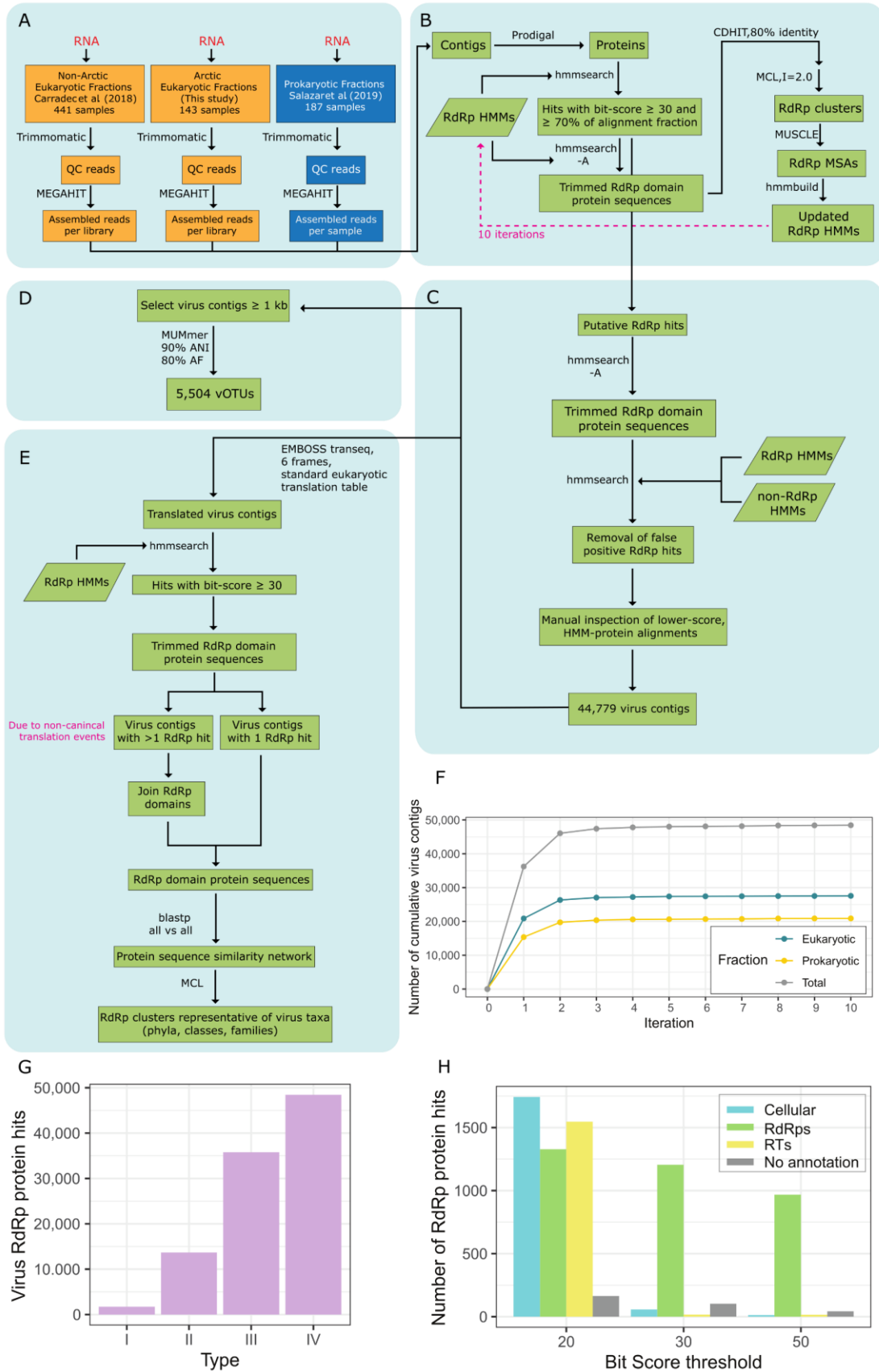


Fig. S2. Bioinformatic workflow and RNA virus identification.

Schematic representation of the process for distinct bioinformatic steps. **(A)** Metatranscriptomic reads from samples collected during the *Tara* Oceans and *Tara* Oceans Polar Circle expeditions were quality-trimmed and assembled into contigs. **(B)** After predicting protein sequences from the contigs, RNA viruses were identified by using an HMM search-and-update pipeline that iteratively improves HMMs to detect highly divergent RdRp domain protein sequences. The dashed, pink line arrow represents the 10 cycles of the process. **(C)** Validation of the RdRp hits was done by competitively searching against RdRp and non-RdRp HMM profiles. **(D)** The contigs for which the RdRps were validated were clustered using 90% of average nucleotide identity (ANI) and 80% of alignment fraction (AF) to obtain the vOTUs for the ecological analyses. **(E)** Problems associated with alternative genetic codes and non-canonical translation events were avoided by using translated contigs (instead of predicted proteins) and reconstructing the RdRp domain sequences from virus contigs. The reconstructed domains were used to build a protein sequence similarity network that was clustered with MCL (applying the benchmarked inflation value thresholds; see **Methods**) to obtain the taxonomic classification of RNA viruses. The colors (blue, orange, or green) of the text boxes indicate the organismal fractions from which the sequences were derived (prokaryotic, eukaryotic, or both, respectively). **(F)** Virus contigs detected in the prokaryotic and eukaryotic fractions along the ten RdRp sequence search/hidden Markov model (HMM) update iterations. **(G)** Comparison of virus RdRp sequence identification methods using the Global Ocean prokaryote-enriched metatranscriptomic dataset: (I) blastp search against nr database using an e-value cutoff of $<10^{-5}$; (II) HMM-based search using the 14 profiles for virus RdRps in Pfam; (III) HMM-based search using the 65 profiles used for virus sequence identification in this study; and (IV) ten iterations of the HMM-based search using the same 65 profiles. **(H)** Benchmarking of the HMMER bitscore threshold using 20 HMMs, derived from the 20 virus RdRp “superfamily” clusters suggested by Shi et al. (2016), searched against protein sequences predicted from the prokaryote-enriched metatranscriptomic sequencing data. Specificity and sensitivity of the viral identification pipeline after five iterations were estimated using different bit score thresholds. Hits were annotated, using blastp with an e-value threshold of 10^{-5} against the NCBI GenBank non-redundant database, into RdRps, reverse transcriptases (RTs), sequences from cellular organisms (Cellular), and sequences without matches (No annotation).

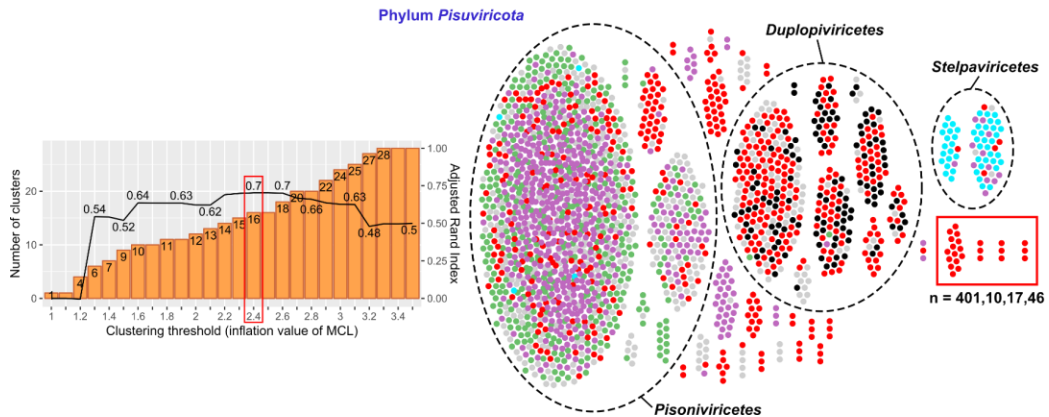
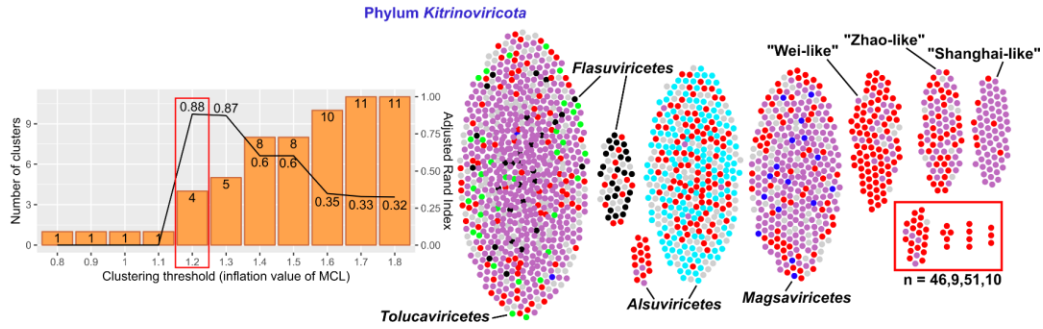
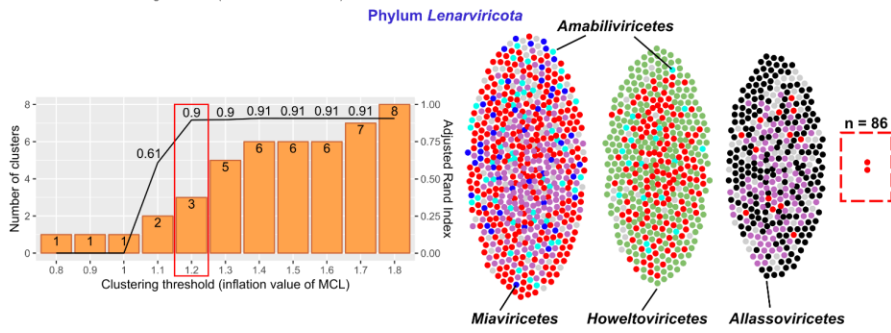
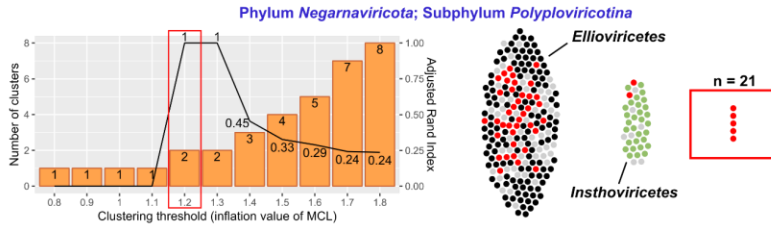
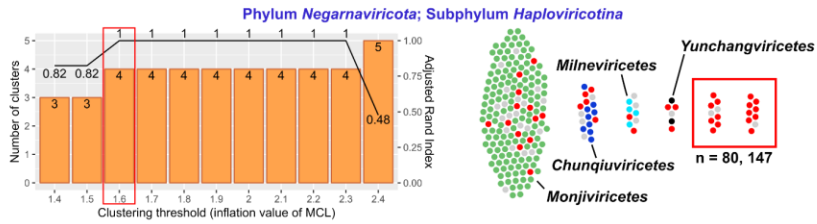
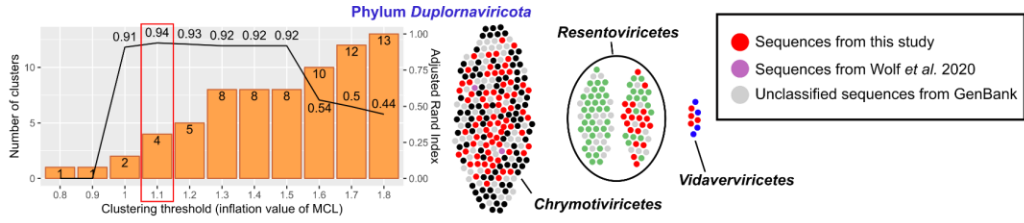


Fig. S3. Establishment of RdRp domain-based class-rank clusters included in previously established orthornaviran phyla.

Similar to **Fig. 1**, percent agreement (line) of our network-guided and phylogeny-based classes at different clustering thresholds are shown per each phylum/subphylum (left). Bars represent the number of clusters of near-complete RdRp domain sequences at these different clustering thresholds. Only virus sequences of established taxa were used for calculating the agreement percentage. Swarm plots of the emerging clusters at the chosen (red-boxed) inflation value are shown on the right of each bar plot. Solid black lines encompass sequences that were exclusively joined at a lower inflation value, whereas the dashed black lines encompass the sequence clusters assigned to phylum *Pisuviricota* that were not exclusively joined at lower inflation values. Dot colors used in each row are randomly assigned to show finer taxonomic resolution within each phylum and are independent from other rows except for the three categories shown in the legend. New classes emerging exclusively from our study are red-boxed, with solid red lines indicating the retrieval of all canonical motifs in the RdRp domain (see **table S10** for domain motifs). Singletons have been removed from this analysis and new classes were required to have at least two sequence representatives (of 50% identity clusters). The total number of sequences from our study (complete and partial) represented by these dots are shown around the red boxes in their respective orders.

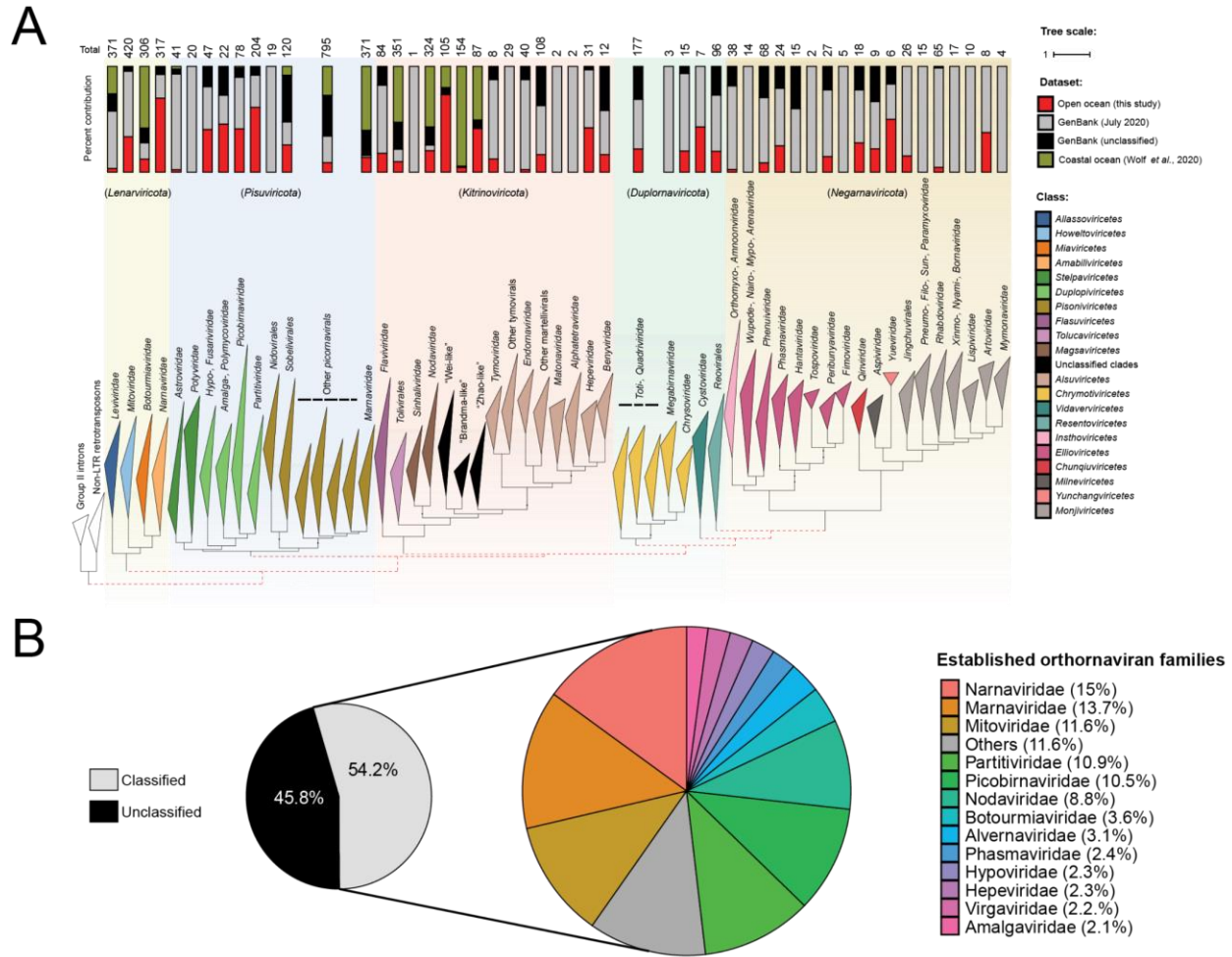


Fig. S4. Marine RNA viruses of established families.

(A) Approximate RdRp tree showing previously established taxa in riboviriad kingdom *Orthornavirae* as inferred in Wolf et al. (2018). Taxa were collapsed at the family or order rank. The stacked bar plot shows the relative dataset-specific contribution to each clade (clustered sequences with <50% identity), with the total number of sequences indicated at the top of each bar plot. Black dots on the branches indicate the removal of unclassified clades to improve visualization. Red dashed lines indicate the phylogenetic relationships revised by our study (shown in Fig. 3). (B) Pie charts showing that after MCL clustering of the near-complete RdRp protein sequence similarity network and building phylogenies within classes, 49 out of the 103 ICTV-established families were assigned to 54.2% of the orthornavirans captured in this work (table S7).

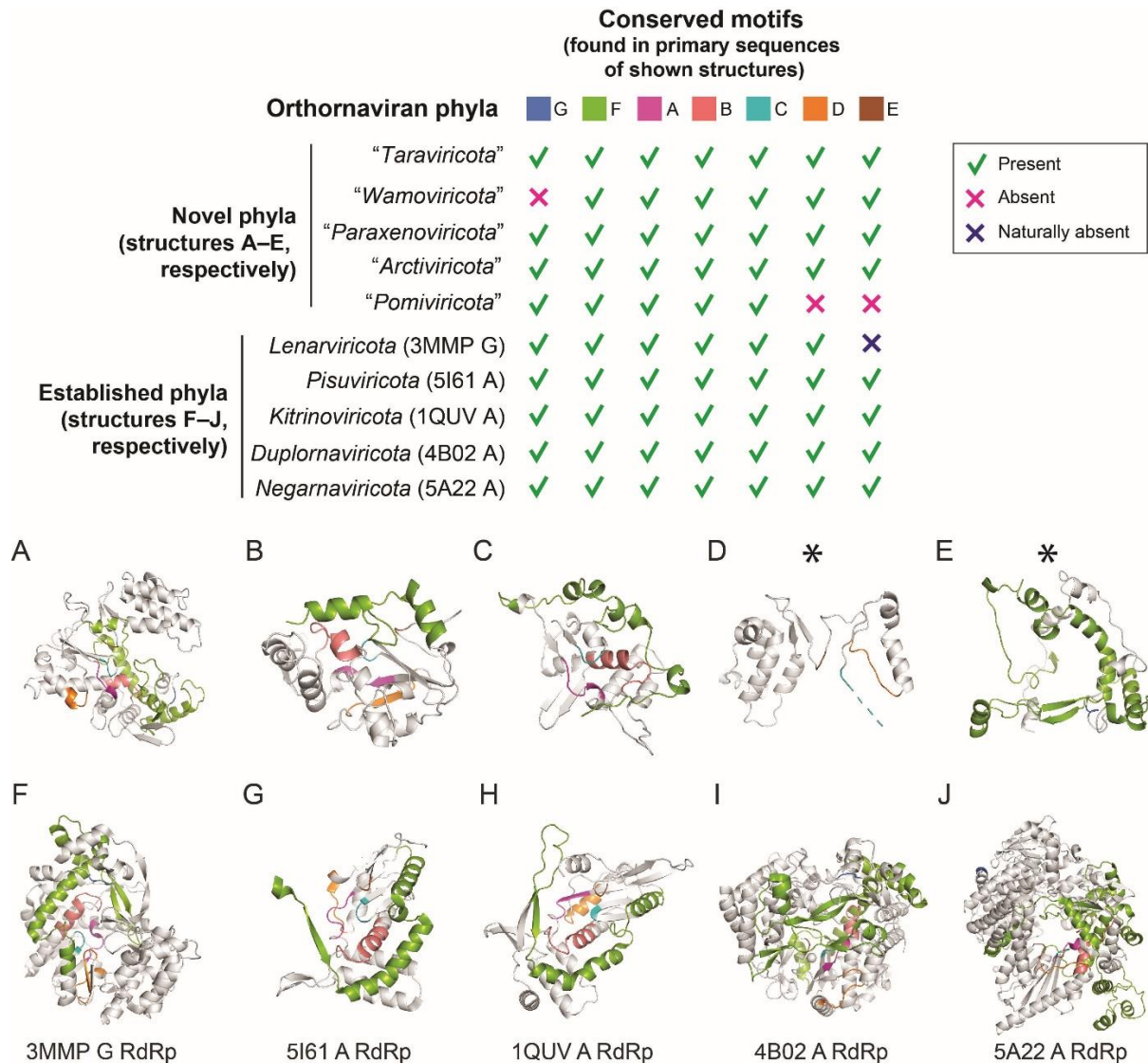


Fig. S5. Representative, modeled regions of RdRp three-dimensional structures of the five novel and five established orthornaviran phyla.

The top table shows the presence of primary sequence-identified conserved motifs of representative RdRp proteins of orthornaviran phyla, where full RdRp sequences known to lack a motif are described as ‘naturally absent’ and partial RdRp sequences are described only as ‘absent’ as missing motifs could be due to limitations of the dataset or real biology. The two rows of protein structures are predicted RdRp ribbon-style models with conserved motifs color coded (as described in **Methods**). For structures of divergent and novel orthornaviran phyla, to most conservatively present the data, only the primary sequence regions that can be modeled in three-dimensional space are shown. Asterisks for D and E indicate low confidence models (<90% confidence; see **Methods**). For the specifics of primary sequence-identified conserved motifs, see **table S10** and **Data S1**.

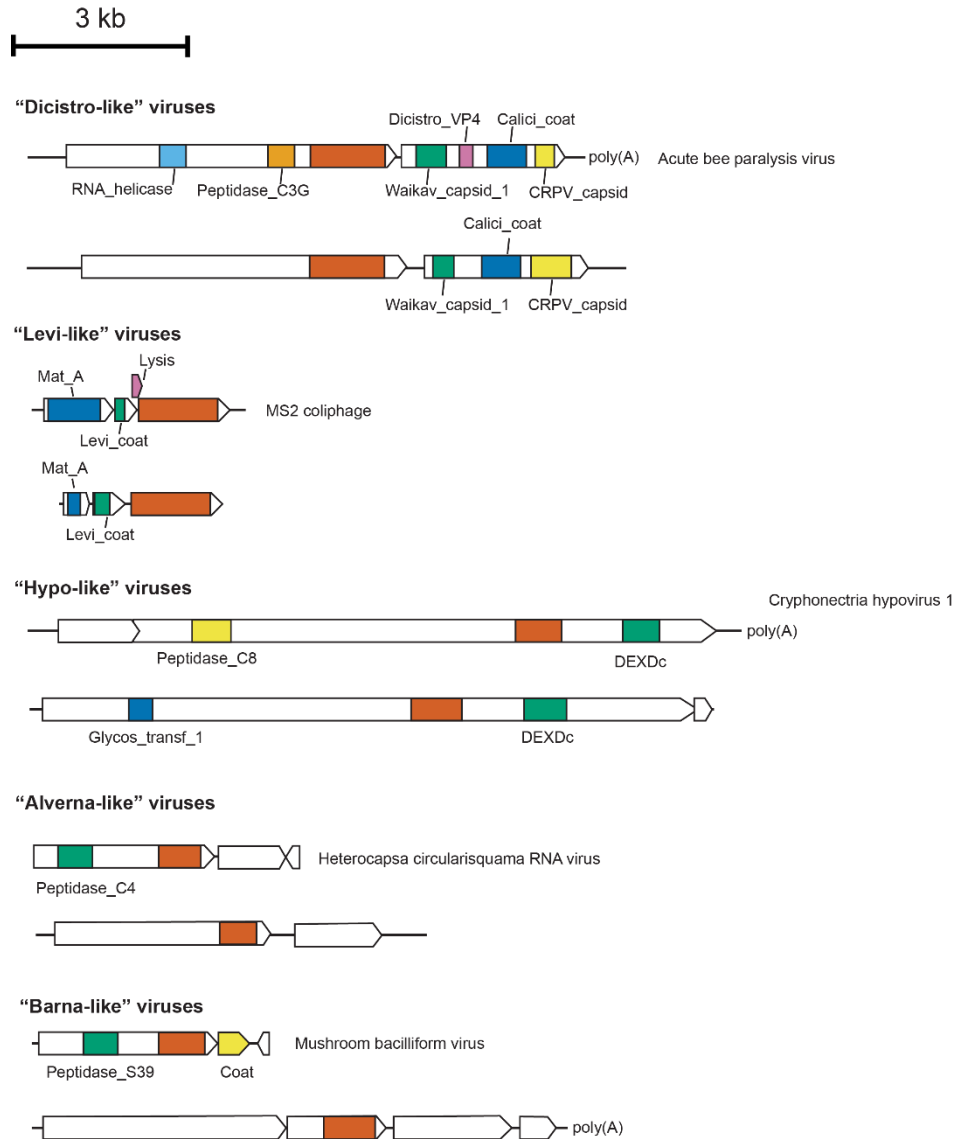


Fig. S6. Representative genome organization of viruses of established orthornaviran families.

Longer genomes representing five known orthornaviran taxonomic groups (approximately at the rank of family) were selected to show the genome organizations, along with the corresponding reference viruses for comparison. Within each virus genome, the white arrow boxes define the ORF boundaries and sense, whereas the inner boxes define signals with blastx matches and/or functional protein domain. Each color represents a different protein/domain region for the same genome. The virus RdRp domain is indicated in dark orange across all genomes. RNA_helicase, virus RNA helicase; Peptidase_C3G, Tungro spherical virus-type protease; Waikav_capsid_1, waikavirus capsid protein 1; Dicistro_VP4, cricket paralysis virus capsid protein VP4; Calici_coat, calicivirus coat protein; CRPV_capsid, cricket paralysis virus capsid protein like; Mat_A, maturation protein A; Levi_coat, levivirus coat protein; Lysis, levivirus lysis protein; Peptidase_C8, hypovirus cysteine peptidase family C8; DEXDc, DEAD-like helicases superfamily protein; Peptidase_C4, peptidase family C4.

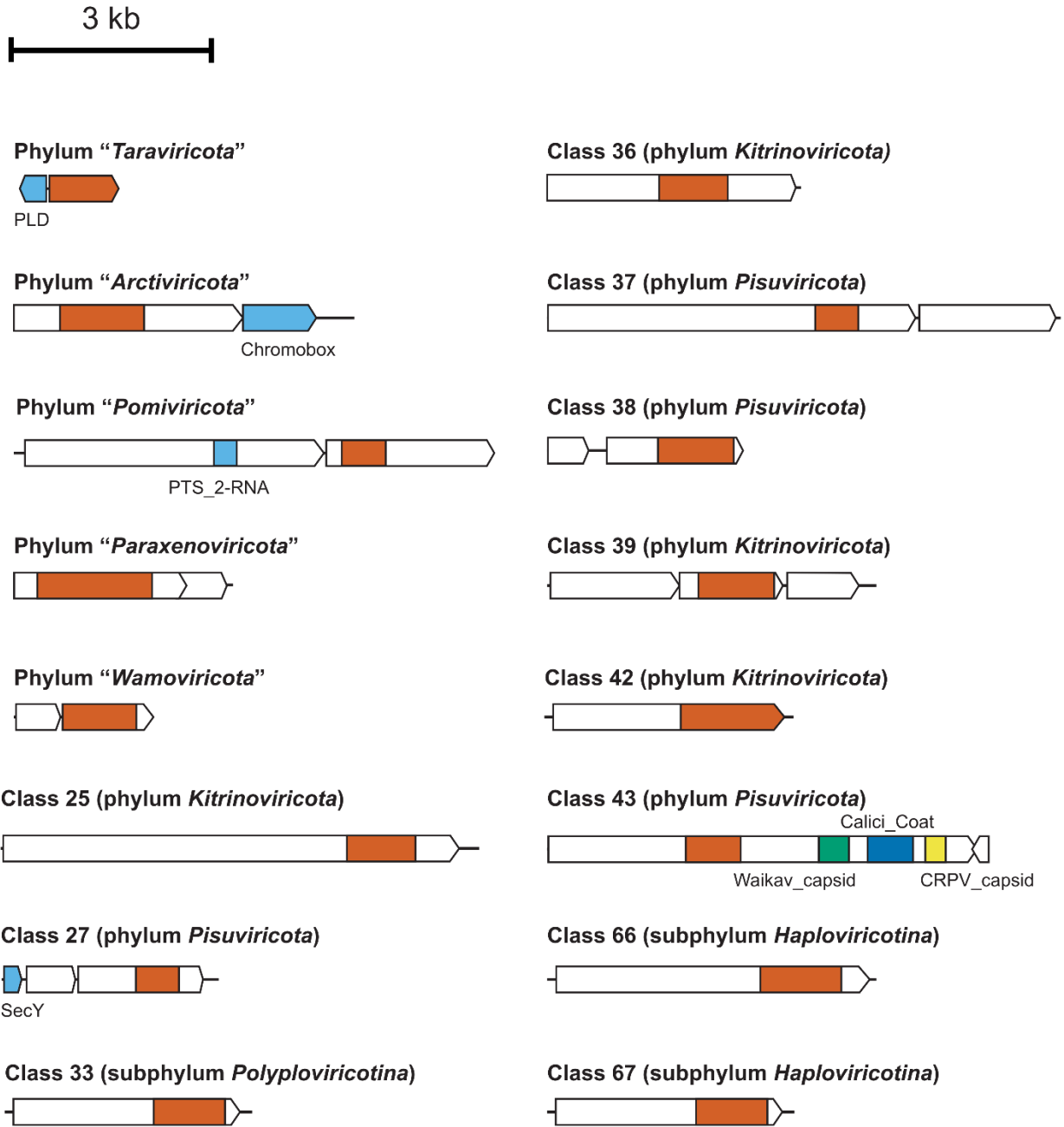


Fig. S7. Representative genome organization of novel orthornavirans.

Shown are longer genomes from novel phyla and classes. Due to virus sequence divergence and limitations of public databases, only a few functions could be assigned to the proteins beyond RdRp domain. The representative genomes for three novel phyla and one novel class encode "cellular" proteins of diverse functions. PLD, phospholipase D α 1; Chromobox, chromobox domain protein; PTS_2-RNA, RNA 2'-phosphotransferase; SecY, SecY subunit of the bacterial Type-II secretion system; Waikav_capsid_1, waikavirus capsid protein 1; Calici_coat, calicivirus coat protein; CRPV_capsid, cricket paralysis virus capsid protein like. Figure legend follows **Fig. S6**.

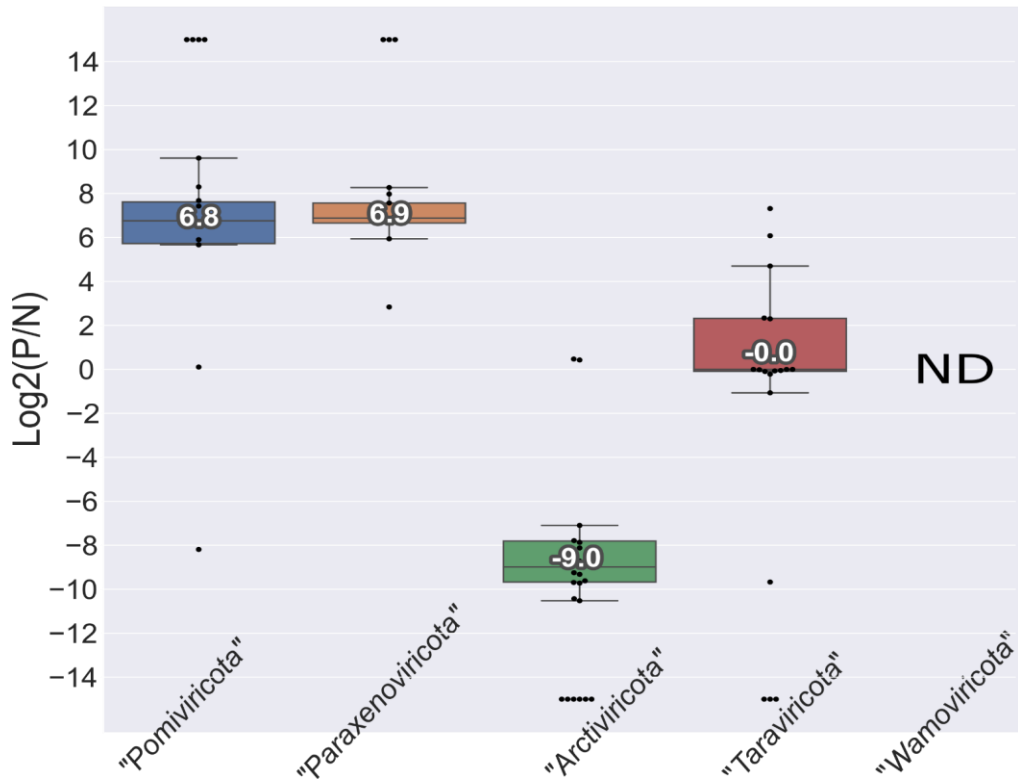


Fig. S8. Genomic strandedness of novel RNA virus phyla.

Boxplots to infer the strandedness (+ssRNA, -ssRNA, or dsRNA viruses) for the new phyla inferred in this study, quantified as \log_2 of number of reads mapping to the positive sense divided by those mapping to the negative sense (higher positive values indicate +ssRNA viruses, whereas negative values indicate -ssRNA viruses; see **Methods**). Each point corresponds to a different vOTU per sample, grouped according to the phylum as shown on the x-axis. For visualization purposes, vOTUs with reads mapping exclusively in either the positive- or negative-sense are shown with values of +15 and -15, respectively (both values are arbitrarily chosen to exceed the maximum and minimum observed values, respectively, and the median for each boxplot was calculated without including those arbitrary values). ND; not determined due to lack of enough coverage by reads (see **Methods**).

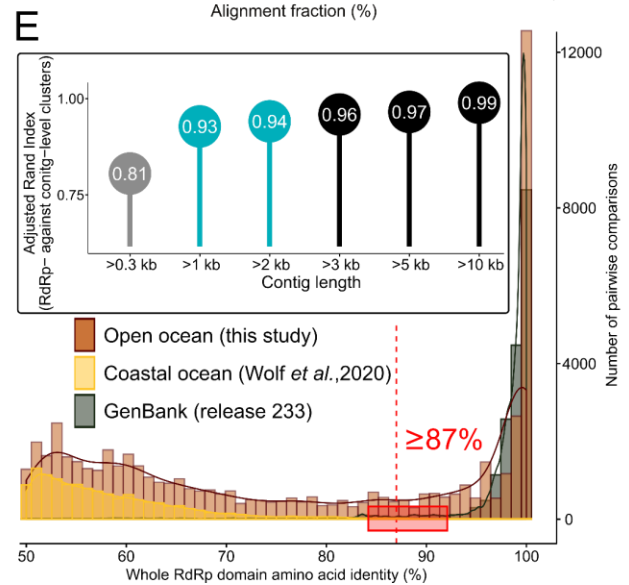
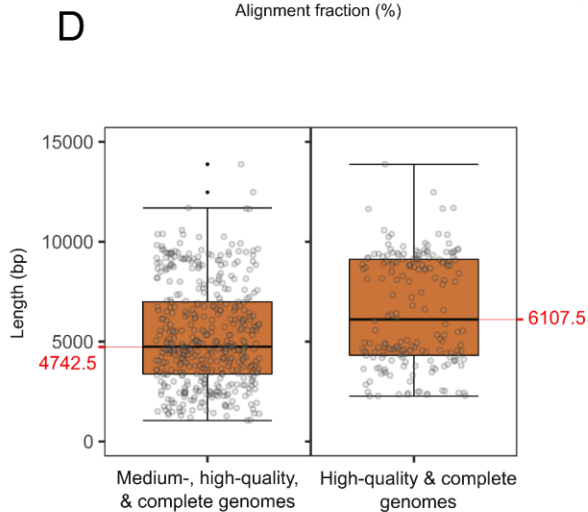
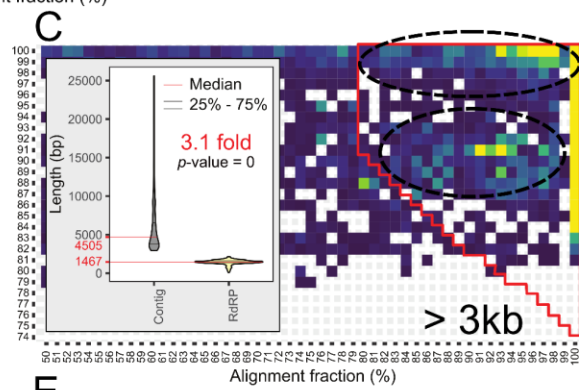
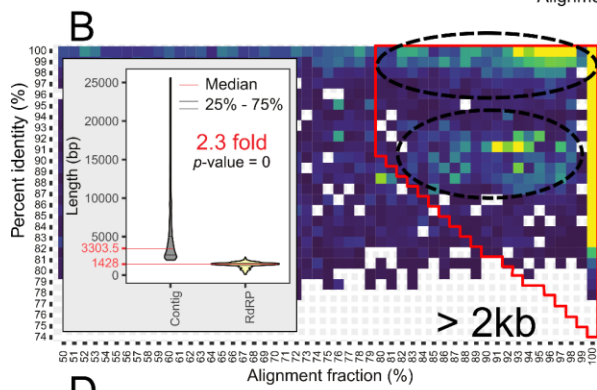
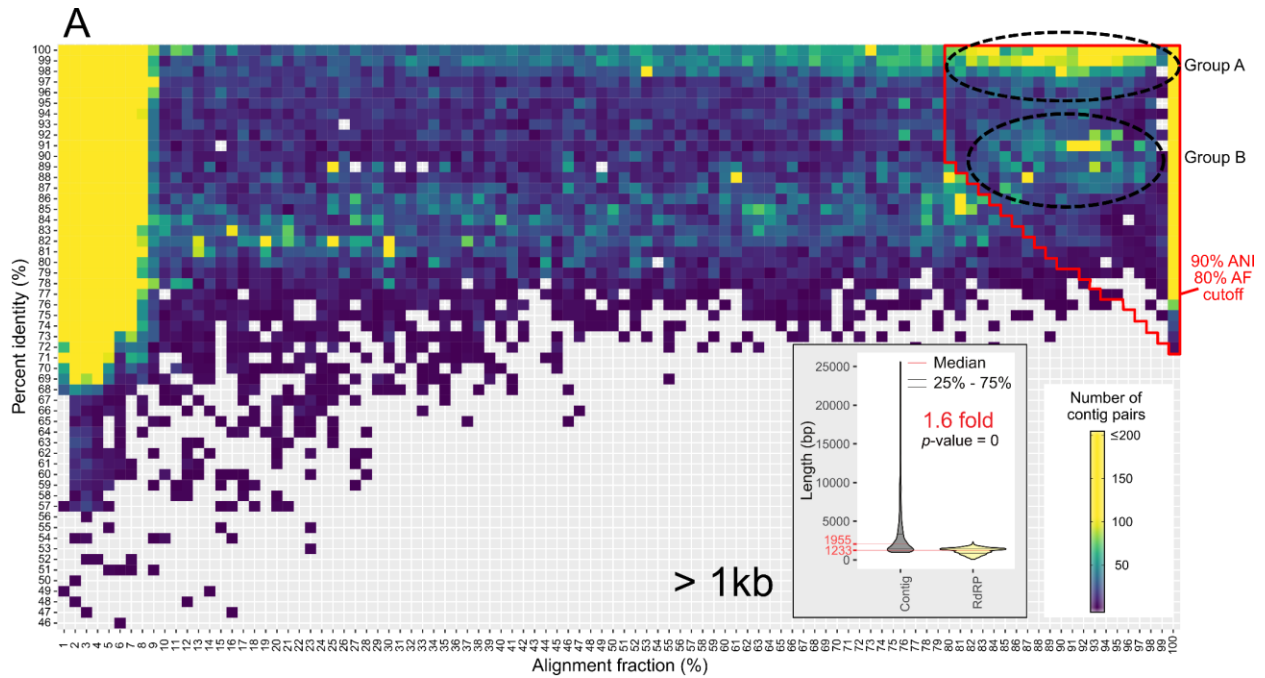


Fig. S9. Establishment of genome-based universal cutoffs for vOTUs. (A–C).

Heatmaps showing the frequency (cell color intensity) of pairwise average nucleotide identity (ANI) and alignment fraction (AF) for all the virus contigs ≥ 1 , ≥ 2 , and ≥ 3 kb identified in this study, respectively, binned at 1% intervals. The two groups of genome pairs circled with black dashed lines would fall within the virus operational taxonomy unit (vOTU) cutoffs (highlighted red) used in this study as previously demonstrated for dsDNA viruses (Roux et al., 2019). The ≥ 2 , and ≥ 3 -kb analyses show that the cutoffs used for the ≥ 1 -kb contigs are the same, but will result in much sparser data for downstream analyses. The insets represent violin plots (function ‘geom_violin’ of ggplot2 in R) comparing the size distribution of the contigs and their RdRps, showing that even for the ≥ 1 kb contigs the genomic information goes beyond the RdRp domain. *P*-values were calculated from a Wilcoxon Rank Sum test (function ‘wilcox.test’ in R) on the medians (red lines). **(D)** Boxplots (function ‘geom_boxplot’ of ggplot2 in R) showing the medians (red lines) of contig length distribution for the high- and medium-quality genomes identified in this study. The ≥ 1 -kb contig length cutoff captures all of these genomes in downstream analyses, whereas longer cutoffs (such as 10 kb and 5 kb) usually used for studying dsDNA viruses (Roux et al., 2017; Gregory et al., 2019) would fail to capture almost all or most of them, respectively. **(E)** Frequency histograms of the RdRp sequence-space similarity scores across reference (GenBank; grey), coastal ocean (Wolf et al., 2020; yellow), and open ocean (this study; brown) datasets. The histogram shows that the sequences from our study provide a balanced representation of the RdRp domain sequence space whereas GenBank and the coastal virome datasets were biased towards (i.e., overrepresented) low and high taxonomic ranks (right and left of the dotted red line), respectively. The red box indicates the range of percent identities commonly used to establish a vOTU, resulting in >90% agreement with our genome-wide ecological unit demarcation (inset), with the highest agreement achieved at 87% percent identity (dotted red line). The inset shows the percent agreement of vOTU delineations from genome-wide versus near-complete RdRp sequences (and partial RdRp sequences for short contigs <1 kb).

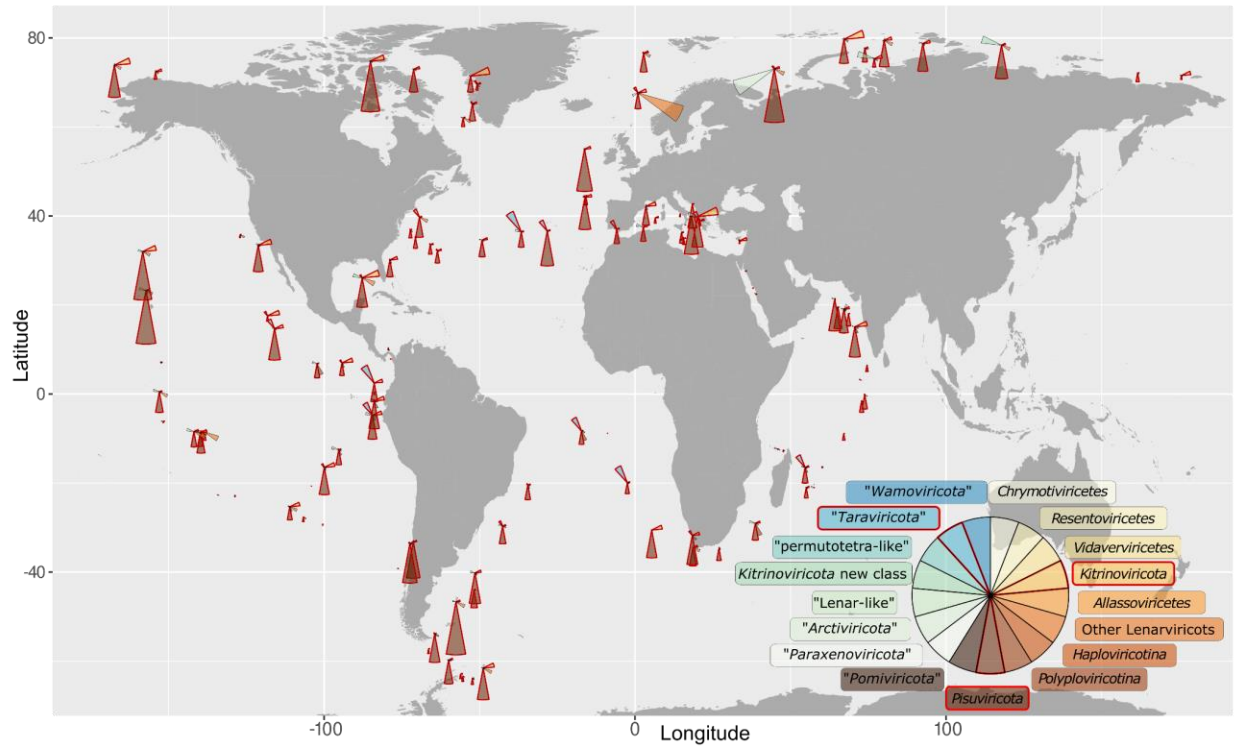


Fig. S10. Biogeography of RNA virus megataxa.

Global map showing the distribution and relative abundance of RNA viruses inferred in this study per megataxon. The position and color of wedges are fixed for the same megataxon across the Global Ocean. Wedge lengths are proportional to the cumulative abundance of all vOTUs belonging to the same megataxon in the sample as well as across the global dataset. The average relative abundances of vOTUs per phylum are shown in **Fig. 4** and **fig. S11**.

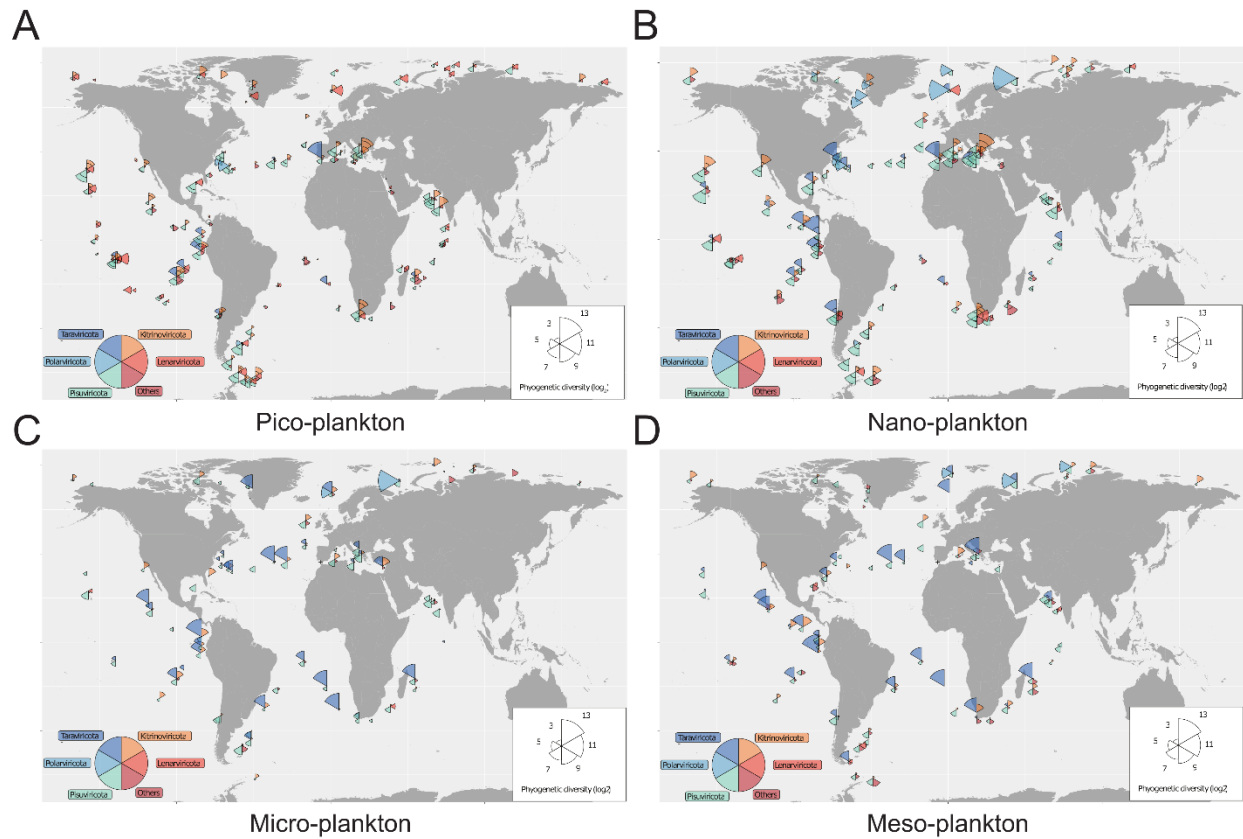


Fig. S11. Biogeography of orthornaviran megataxa per size fraction.

Global map showing the distribution and average relative abundance (on a \log_2 scale) of vOTUs inferred in this study per phylum per size fraction. The position and color of wedges are fixed for the same megataxon across the Global Ocean. Wedge lengths are proportional to the average abundance of all vOTUs belonging to the same megataxon in the sample as well as across the global dataset. Similar to the overall average relative abundances of vOTUs per phylum (**Fig. 4**), “*Taraviricota*” and “*Arctiviricota*” were on average the most abundant (except for the smallest size; the pico-plankton fraction), consistent with predicted eukaryotic hosts.

References and Notes

1. M. E. J. Woolhouse, L. Brierley, Epidemiological characteristics of human-infective RNA viruses. *Sci. Data* **5**, 180017 (2018).
2. K.-B. G. Scholthof, S. Adkins, H. Czosnek, P. Palukaitis, E. Jacquot, T. Hohn, B. Hohn, K. Saunders, T. Candresse, P. Ahlquist, C. Hemenway, G. D. Foster, Top 10 plant viruses in molecular plant pathology. *Mol. Plant Pathol.* **12**, 938–954 (2011).
3. A. Brun, Vaccines and Vaccination for Veterinary Viral Diseases: A General Overview. *Methods Mol. Biol.* **1349**, 1–24 (2016).
4. 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).
5. 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. *Nat. Microbiol.* **5**, 1262–1270 (2020).
6. M. Shi, X.-D. Lin, X. Chen, J.-H. Tian, L.-J. Chen, K. Li, W. Wang, J.-S. Eden, J.-J. Shen, L. Liu, E. C. Holmes, Y.-Z. Zhang, The evolutionary history of vertebrate RNA viruses. *Nature* **556**, 197–202 (2018).
7. 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–e18 (2018).
8. M. Krupovic, V. V. Dolja, E. V. Koonin, Plant viruses of the Amalgaviridae family evolved via recombination between viruses with double-stranded and negative-strand RNA genomes. *Biol. Direct* **10**, 12 (2015).
9. V. V. Dolja, E. V. Koonin, Metagenomics reshapes the concepts of RNA virus evolution by revealing extensive horizontal virus transfer. *Virus Res.* **244**, 36–52 (2018).
10. C. W. Carter Jr., What RNA World? Why a Peptide/RNA Partnership Merits Renewed Experimental Attention. *Life* **5**, 294–320 (2015).
11. S. Chatterjee, S. Yadav, The Origin of Prebiotic Information System in the Peptide/RNA World: A Simulation Model of the Evolution of Translation and the Genetic Code. *Life* **9**, 25 (2019).
12. A. Pereira Dos Santos Jr., M. V. José, S. Torres de Farias, From RNA to DNA: Insights about the transition of informational molecule in the biological systems based on the structural proximity between the polymerases. *Biosystems* **206**, 104442 (2021).
13. S. T. de Farias, A. P. Dos Santos Jr., T. G. Rêgo, M. V. José, Origin and Evolution of RNA-Dependent RNA Polymerase. *Front. Genet.* **8**, 125 (2017).
14. P. Forterre, The origin of viruses and their possible roles in major evolutionary transitions. *Virus Res.* **117**, 5–16 (2006).
15. P. Forterre, D. Prangishvili, The origin of viruses. *Res. Microbiol.* **160**, 466–472 (2009).

16. J. H. Kuhn, Y. I. Wolf, M. Krupovic, Y.-Z. Zhang, P. Maes, V. V. Dolja, E. V. Koonin, Classify viruses - the gain is worth the pain. *Nature* **566**, 318–320 (2019).
17. E. V. Koonin, V. V. Dolja, M. Krupovic, A. Varsani, Y. I. Wolf, N. Yutin, F. M. Zerbini, J. H. Kuhn; Global Organization and Proposed Megataxonomy of the Virus World, Global Organization and Proposed Megataxonomy of the Virus World. *Microbiol. Mol. Biol. Rev.* **84**, e00061–e19 (2020).
18. J. B. Whitfield, P. J. Lockhart, Deciphering ancient rapid radiations. *Trends Ecol. Evol.* **22**, 258–265 (2007).
19. M. Krupovic, V. V. Dolja, E. V. Koonin, The LUCA and its complex virome. *Nat. Rev. Microbiol.* **18**, 661–670 (2020).
20. D. M. Kristensen, A. R. Mushegian, V. V. Dolja, E. V. Koonin, New dimensions of the virus world discovered through metagenomics. *Trends Microbiol.* **18**, 11–19 (2010).
21. E. C. Holmes, S. Duchêne, Can Sequence Phylogenies Safely Infer the Origin of the Global Virome? *mBio* **10**, 1–2 (2019).
22. A. M. Burroughs, Y. Ando, L. Aravind, New perspectives on the diversification of the RNA interference system: Insights from comparative genomics and small RNA sequencing. *Wiley Interdiscip. Rev. RNA* **5**, 141–181 (2014).
23. L. M. Iyer, E. V. Koonin, L. Aravind, Evolutionary connection between the catalytic subunits of DNA-dependent RNA polymerases and eukaryotic RNA-dependent RNA polymerases and the origin of RNA polymerases. *BMC Struct. Biol.* **3**, 1 (2003).
24. D. J. Obbard, M. Shi, K. E. Roberts, B. Longdon, A. B. Dennis, A new lineage of segmented RNA viruses infecting animals. *Virus Evol.* **6**, vez061 (2020).
25. A. Dance, Beyond coronavirus: The virus discoveries transforming biology. *Nature* **595**, 22–25 (2021).
26. A. J. W. te Velthuis, Common and unique features of viral RNA-dependent polymerases. *Cell. Mol. Life Sci.* **71**, 4403–4420 (2014).
27. Y. I. Wolf, D. Kazlauskas, J. Iranzo, A. Lucía-Sanz, J. H. Kuhn, M. Krupovic, V. V. Dolja, E. V. Koonin, Reply to Holmes and Duchêne, “can sequence phylogenies safely infer the origin of the global virome?”: Deep phylogenetic analysis of RNA viruses is highly challenging but not meaningless. *mBio* **10**, e00542–e19 (2019).
28. B. Rost, Twilight zone of protein sequence alignments. *Protein Eng.* **12**, 85–94 (1999).
29. P. M. Zlotoff, M. J. Gibbs, E. A. Gould, E. C. Holmes, A reevaluation of the higher taxonomy of viruses based on RNA polymerases. *J. Virol.* **70**, 6083–6096 (1996).
30. S. T. de Farias, T. G. Rêgo, M. V. José, tRNA Core Hypothesis for the Transition from the RNA World to the Ribonucleoprotein World. *Life (Basel)* **6**, 15 (2016).
31. K. Illergård, D. H. Ardell, A. Elofsson, Structure is three to ten times more conserved than sequence—A study of structural response in protein cores. *Proteins* **77**, 499–508 (2009).
32. 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. A. Daly, C. Desnues, M. B. Duhaime, J. B.

- Emerson, F. Enault, J. A. Fuhrman, P. Hingamp, P. Hugenholtz, B. L. Hurwitz, N. N. Ivanova, J. M. Labonté, K.-B. Lee, R. R. Malmstrom, M. Martinez-Garcia, I. K. Mizrachi, H. Ogata, D. Páez-Espino, M.-A. Petit, C. Putonti, T. Rattei, A. Reyes, F. Rodriguez-Valera, K. Rosario, L. Schriml, F. 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. C. Wrighton, P. Yilmaz, T. Yoshida, M. J. Young, N. Yutin, L. Z. Allen, N. C. Kyrpides, E. A. Elie-Fadrosh, Minimum information about an uncultivated virus genome (MIUVIG). *Nat. Biotechnol.* **37**, 29–37 (2019).
33. 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).
34. A. I. Culley, A. S. Lang, C. A. Suttle, Metagenomic analysis of coastal RNA virus communities. *Science* **312**, 1795–1798 (2006).
35. A. Culley, New insight into the RNA aquatic virosphere via viromics. *Virus Res.* **244**, 84–89 (2018).
36. S. Ghosh, Y. S. Malik, The True Host/s of Picobirnaviruses. *Front. Vet. Sci.* **7**, 615293 (2021).
37. 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, M. Babin, C. Bowler, A. I. Culley, C. de Vargas, B. E. Dutilh, D. Iudicone, L. Karp-Boss, S. Roux, S. Sunagawa, P. Wincker, M. B. Sullivan, Tara Oceans Coordinators, Marine DNA Viral Macro- and Microdiversity from Pole to Pole. *Cell* **177**, 1109–1123.e14 (2019).
38. R. K. French, E. C. Holmes, An Ecosystems Perspective on Virus Evolution and Emergence. *Trends Microbiol.* **28**, 165–175 (2020).
39. 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).
40. 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:10.25739/qhr0-8j16.
41. J. M. Wainaina, A. A. Zayed, G. Dominguez-Huerta, B. Bolduc, M. B. Sullivan, Supporting trees and alignments for the publication: Cryptic and abundant marine viruses at the evolutionary origins of Earth's RNA virome. DRYAD (2022). doi:10.5061/dryad.gb5mkkwqx.
42. J. Guo, rdrpsearch, Version 0.1. Zenodo (2021). doi:10.5281/zenodo.5731488
43. R. Schlitzer, Ocean Data View (2018); <https://odv.awi.de/>
44. A. I. Culley, A. S. Lang, C. A. Suttle, High diversity of unknown picorna-like viruses in the sea. *Nature* **424**, 1054–1057 (2003).

45. 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–e01214 (2014).
46. T. Zhang, M. Breitbart, W. H. Lee, J.-Q. Run, C. L. Wei, S. W. L. Soh, M. L. Hibberd, E. T. Liu, F. Rohwer, Y. Ruan, RNA viral community in human feces: Prevalence of plant pathogenic viruses. *PLOS Biol.* **4**, e3 (2006).
47. A. I. Culley, A. S. Lang, C. A. Suttle, The complete genomes of three viruses assembled from shotgun libraries of marine RNA virus communities. *Viol. J.* **4**, 69–78 (2007).
48. 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).
49. A. I. Culley, G. F. Steward, New genera of RNA viruses in subtropical seawater, inferred from polymerase gene sequences. *Appl. Environ. Microbiol.* **73**, 5937–5944 (2007).
50. K. Nagasaki, Dinoflagellates, diatoms, and their viruses. *J. Microbiol.* **46**, 235–243 (2008).
51. A. Djikeng, R. Kuzmickas, N. G. Anderson, D. J. Spiro, Metagenomic analysis of RNA viruses in a fresh water lake. *PLOS ONE* **4**, e7264 (2009).
52. K. Rosario, C. Nilsson, Y. W. Lim, Y. Ruan, M. Breitbart, Metagenomic analysis of viruses in reclaimed water. *Environ. Microbiol.* **11**, 2806–2820 (2009).
53. A. S. Lang, M. L. Rise, A. I. Culley, G. F. Steward, RNA viruses in the sea. *FEMS Microbiol. Rev.* **33**, 295–323 (2009).
54. K. Rosario, M. Breitbart, Exploring the viral world through metagenomics. *Curr. Opin. Virol.* **1**, 289–297 (2011).
55. P. G. Cantalupo, B. Calgua, G. Zhao, A. Hundesa, A. D. Wier, J. P. Katz, M. Grabe, R. W. Hendrix, R. Girones, D. Wang, J. M. Pipas, Raw sewage harbors diverse viral populations. *mBio* **2**, e00180–e11 (2011).
56. S. M. Short, The ecology of viruses that infect eukaryotic algae. *Environ. Microbiol.* **14**, 2253–2271 (2012).
57. B. Bolduc, D. P. Shaughnessy, Y. I. Wolf, E. V. Koonin, F. F. Roberto, M. Young, Identification of novel positive-strand RNA viruses by metagenomic analysis of archaea-dominated Yellowstone hot springs. *J. Virol.* **86**, 5562–5573 (2012).
58. T. F. F. Ng, R. Marine, C. Wang, P. Simmonds, B. Kapusinszky, L. Bodhidatta, B. S. Oderinde, K. E. Wommack, E. Delwart, High variety of known and new RNA and DNA viruses of diverse origins in untreated sewage. *J. Virol.* **86**, 12161–12175 (2012).
59. 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).
60. A. López-Bueno, A. Rastrojo, R. Peiró, M. Arenas, A. Alcamí, Ecological connectivity shapes quasispecies structure of RNA viruses in an Antarctic lake. *Mol. Ecol.* **24**, 4812–4825 (2015).

61. R. Cavicchioli, S. Erdmann, The discovery of Antarctic RNA viruses: A new game changer. *Mol. Ecol.* **24**, 4809–4811 (2015).
62. A. L. Greninger, J. L. DeRisi, Draft genome sequences of marine RNA viruses SF-1, SF-2, and SF-3 recovered from San Francisco wastewater. *Genome Announc.* **3**, e00653–e15 (2015).
63. T. Lachnit, T. Thomas, P. Steinberg, Expanding our understanding of the seaweed holobiont: RNA viruses of the red alga *Delisea pulchra*. *Front. Microbiol.* **6**, 1489 (2016).
64. T. Engelhardt, W. D. Orsi, B. B. Jørgensen, Viral activities and life cycles in deep seafloor sediments. *Environ. Microbiol. Rep.* **7**, 868–873 (2015).
65. S. R. Krishnamurthy, A. B. Janowski, G. Zhao, D. Barouch, D. Wang, Hyperexpansion of RNA bacteriophage diversity. *PLOS Biol.* **14**, e1002409 (2016).
66. J. A. Miranda, A. I. Culley, C. R. Schvarcz, G. F. Steward, RNA viruses as major contributors to Antarctic viroplankton. *Environ. Microbiol.* **18**, 3714–3727 (2016).
67. 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**, 16054 (2017).
68. 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–e16 (2017).
69. R. V. Thurber, J. P. Payet, A. R. Thurber, A. M. S. Correa, Virus-host interactions and their roles in coral reef health and disease. *Nat. Rev. Microbiol.* **15**, 205–216 (2017).
70. S.-I. Urayama, Y. Takaki, S. Nishi, Y. Yoshida-Takashima, S. Deguchi, K. Takai, T. Nunoura, Unveiling the RNA virosphere associated with marine microorganisms. *Mol. Ecol. Resour.* **18**, 1444–1455 (2018).
71. P. G. Cantalupo, J. M. Pipas, Complete genome sequence of Pittsburgh sewage-associated virus 1. *Genome Announc.* **6**, e01460–e17 (2018).
72. 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).
73. I. Hewson, K. S. I. Bistolas, J. B. Button, E. W. Jackson, Occurrence and seasonal dynamics of RNA viral genotypes in three contrasting temperate lakes. *PLOS ONE* **13**, e0194419 (2018).
74. S. Yau, M. Seth-Pasricha, Viruses of polar aquatic environments. *Viruses* **11**, 189 (2019).
75. Y.-Z. Zhang, Y.-M. Chen, W. Wang, X.-C. Qin, E. C. Holmes, Expanding the RNA virosphere by unbiased metagenomics. *Annu. Rev. Virol.* **6**, 119–139 (2019).
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).

77. M. Vlok, A. S. Lang, C. A. Suttle, Marine RNA virus quasispecies are distributed throughout the oceans. *MSphere* **4**, 1–18 (2019).
78. S. M. Short, M. A. Staniewski, Y. V. Chaban, A. M. Long, D. Wang, Diversity of Viruses Infecting Eukaryotic Algae. *Curr. Issues Mol. Biol.* **39**, 29–62 (2020).
79. J. A. Gustavsen, C. A. Suttle, Role of phylogenetic structure in the dynamics of coastal viral assemblages. *Appl. Environ. Microbiol.* **87**, e02704–e02720 (2021).
80. M. Sadeghi, Y. Tomaru, T. Ahola, RNA viruses in aquatic unicellular eukaryotes. *Viruses* **13**, 362 (2021).
81. J. Zong, X. Yao, J. Yin, D. Zhang, H. Ma, Evolution of the RNA-dependent RNA polymerase (RdRP) genes: Duplications and possible losses before and after the divergence of major eukaryotic groups. *Gene* **447**, 29–39 (2009).
82. J. Callanan, S. R. Stockdale, A. Shkoporov, L. A. Draper, R. P. Ross, C. Hill, Expansion of known ssRNA phage genomes: From tens to over a thousand. *Sci. Adv.* **6**, eaay5981 (2020).
83. S. Sunagawa, M. K. DeSalvo, C. R. Voolstra, A. Reyes-Bermudez, M. Medina, Identification and gene expression analysis of a taxonomically restricted cysteine-rich protein family in reef-building corals. *PLOS ONE* **4**, e4865 (2009).
84. L. Wang, J. Zhang, H. Zhang, D. Qiu, L. Guo, Two novel relative double-stranded RNA mycoviruses infecting *Fusarium poae* strain SX63. *Int. J. Mol. Sci.* **17**, 641 (2016).
85. J. M. Arjona-Lopez, P. Telengech, A. Jamal, S. Hisano, H. Kondo, M. D. Yelin, I. Arjona-Girona, S. Kanematsu, C. J. Lopez-Herrera, N. Suzuki, Novel, diverse RNA viruses from Mediterranean isolates of the phytopathogenic fungus, *Rosellinia necatrix*: Insights into evolutionary biology of fungal viruses. *Environ. Microbiol.* **20**, 1464–1483 (2018).
86. S. I. Urayama, Y. Takaki, D. Hagiwara, T. Nunoura, dsRNA-seq reveals novel RNA virus and virus-like putative complete genome sequences from hymeniacidon sp. *Microbes Environ.* **35**, ME19132 (2020).
87. P. J. Keeling, F. Burki, H. M. Wilcox, B. Allam, E. E. Allen, L. A. Amaral-Zettler, E. V. Armbrust, J. M. Archibald, A. K. Bharti, C. J. Bell, B. Beszteri, K. D. Bidle, C. T. Cameron, L. Campbell, D. A. Caron, R. A. Cattolico, J. L. Collier, K. Coyne, S. K. Davy, P. Deschamps, S. T. Dyrman, B. Edvardsen, R. D. Gates, C. J. Gobler, S. J. Greenwood, S. M. Guida, J. L. Jacobi, K. S. Jakobsen, E. R. James, B. Jenkins, U. John, M. D. Johnson, A. R. Juhl, A. Kamp, L. A. Katz, R. Kiene, A. Kudryavtsev, B. S. Leander, S. Lin, C. Lovejoy, D. Lynn, A. Marchetti, G. McManus, A. M. Nedelcu, S. Menden-Deuer, C. Miceli, T. Mock, M. Montresor, M. A. Moran, S. Murray, G. Nadathur, S. Nagai, P. B. Ngam, B. Palenik, J. Pawlowski, G. Petroni, G. Piganeau, M. C. Posewitz, K. Rengefors, G. Romano, M. E. Rumpho, T. Rynearson, K. B. Schilling, D. C. Schroeder, A. G. B. Simpson, C. H. Slamovits, D. R. Smith, G. J. Smith, S. R. Smith, H. M. Sosik, P. Stief, E. Theriot, S. N. Twary, P. E. Umale, D. Vaultot, B. Wawrik, G. L. Wheeler, W. H. Wilson, Y. Xu, A. Zingone, A. Z. Worden, The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): Illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. *PLOS Biol.* **12**, e1001889 (2014).

88. 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).
89. P. Skewes-Cox, T. J. Sharpton, K. S. Pollard, J. L. DeRisi, Profile hidden Markov models for the detection of viruses within metagenomic sequence data. *PLOS ONE* **9**, e105067 (2014).
90. P. Rice, I. Longden, A. Bleasby, EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet.* **16**, 276–277 (2000).
91. B. J. Walker, T. Abeel, T. Shea, M. Priest, A. Abouelliel, S. Sakthikumar, C. A. Cuomo, Q. Zeng, J. Wortman, S. K. Young, A. M. Earl, Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLOS ONE* **9**, e112963 (2014).
92. S. Nayfach, A. P. Camargo, F. Schulz, E. Eloë-Fadrosch, S. Roux, N. C. Kyrpides, CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nat. Biotechnol.* **39**, 578–585 (2021).
93. R. C. Edgar, Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461 (2010).
94. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
95. A. J. Enright, S. Van Dongen, C. A. Ouzounis, An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **30**, 1575–1584 (2002).
96. K. Katoh, D. M. Standley, MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
97. S. Capella-Gutiérrez, J. M. Silla-Martínez, T. Gabaldón, trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
98. M. F. Boni, D. Posada, M. W. Feldman, An exact nonparametric method for inferring mosaic structure in sequence triplets. *Genetics* **176**, 1035–1047 (2007).
99. S. Kalyaanamoorthy, B. Q. Minh, T. K. F. Wong, A. von Haeseler, L. S. Jermin, ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).
100. L.-T. Nguyen, H. A. Schmidt, A. von Haeseler, B. Q. Minh, IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
101. N. Grandi, M. P. Pisano, M. Demurtas, J. Blomberg, G. Magiorkinis, J. Mayer, E. Tramontano, Identification and characterization of ERV-W-like sequences in Platyrrhini species provides new insights into the evolutionary history of ERV-W in primates. *Mob. DNA* **11**, 6 (2020).

102. N. Grandi, M. Cadeddu, J. Blomberg, J. Mayer, E. Tramontano, HERV-W group evolutionary history in non-human primates: Characterization of ERV-W orthologs in Catarrhini and related ERV groups in *Platyrrhini*. *BMC Evol. Biol.* **18**, 6 (2018).
103. L. Vargiu, P. Rodriguez-Tomé, G. O. Sperber, M. Cadeddu, N. Grandi, V. Blikstad, E. Tramontano, J. Blomberg, Classification and characterization of human endogenous retroviruses: mosaic forms are common. *Retrovirology* **13**, 7 (2016).
104. M. Chen, Y. Ma, C. Yang, L. Yang, H. Chen, L. Dong, J. Dai, M. Jia, L. Lu, The combination of phylogenetic analysis with epidemiological and serological data to track HIV-1 transmission in a sexual transmission case. *PLOS ONE* **10**, e0119989 (2015).
105. B. Fernández-Caso, J. Á. Fernández-Caballero, N. Chueca, E. Rojo, A. de Salazar, L. García Buey, L. Cardeñoso, F. García, Infection with multiple hepatitis C virus genotypes detected using commercial tests should be confirmed using next generation sequencing. *Sci. Rep.* **9**, 9264 (2019).
106. A. Alipour, S. Tsuchimoto, H. Sakai, N. Ohmido, K. Fukui, Structural characterization of copia-type retrotransposons leads to insights into the marker development in a biofuel crop, *Jatropha curcas* L. *Biotechnol. Biofuels* **6**, 129 (2013).
107. X. Zhang, S. Firestein, The olfactory receptor gene superfamily of the mouse. *Nat. Neurosci.* **5**, 124–133 (2002).
108. J. J. Wiens, Missing data and the design of phylogenetic analyses. *J. Biomed. Inform.* **39**, 34–42 (2006).
109. I. Letunic, P. Bork, Interactive tree of life (iTOL) v3: An online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res.* **44** (W1), W242-5 (2016).
110. L. A. Kelley, S. Mezulis, C. M. Yates, M. N. Wass, M. J. E. Sternberg, The Phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protoc.* **10**, 845–858 (2015).
111. T. Kawabata, MATRAS: A program for protein 3D structure comparison. *Nucleic Acids Res.* **31**, 3367–3369 (2003).
112. P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, T. Ideker, Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
113. A. L. Delcher, S. L. Salzberg, A. M. Phillippy, Using MUMmer to identify similar regions in large sequence sets. *Curr. Protoc. Bioinformatics* **Chapter 10**, 3 (2003).
114. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
115. L. Guidi, S. Chaffron, L. Bittner, D. Eveillard, A. Larhlimi, S. Roux, Y. Darzi, S. Audic, L. Berline, J. 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, P. 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; Tara Oceans coordinators, Plankton networks driving carbon export in the oligotrophic ocean. *Nature* **532**, 465–470 (2016).

116. M. Steinegger, M. Meier, M. Mirdita, H. Vöhringer, S. J. Haunsberger, J. Söding, HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics* **20**, 473 (2019).
117. M. Mirdita, L. von den Driesch, C. Galiez, M. J. Martin, J. Söding, M. Steinegger, Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic Acids Res.* **45**, D170–D176 (2017).
118. B. Buchfink, C. Xie, D. H. Huson, Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).
119. T. Rognes, T. Flouri, B. Nichols, C. Quince, F. Mahé, VSEARCH: A versatile open source tool for metagenomics. *PeerJ* **4**, e2584 (2016).
120. J. Starmer, A. Stomp, M. Vouk, D. Bitzer, Predicting Shine-Dalgarno sequence locations exposes genome annotation errors. *PLOS Comput. Biol.* **2**, e57 (2006).
121. 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 (2020).
122. J. Tackmann, J. F. Matias Rodrigues, C. von Mering, Rapid inference of direct interactions in large-scale ecological networks from heterogeneous microbial sequencing data. *Cell Syst.* **9**, 286–296.e8 (2019).
123. A. Saberi, A. A. Gulyaeva, J. L. Brubacher, P. A. Newmark, A. E. Gorbalenya, A planarian nidovirus expands the limits of RNA genome size. *PLOS Pathog.* **14**, e1007314 (2018).
124. G. Salazar, L. Paoli, A. Alberti, J. Huerta-Cepas, H.-J. Ruscheweyh, M. Cuenca, C. M. Field, L. P. Coelho, C. Cruaud, S. Engelen, A. C. Gregory, K. Labadie, C. Marec, E. Pelletier, M. Royo-Llonch, S. Roux, P. Sánchez, H. Uehara, A. A. Zayed, G. Zeller, M. Carmichael, C. Dimier, J. Ferland, S. Kandels, M. Picheral, S. Pisarev, J. Poulain, S. G. Acinas, M. Babin, P. Bork, C. Bowler, C. de Vargas, L. Guidi, P. Hingamp, D. Iudicone, L. Karp-Boss, E. Karsenti, H. Ogata, S. Pesant, S. Speich, M. B. Sullivan, P. Wincker, S. Sunagawa, Tara Oceans Coordinators, Gene expression changes and community turnover differentially shape the global ocean metatranscriptome. *Cell* **179**, 1068–1083.e21 (2019).
125. Z. Yuan, X. Ye, L. Zhu, N. Zhang, Z. An, W. J. Zheng, Virome assembly and annotation in brain tissue based on next-generation sequencing. *Cancer Med.* **9**, 6776–6790 (2020).
126. A. M. Bolger, M. Lohse, B. Usadel, Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
127. D. Li, C. M. Liu, R. Luo, K. Sadakane, T. W. Lam, MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674–1676 (2015).
128. B. Nowinski, C. B. Smith, C. M. Thomas, K. Esson, R. Marin III, C. M. Preston, J. M. Birch, C. A. Scholin, M. Huntemann, A. Clum, B. Foster, B. Foster, S. Roux, K. Palaniappan, N. Varghese, S. Mukherjee, T. B. K. Reddy, C. Daum, A. Copeland, I. A.

- Chen, N. N. Ivanova, N. C. Kyrpides, T. Glavina Del Rio, W. B. Whitman, R. P. Kiene, E. A. Elie-Fadrosh, M. A. Moran, Microbial metagenomes and metatranscriptomes during a coastal phytoplankton bloom. *Sci. Data* **6**, 129 (2019).
129. B. C. Crump, J. M. Wojahn, F. Tomas, R. S. Mueller, Metatranscriptomics and Amplicon Sequencing Reveal Mutualisms in Seagrass Microbiomes. *Front. Microbiol.* **9**, 388 (2018).
130. Y. W. Chung, H.-J. Gwak, S. Moon, M. Rho, J.-H. Ryu, Functional dynamics of bacterial species in the mouse gut microbiome revealed by metagenomic and metatranscriptomic analyses. *PLOS ONE* **15**, e0227886 (2020).
131. 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).
132. D. Hyatt, G. L. Chen, P. F. Locascio, M. L. Land, F. W. Larimer, L. J. Hauser, Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).
133. R. Islam, R. S. Raju, N. Tasnim, I. H. Shihab, M. A. Bhuiyan, Y. Araf, T. Islam, Choice of assemblers has a critical impact on de novo assembly of SARS-CoV-2 genome and characterizing variants. *Brief. Bioinform.* **22**, bbab102 (2021).
134. J.-L. Zeddiam, K. H. J. Gordon, C. Lauber, C. A. F. Alves, B. T. Luke, T. N. Hanzlik, V. K. Ward, A. E. Gorbalenya, Euprosterona elaeasa virus genome sequence and evolution of the Tetraviridae family: Emergence of bipartite genomes and conservation of the VPg signal with the dsRNA Birnaviridae family. *Virology* **397**, 145–154 (2010).
135. 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).
136. 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).
137. S. Roux, J. B. Emerson, E. A. Elie-Fadrosh, M. B. Sullivan, Benchmarking viromics: An in silico evaluation of metagenome-enabled estimates of viral community composition and diversity. *PeerJ* **5**, e3817 (2017).
138. 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). *J. Phycol.* **39**, 343–352 (2003).
139. 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).
140. 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*. *Appl. Environ. Microbiol.* **70**, 704–711 (2004).

141. 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*. *Aquat. Microb. Ecol.* **34**, 207–218 (2004).
142. C. P. Brussaard, Optimization of procedures for counting viruses by flow cytometry. *Appl. Environ. Microbiol.* **70**, 1506–1513 (2004).
143. K. Nagasaki, Y. Tomaru, Y. Takao, K. Nishida, Y. Shirai, H. Suzuki, T. Nagumo, Previously unknown virus infects marine diatom. *Appl. Environ. Microbiol.* **71**, 3528–3535 (2005).
144. Y. Takao, K. Mise, K. Nagasaki, T. Okuno, D. Honda, Complete nucleotide sequence and genome organization of a single-stranded RNA virus infecting the marine fungoid protist *Schizochytrium* sp. *J. Gen. Virol.* **87**, 723–733 (2006).
145. 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. *Appl. Environ. Microbiol.* **74**, 4022–4027 (2008).
146. 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).
147. 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*. *Front. Microbiol.* **9**, 3235 (2019).
148. C. de Vargas, S. Audic, N. Henry, J. Decelle, F. Mahé, 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. Horák, O. Jaillon, G. Lima-Mendez, J. Lukeš, 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, Tara Oceans Coordinators, Eukaryotic plankton diversity in the sunlit ocean. *Science* **348**, 1261605 (2015).
149. 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. Méheust, J. Poulain, S. Romac, D. J. Richter, G. Yoshikawa, C. Dimier, S. Kandels-Lewis, M. Picheral, S. Searson, 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; Tara Oceans Coordinators, A global ocean atlas of eukaryotic genes. *Nat. Commun.* **9**, 373 (2018).
150. 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; Tara Oceans Consortium Coordinators, Open science resources for the discovery and analysis of Tara Oceans data. *Sci. Data* **2**, 150023 (2015).
151. A. Alberti, J. Poulain, S. Engelen, K. Labadie, S. Romac, I. Ferrera, G. Albin, 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, Genoscope Technical Team, Tara Oceans Consortium Coordinators, Viral to metazoan marine plankton nucleotide sequences from the Tara Oceans expedition. *Sci. Data* **4**, 170093 (2017).
152. J. Callanan, S. R. Stockdale, E. M. Adriaenssens, J. H. Kuhn, J. Rumnieks, M. J. Pallen, A. N. Shkoporov, L. A. Draper, R. P. Ross, C. Hill, *Leviviricetes*: Expanding and restructuring the taxonomy of bacteria-infecting single-stranded RNA viruses. *Microb. Genom.* **7**, 11 (2021).
153. A. J. van der Walt, M. W. van Goethem, J.-B. Ramond, T. P. Makhalanyane, O. Reva, D. A. Cowan, Assembling metagenomes, one community at a time. *BMC Genomics* **18**, 521 (2017).