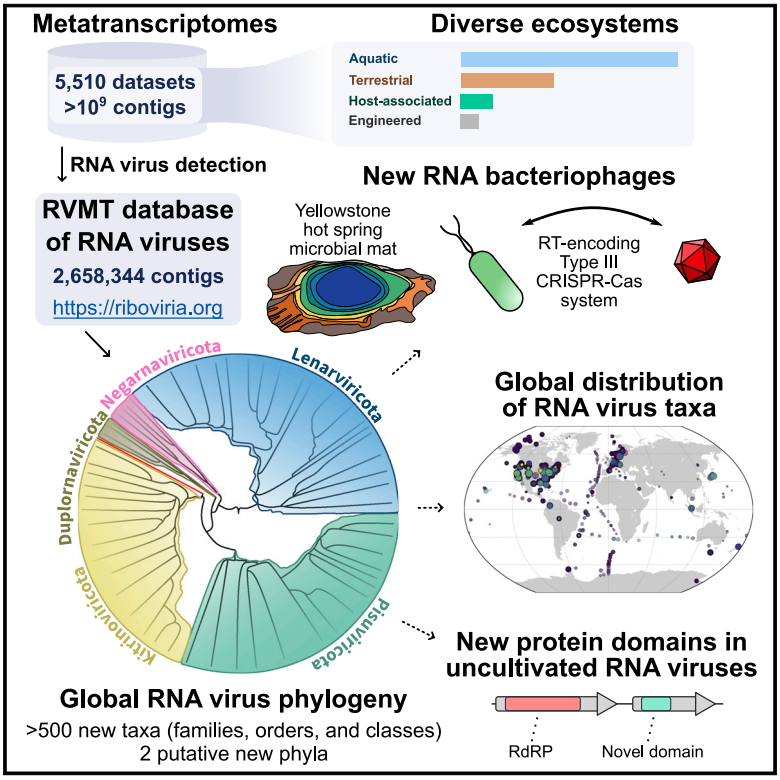
Cell

Expansion of the global RNA virome reveals diverse clades of bacteriophages

Graphical abstract



Highlights

- Metatranscriptome mining reveals a major expansion of RNA virus diversity
- A putative new phylum of RNA bacteriophages encodes distinct lysis proteins
- Partiti-like RNA phages are targeted by a bacterial CRISPR system
- Protein domains implicated in virus-host interactions are identified

Authors

Uri Neri, Yuri I. Wolf, Simon Roux, ..., Nikos C. Kyrpides, Eugene V. Koonin, Uri Gophna

Correspondence

uri.neri@gmail.com (U.N.), doljav@oregonstate.edu (V.V.D.), nckyrpides@lbl.gov (N.C.K.), koonin@ncbi.nlm.nih.gov (E.V.K.), urigo@tauex.tau.ac.il (U.G.)

In brief

Analysis of viral RNA genomes from thousands of diverse ecosystems substantially expands the known diversity of RNA viruses and show that RNA bacteriophages account for a much greater fraction of the global RNA virome.

Neri et al., 2022, Cell *185*, 4023–4037
October 13, 2022
https://doi.org/10.1016/j.cell.2022.08.023

CellPress

Cell

CellPress
OPEN ACCESS

Resource

Expansion of the global RNA virome reveals diverse clades of bacteriophages

Uri Neri,[1,12,*] Yuri I. Wolf,[2] Simon Roux,[3] Antonio Pedro Camargo,[3] Benjamin Lee,[2,4] Darius Kazlauskas,[5] I. Min Chen,[3] Natalia Ivanova,[3] Lisa Zeigler Allen,[6,7] David Paez-Espino,[3] Donald A. Bryant,[8] Devaki Bhaya,[9] RNA Virus Discovery Consortium, Mart Krupovic,[10] Valerian V. Dolja,[2,11,*] Nikos C. Kyrpides,[3,*] Eugene V. Koonin,[2,*] and Uri Gophna[1,*]

[1]The Shmunis School of Biomedicine and Cancer Research, Tel Aviv University, Tel Aviv 6997801, Israel
[2]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA
[3]Department of Energy Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA
[4]Nuffield Department of Medicine, University of Oxford, Oxford OX3 7BN, UK
[5]Institute of Biotechnology, Life Sciences Center, Vilnius University, Saulėtekio av. 7, Vilnius 10257, Lithuania
[6]Microbial and Environmental Genomics, J. Craig Venter Institute, La Jolla, CA, USA
[7]Marine Biology Research Division, Scripps Institution of Oceanography, La Jolla, CA, USA
[8]Department of Biochemistry and Molecular Biology, The Pennsylvania State University, University Park, PA 16802, USA
[9]Department of Plant Biology, Carnegie Institution for Science, Stanford, CA 94305, USA
[10]Institut Pasteur, Université Paris Cité, CNRS UMR 6047, Archaeal Virology Unit, 75015 Paris, France
[11]Department of Botany and Plant Pathology, Oregon State University, Corvallis, OR 97331, USA
[12]Lead contact
*Correspondence: uri.neri@gmail.com (U.N.), doljav@oregonstate.edu (V.V.D.), nckyrpides@lbl.gov (N.C.K.), koonin@ncbi.nlm.nih.gov (E.V.K.), urigo@tauex.tau.ac.il (U.G.)
https://doi.org/10.1016/j.cell.2022.08.023

SUMMARY

High-throughput RNA sequencing offers broad opportunities to explore the Earth RNA virome. Mining 5,150 diverse metatranscriptomes uncovered >2.5 million RNA virus contigs. Analysis of >330,000 RNA-dependent RNA polymerases (RdRPs) shows that this expansion corresponds to a 5-fold increase of the known RNA virus diversity. Gene content analysis revealed multiple protein domains previously not found in RNA viruses and implicated in virus-host interactions. Extended RdRP phylogeny supports the monophyly of the five established phyla and reveals two putative additional bacteriophage phyla and numerous putative additional classes and orders. The dramatically expanded phylum *Lenarviricota*, consisting of bacterial and related eukaryotic viruses, now accounts for a third of the RNA virome. Identification of CRISPR spacer matches and bacteriolytic proteins suggests that subsets of picobirnaviruses and partitiviruses, previously associated with eukaryotes, infect prokaryotic hosts.

INTRODUCTION

Viruses are obligate intracellular parasites of living organisms and are regarded as the most numerous biological entities on Earth (Mushegian, 2020). Historically, only viruses causing disease in humans, livestock, and crops along with model bacterial viruses (phages) have been studied in detail. Recently, a previously unsuspected diversity of DNA viruses has been identified, thanks to advances in genome sequencing and metagenomics (Call et al., 2021; Roux et al., 2021). Recognizing metagenomics role in virus discovery, the International Committee for Taxonomy of Viruses (ICTV) approved formal recognition of new virus taxa on the basis of metagenomic sequence analysis (Simmonds et al., 2017).

Compared with DNA viruses, the diversity and role of RNA viruses in microbial ecosystems is poorly understood. Recently, however, metatranscriptome surveys (bulk RNA sequencing of entire microbial communities) uncovered massive amounts of previously undetected RNA viruses (Krishnamurthy et al., 2016; Zeigler Allen et al., 2017; Dolja and Koonin, 2018). In particular, analysis of invertebrate transcriptomes resulted in doubling the number of known RNA viruses (Shi et al., 2016), followed by a further 2-fold expansion through analysis of the RNA sequences in the metavirome (sequencing of the subcellular size fraction) from a single site, implying a vast, barely sampled global RNA virome (Wolf et al., 2020). Other forays into RNA viromes include analysis of fungal transcriptomes (Sutela et al., 2020), metatranscriptomes of various types of soil (Starr et al., 2019; Wu et al., 2021), and expansion of the RNA phageome of aquatic environments (Callanan et al., 2020).

Apart from deltaviruses, all RNA viruses share a single hallmark protein, the RNA-dependent RNA polymerase (RdRP) (Koonin et al., 2020). Thus, study of the diversity and evolution of RNA viruses hinges on detection and analysis of RdRPs. Although due to the extreme sequence divergence of the RdRPs, the confidence in the deepest branchings in the phylogenetic tree is low, five well-separated, major clades were identified (Wolf et al., 2018; Holmes

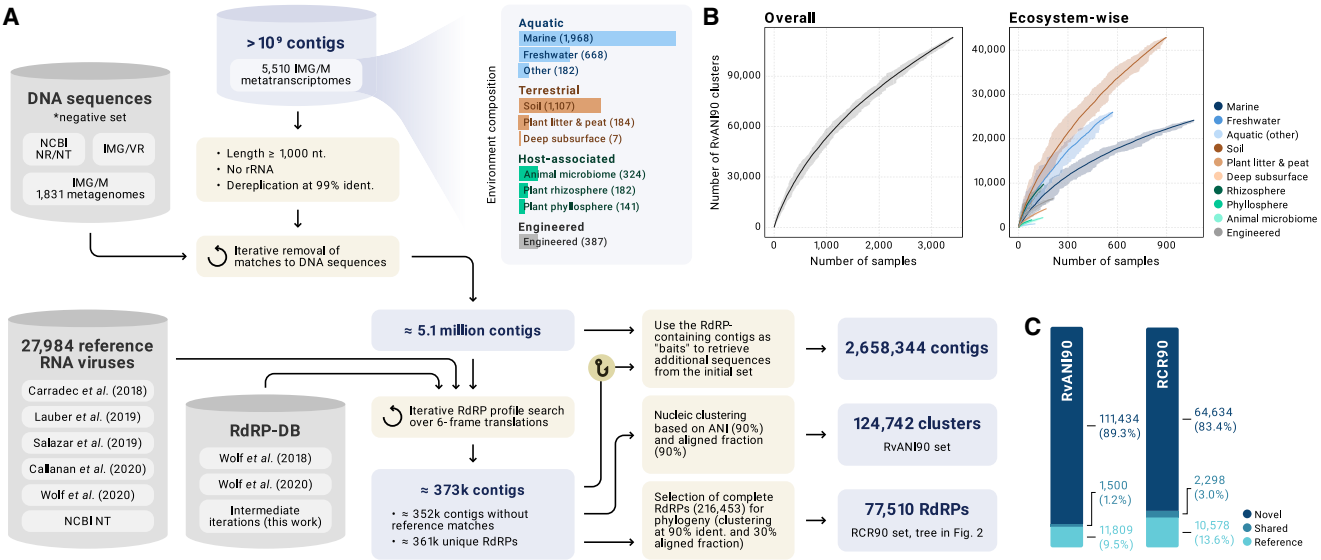


Figure 1. RNA virus discovery pipeline

(A) RNA virus discovery pipeline.

(B) RvANI90 Rarefaction curves: accumulation of unique clusters as a function of the number of analyzed samples (GOLD field—ITS.PIDs). These values were obtained via bootstrapping; semi-opaque segments represent the range of measured unique RvANI90 clusters across 25 random subsamplings. The central line represents the mean of 25 random samples. Colors indicate the environment type (right chart).

(C) Number of RCR90 clusters (left) and RvANI90 (right), whose members are either entirely "reference" (contigs from the "reference set" only), "novel" (only identified in the analyzed metatranscriptomes), or "shared" (contains members of each type).

See also Figure S1.

and Duchêne, 2019) and subsequently recognized as phyla comprising the kingdom *Orthornavirae* within the realm *Riboviria* (International Committee on Taxonomy of Viruses Executive Committee, 2020; Koonin et al., 2020).

Clearly, an extensive census of RNA virus genomes from diverse habitats and hosts is crucial for understanding RNA virus evolution. Here, mining 5,150 metatranscriptomes from various environments, we expanded RNA virus diversity from 13,282 to 124,873 distinct clusters at a granularity level between species and genus. We identified two candidate additional phyla and numerous tentative classes, orders, and families. These include unreported lineages likely infecting bacteria. Additionally, we report multiple unexpected protein domains, some of which are likely to counter antiviral defense.

RESULTS

Identification of RNA viruses from diverse metatranscriptomes

Here, we devised a computational pipeline for sensitive RNA virus detection suitable for analysis of thousands of metatranscriptomes (Figure 1; see STAR Methods). Briefly, the pipeline first filters out sequences likely encoded by DNA entities by comparing the metatranscriptomic contigs to a diverse set of DNA genomes and metagenomes. Subsequently, the much reduced sequence set (<1% of the initial set) is iteratively searched for RdRPs, and confident matches are treated as putative RNA viruses (see STAR Methods). 3,598 of the 5,150 metatranscriptomes queried, contained one or more contigs coding an RdRP of sufficient completeness for further analyses (see STAR Methods). We then used the RdRP-encoding contigs as bait to identify additional metatranscriptomic contigs sharing high nucleic similarity with the RdRP-encoding ones (including outside of the RdRP region). Altogether, 2,658,344 RNA virus contigs were identified and supplemented with 27,984 sequences from published sources (Figure 1A). Of these, 348,762 contigs represented a deduplicated, non-redundant sequence set of length $\geq$1 kbp. These were grouped into 124,743 clusters sharing 90% average nucleotide identity (RNA Virus ANI90 clusters [hereafter RvANI90]), of which only 13,308 (10.7%) contained at least one previously known sequence, translating into a roughly 9-fold expansion of the global RNA virome, at the ANI90 level of diversity.

The RNA virus sequence clusters showed a power law-like distribution by size, dominated by small clusters, with a long tail of large clusters, the largest one including 429 contigs (Figure S1). Based on the accumulation curve, the global diversity of RNA viruses evaluated at the RvANI90 level showed no sign of saturation (Figure 1B), with a particularly high richness in soil environments (Figure 1B). About 5.8% of the RdRP-encoding contigs showed evidence of utilizing alternative genetic codes (Figure 2), and about 0.5% showed shuffling of the conserved motifs (domain permutation) within the RdRP (Figure 2).

RdRP phylogeny and major expansion of RNA virus diversity

To build a global RNA virus phylogeny, we first collected full-length RdRP core domain sequences and clustered them at 90% amino acid identity threshold, arriving at 77,510

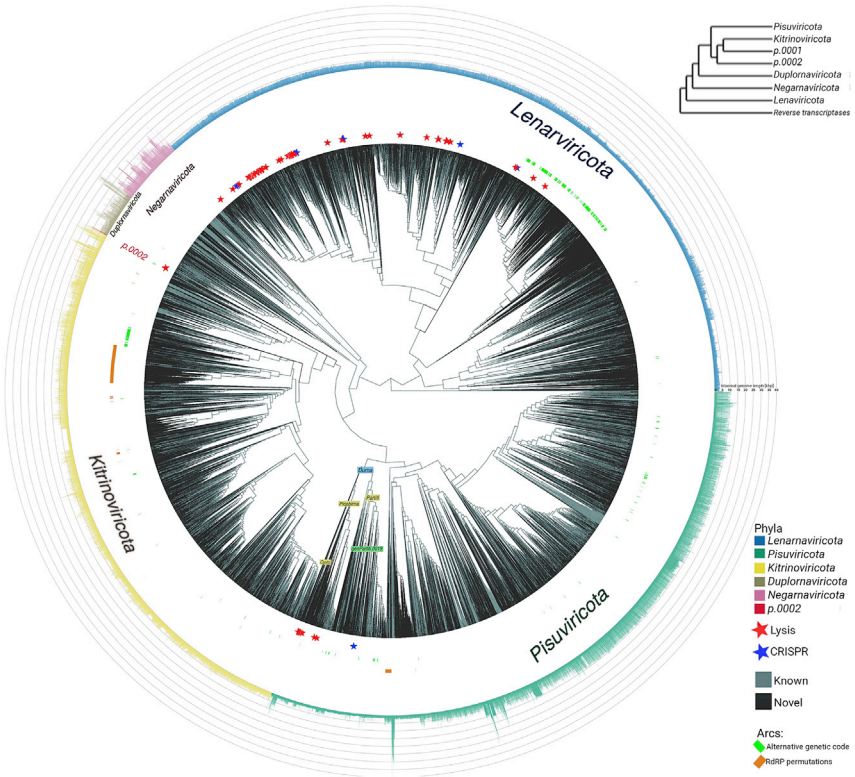


Figure 2. Phylogenetic reconstruction of the global RNA virosphere

An ultrametrized RdRP tree rooted using reverse transcriptases as an outgroup and visualized with ggtree and ggtreeExtra (Xu et al., 2021; Yu et al., 2018). Branches are colored black unless any of their descendants contain at least one sequence from the "reference set" (cyan). Tips aligned with stars indicate evidence of prokaryotic host—CRISPR spacer match in blue and bacteriolytic domain in red. Green arcs indicate clades with an alternative genetic code in ≥50% of the sequences. Orange arcs indicate clades with motif permutation in ≥50% of the RdRPs. The 5 established phyla and the proposed candidate phylum p.0002 are color coded in both the text and the bar-plot in the outermost ring, which represents the maximum genome length observed for each RCR90 cluster (i.e., tree tip). Key taxa are labeled directly on the tree. Additional visualizations of the tree are available in the project's Zenodo repository (see data and code availability).
See also Figure S2.

representatives (RCR90 set). Even when reduced to the RCR90 granularity, the set remained too large and diverse to be directly amenable for multiple sequence alignment and phylogenetic analysis with advanced maximum likelihood phylogenetic methods. Therefore, we employed an iterative procedure in which the tree was reconstructed using an alignment of consensuses of sequence cluster alignments (see STAR Methods). The resulting RdRP tree comprised 77,520 representative sequences (77,510 RCR90 sequences and 10 reverse transcriptases [RTs] included as an outgroup; Figure 2). Despite this dramatic expansion, the 5 previously established phyla (Wolf et al., 2018) remained largely monophyletic. In addition, the tree included two groups below the base of the phylum *Kitrinoviricota*, which were analyzed in detail (see below).

Monophyly of the major branches in the RdRP tree, in particular the 5 phyla, was verified by subsampling. Representatives of virus families were repeatedly randomly sampled, phylogeny was reconstructed from the multiple alignment of each sample, the positions of the phyla clades were traced, and a quantitative measure of their monophyly was calculated (see STAR Methods). In most of the samples, the 5 phyla stayed largely monophyletic (Figure S2A). Sequences that tended to break the phylum-level monophyly formed a sharply biased subset, with *Flasuviricetes* being the most common "offender." In this work, *Flasuviricetes* was placed inside *Pisuviricota*, whereas in previous analyses, it was the basal clade of *Kitrinoviricota*. Nevertheless, the inconsistent position of flaviviruses in subsampled trees indicates that their phylogenetic placement remains uncertain. The families *Reoviridae*, *Picobirnaviridae*, *Cystoviridae*, and several candidate families also often broke away from their respective phyla, although the consensus tree placed *Picobirnaviridae* and *Cystoviridae* confidently within *Pisuviricota* (see below).

When the subsampled trees were reduced to the lowest common ancestor of each of the five phyla, the deepest branching order was found to be robust, with *Pisuviricota* and *Kitrinoviricota* forming a crown group in the consensus tree, and *Lenarviricota* and *Negarnaviricota* occupying basal positions (Figure 2, top right inset). As in previous analyses (Wolf et al., 2018), when the tree was rooted by RTs, the deepest branch within *Orthornavirae* was the phylum *Lenarviricota* that includes leviviruses (positive-sense RNA phages; class *Allassoviricetes*) and their apparent direct descendants among the viruses of eukaryotes, mitoviruses (*Howeltoviricetes*), narnaviruses (*Amabilivirecetes*), and botourmiaviruses (*Miaviricetes*). Although validating this branching order definitively may not be feasible, this position of *Lenarviricota* is biologically plausible, placing the origin of *Orthornavirae* in the bacterial domain. In contrast, the deep placement of *Negarnaviricota* was unexpected, given that -ssRNA viruses have been isolated almost exclusively from animals and plants. *Negarnaviricota* position might reflect an ancient origin, but more likely, is a phylogenetic artifact, perhaps caused by acceleration of evolution at the base of *Negarnaviricota*.

Comparison of the phylogenetic depths of the present RdRP phylogeny and the previously reported tree (Wolf et al., 2020) reflected a roughly 5-fold expansion of the global RNA virome as measured by the total-branch-length (TBL). To convert the RdRP phylogeny into a tentative taxonomic scheme, we developed a semi-quantitative approach for assigning taxonomic ranks to unclassified nodes based on neighboring well-established taxa (see STAR Methods). Taxa were designated to rank and prefixed by *p*, *c*, *o*, *f*, and *g* for phylum, class, order, family, and genus, respectively, followed by an ordinal number for proposed taxa of

Table 1. Expansion of the global RNA virome

Rank	Number of known taxa	Updated number of taxa	Fold increase
RvANI90 cluster	13,282	124,873	9.4
RCR90 cluster	12,862	77,510	6.0
Family	98	489	4.9
Order	26	121	4.7
Class	19	93	4.9
Phylum	5	7	1.4

that rank. Taxa that associated with a previously described taxon were terminated with "base," e.g., *f.0127.base-Noda* is the 127th new family that is basal to *Nodaviridae* in the RdRP tree (Table S1).

This approach resulted in a roughly 5-fold expansion of diversity at all ranks below phylum, compared with the results of the latest RNA virome analysis (Wolf et al., 2020; Table 1). However, it has to be emphasized that this estimate was obtained without taking into account the results of two large-scale RNA virus surveys published since this analysis was performed (see limitations of the study section) (Edgar et al., 2022; Zayed et al., 2022).

When broken down by phyla, the largest expansion at all ranks was within *Lenarviricota*, followed by *Kitrinoviricota* and *Pisuviricota*. By contrast, only a few taxa were added to *Duplornaviricota* and *Negarnaviricota* (Figure 2; Table S1).

In addition to the expansion reflected in the RdRP phylogenetic tree, some of the RNA viruses (39,000 contigs that formed 24,742 RvANI90 clusters) identified in this work via the RdRP-based profile searches were discarded from the phylogenetic analysis as the boundaries and some of the motifs of the core RdRP domain could not be reliably identified.

Putative additional phyla and classes

As there is currently no official guidance from the ICTV for the formation of RNA virus phyla and classes, we opted for criteria similar to the ones used for shallower ranks (see STAR Methods), that is, to form a phylum or class, a group was required to branch outside of the existing phyla or classes. Two of the most divergent clades identified here were positioned below the base of *Kitrinoviricota* in the RdRP phylogeny and, in principle, can be included in an expanded version of this phylum. The first of these deep branches, *p.0001*, included only 3 RCR90 clusters and therefore was not analyzed further. The second one, *p.0002*, possess distinct features that appear more compatible with a candidate phylum designation rather than expansion of *Kitrinoviricota*. This putative phylum consisted of 234 contigs from 30 RCR90 clusters, the most complete ones encoding ~10 ORFs with mean length of about 12 kb. Except the RdRP, only one of the ORFs (conserved in one of the two tentative families in *p.0002*) had significant similarity to a known protein domain, specifically to M15 or M35 family of zinc metallopeptidases implicated in cell lysis (see below). The ORFs in *p.0002* genomes are tightly spaced and preceded by ribosome-binding motifs (Shine Dalgarno [SD]) involved in prokaryotic translation initiation (Figure 3A). Taken together, *p.0002* appears to consist of bacteriophages, supporting the group's phylum designation as all isolated *Kitrinoviricota* members infect eukaryotes.

Another highly divergent candidate RNA phage phylum was RvANI90_0011770, one of the viral clusters omitted from the phylogeny effort as they distorted the RdRP alignment (hence, no *p* designation). All RvANI90_0011770 members originated from 27 different active sludge samples, where the largest of these 55 contigs were 10–12-kb long, encoding 7–9 closely spaced ORFs with no conserved SD motifs. Similarly to p.0002, the only recognized protein domains included the RdRP and a predicted lysis enzyme (see below).

A substantial increase in class-level diversity (see STAR Methods) was observed in 4 of the 5 established phyla, including 14 classes versus 4 known in *Lenarviricota*, 18 classes over the 4 known in *Pisuviricota*, 20 classes versus 3 known in *Kitrinoviricota*, and 18 classes versus 6 known in *Negarnaviricota*. In *Duplornaviricota*, only two candidate class-level clades were identified in addition to the two recognized classes. Overall, the 5 phyla of *Orthornavirae* contained 91 classes compared with the 19 previously established ones and 489 families compared with the previously recognized 98 (Table 1; Table S1). Some of these additional candidate taxa included previously reported, divergent viruses that so far eluded placement and lacked ICTV designation.

Major expansion of the range of RNA viruses associated with bacteria

So far, most RNA viruses have been associated with eukaryotic hosts, with only two groups known to infect bacteria, leviviruses (*Leviviricetes*), and cystoviruses (*Vidaverviricetes*). Until recently, leviviruses and particularly cystoviruses, included small numbers of viruses with narrow host ranges. Here, we expand *Cystoviridae* diversity from the 8 published RCR90 clusters to 132 RCR90 clusters. Levivirus diversity, which was recently expanded (Callanan et al., 2020) to 1,940 RCR90 clusters, was further increased here by an additional 13,512 RCR90 clusters.

The expanded phylum *Lenarviricota* now accounts for over a third of the RNA virus RCR90 clusters, including the four largest families (Figure 2; Table S1), the first and fourth of these, *Steitzviridae* and *Fiersviridae*, respectively, are bona fide *Leviviricetes* phages. The second-largest family, *Botourmiaviridae*, consists of eukaryotic viruses that appear to have evolved from a common ancestor with *Leviviricetes*, with the capsid-less *Narnaviridae* and *Mitoviridae* (the third largest family of RNA viruses) as intermediates (Koonin et al., 2020). In addition to the major expansion of *Lenarviricota*, converging lines of evidence suggested reassignment to bacterial hosts for several groups of viruses previously thought to solely infect eukaryotes (Figure 3B). Phages now appear to be interspersed with those infecting eukaryotes within *Pisuviricota*. Specifically, the family *Cystoviridae*, which migrated from *Duplornaviricota* to *Pisuviricota* in the current RdRP phylogeny, forms a strongly supported branch with picobirnaviruses and partitiviruses (double-stranded RNA [dsRNA] families embedded in the midst of the +ssRNA viruses [Figure 2]). Within this *Durnavirales* order, several clades showed unexpected conservation of SD motifs in the 5′ untranslated regions (UTRs), suggesting that these viruses infect bacteria (Bahiri Elitzur et al., 2021; Hockenberry et al., 2018). These putative phages include members of *Picobirnaviridae*, for which presence

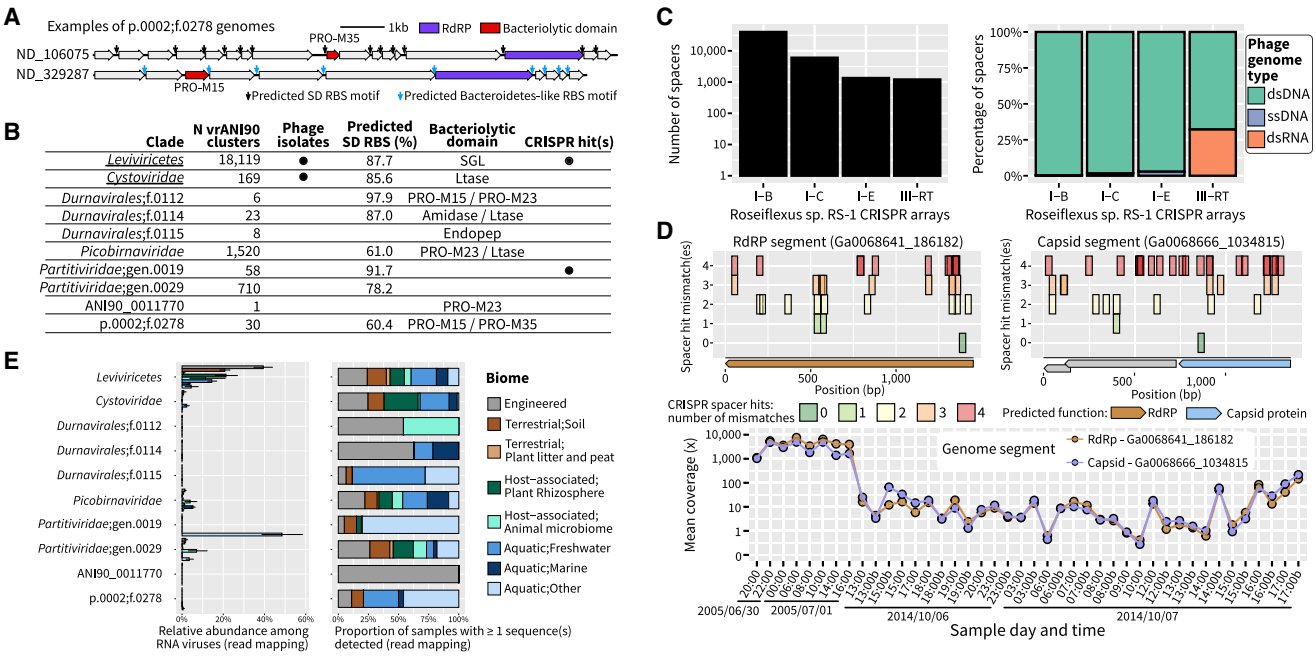


Figure 3. Diversity and abundance of prokaryotic RNA viruses

(A) Genome map of viruses from the tentative family *f.0278* of the proposed phylum *p.0002*. ORFs are colored based on functional annotation and predicted SD motifs are indicated with colored arrows.

(B) Overview of recognized (underlined) and predicted prokaryotic RNA viruses. For each group, the type of evidence supporting its association with prokaryotic hosts is indicated. Clades were considered as likely composed primarily of phages if ≥50% of the predicted ORFs were associated with an SD motif, after excluding genes predicted on the edges of contigs. Ltase, lytic transglycosylase, lysozyme superfamily fold; SGL, "single-gene lysis" (cell wall synthesis inhibitors); PRO-M15, Zn-DD-carboxypeptidase (sensu PF08291.13); PRO-M35, M35 family zinc metalloendopeptidase; PRO-M23, M23-family metallopeptidases; Amidase, N-acetylmuramoyl-L-alanine amidase; Endopep, L-alanyl-D-glutamate endopeptidase.

(C) CRISPR spacer landscape of *Roseiflexus* sp. RS-1 in Yellowstone hot springs, including spacers matching *genPartiti.0019* genomes. Left panel displays the total number of spacers identified for each type of Roseiflexus sp. RS-1 CRISPR arrays (see Figure S3). The right panel presents phage type (dsDNA, ssDNA, or RNA) for which hits to CRISPR spacers were identified for each CRISPR type.

(D) Example of a predicted pair of RdRP and capsid-encoding segments from a *genPartiti.0019* phage. Top panel: CRISPR spacer matches are indicated alongside a genomic map for each segment. The number of mismatches is shown on the y axis, and the position of the hit is indicated on the x axis. The bottom panel displays the relative abundance of both segments across a metatranscriptome time series.

(E) Relative abundance of different prokaryotic RNA virus groups across biomes. Only datasets dominated by prokaryotic sequences ("P-dominated") containing at least 10 prokaryotic RNA viruses were considered. The right panel shows a breakdown of the biome distribution for each group, calculated from a balanced dataset composed of random subsamples of 50 samples per environment (random subsampling was performed 100 times, and the mean values were plotted). See also Figures S3 and S4.

of SD motifs was previously noted (Boros et al., 2018; Krishnamurthy and Wang, 2018), along with two cysto-like families (*f.0114.base-Cysto* and *f.0112.base-Cysto*) and two additional genera within *Partitiviridae* (genPartiti.0029, genPartiti.0019.-base-Deltapartitivirus) (Table S2; Figure 3B).

Another evidence of bacterial association for some of the identified viral groups is the conserved occurrence of bacteriolytic proteins (Figure 3B). Many dsDNA Phages and dsRNA cystoviruses encode lytic enzymes (endolysins) degrading bacterial peptidoglycan (Cahill and Young, 2019). In contrast, leviviruses induce host lysis by inhibiting peptidoglycan synthesis via small proteins termed single-gene lysis (Sgl) (Cahill and Young, 2019). Leviviruses *sgl* are typically overlapping or nested within other genes (Chamakura and Young, 2020). Here, we used a collection of such lysis domains to detect metatranscriptomic viral genomes potentially infecting bacteria (see STAR Methods) (Figure 3B). This search yielded 546 significant matches to lysis protein profiles, mostly in *Leviviricetes* (469) and *Cystoviridae* (17). Although known cystoviruses encode lytic transglycosylases of the lysozyme superfamily (SF) fold (Dessau et al., 2012), some of the cysto-like families identified here encoded other peptidoglycan-digesting enzymes. Specifically, some *f.0114.base-Cysto* viruses encode N-acetylmuramoyl-L-alanine amidases, whereas viruses of *f.0112.base-Cysto* encoded metallopeptidases of the M15 or M23 families (Table S2), both often found in dsDNA phages and are known to cleave bonds of cross-linking peptides (Oliveira et al., 2013). Some *f.0112.base-Cysto* viruses also encoded lipases that may further induce host lysis. Finally, *f.0115.base-Cysto* viruses encoded an L-alanyl-D-glutamate endopeptidase that commonly functions as endolysins in dsDNA phages (Cahill and Young, 2019; Oliveira et al., 2013). This clade-specific distribution of endolysins in cystoviruses indicates that, as in dsDNA phages, lysis genes are subject to frequent non-homologous replacement, potentially linked to host range change.

Two other groups of RNA viruses were found to encode lysis proteins, picobirnaviruses and family *f.0278* in the proposed phylum *p.0002*. Six picobirnaviruses encode either lytic transglycosylases or M23-family metallopeptidases. Members of *f.0278* encode either M15 or M35 family zinc metallopeptidases (Table S2). M15 family enzymes are involved in host lysis in some dsDNA phages (Kutyshenko et al., 2021) and in some ssDNA bacteriophages (Roux et al., 2012), whereas M35-family enzymes have not been previously linked to phage egress. Given that the two enzymes are mutually exclusive in *f.0278* and the corresponding genes occupy equivalent positions, we propose that M15 and M35 family proteins function as endolysins. The conservation of M15 and M35 proteins in *f.0278* strongly supports bacterial host assignment. Finally, RvANI90_0011770, a putative phylum of RNA bacteriophages identified by the RdRP searches not included in the present phylogeny, showed similar conservation of M23-family metallopeptidases.

The final line of evidence for prokaryotic host assignment was the detection of matches between RNA viruses and CRISPR spacers. Although most known CRISPR systems target DNA templates, a large subset of type III CRISPR systems encode RT and can protect bacteria against RNA bacteriophages (Makarova et al., 2020; Silas et al., 2017). We compared all identified RNA virus genomes with the IMG database of $\geq$50 million spacers (see STAR Methods), detecting spacer matches for 161 RNA viruses from 23 RvANI90 clusters, across two clades: *Leviviricetes*, and *genPartiti.0019* (Figure 3B; Table S2). All matches to *Leviviricetes* viruses were from short contigs derived from IMG metagenomes, with no reliable taxonomic information or adjacent *cas* genes (Table S3). By contrast, matches to *genPartiti.0019* viruses were specifically associated with populations of *Roseiflexus* sp. RS-1 and were further analyzed. This filamentous anoxygenic phototrophic bacterium of the phylum *Chloroflexi* is a dominant member of microbial mats in Mushroom Spring (Davison et al., 2016), from which the *genPartiti.0019* sequences were obtained. The genome of *Roseiflexus* sp. RS-1 contains four CRISPR loci, with one subtype III-B encoding a RT fused to the Cas1 protein (see Figure S3) (van der Meer et al., 2010). Compiling spacers across 16 metagenomes, each of the CRISPR arrays could be associated with $\approx$1,000–40,000 spacers, yet all but one spacers matching *genPartiti.0019* sequences were detected in the RT-encoding III-B array, suggesting that these were acquired from RNA templates (Figure 3C). These CRISPR spacer matches were observed in samples spanning 9 years and showed dynamic spacer gain/loss through time, indicative of virus-host association (Figure S3).

Because all *genPartiti.0019* contigs encoded RdRP alone, whereas related partitiviruses have segmented genomes, where the capsid and other proteins are encoded in separate segments, we searched the Mushroom Spring metatranscriptomes for contigs encoding the corresponding capsid proteins (CPs). Combining matches to spacers from the RT-encoding type III-B array of *Roseiflexus* sp. RS-1, the absence of corresponding sequences in the Mushroom Springs DNA metagenome, and strong relative abundance correlation (>0.9) to at least one *genPartiti.0019* RdRP-encoding sequence, we identified 88 potential capsid-encoding contigs (Figure 3D; Table S3), of which 86 encoded proteins with best alignment to HMM profiles of known partitiviruses capsids (Figure S3). Thus, *genPartiti.0019* members are most likely segmented RNA phages infecting *Roseiflexus* sp. RS-1.

Interestingly, in datasets dominated by prokaryotic hosts (“P-dominated,” see below), most potential RNA phages were detected across a broad range of biomes, where *Leviviricetes* was by far the most abundant group of prokaryotic RNA viruses, except in some Yellowstone hot springs dominated by *genPartiti.0019* (Figure 3E).

Differential distribution of RNA viruses across samples and habitats

Our RNA virus survey spanned the entire globe, reflecting the ubiquity of RNA viruses on Earth (Figure 4A). Metagenomic studies have shown that DNA virus distribution is shaped by the environment type and host community composition (Gregory et al., 2019; Martinez-Hernandez et al., 2017; Roux et al., 2016), and the same factors likely determine the RNA virus distribution. For metatranscriptomes, the sample processing protocol can be another factor, namely, whether the total RNA was sequenced or whether any specific preprocessing was used (such as mRNAs enrichment via poly(A) amplification, or rRNAs depletion) (Gann et al., 2021). Here, most of the datasets analyzed were rRNA-depleted (67%, Figure S4). Although the poly(A)-enriched and total RNA datasets were dominated by eukaryotic sequences, the rRNA-depleted datasets consisted mostly of sequences from prokaryotes (Figure S4). The datasets were separated into three groups: “Eukaryote(E)-dominated” (811), “Prokaryote(P)-dominated” (2,706), and “Mixed” (452), based on the taxonomic composition of non-viral contigs. Most RNA virus classes showed clear distribution patterns across dataset types and environments, likely reflecting the distribution of their primary host groups (Figures 4B and 4E). For instance, *Leviviricetes* were consistently enriched in P-dominated samples from engineered, rhizosphere, and soil habitats (Figure 4B). This implies an uneven global ecological distribution of RNA phages, supporting previous findings (Callanan et al., 2020). Also among *Lenarviricota*, *Miaviricetes* which infect mostly fungi, invertebrates, and plants were associated with E-dominated and Mixed datasets, whereas *Howeltoviricetes* members, including mitoviruses, were common in all sample types but found preferentially in plant-associated datasets also rich in fungi.

Although assigning specific eukaryote hosts to RNA viruses is a challenging task not addressed in this work, we suspect that many of the detected viruses infect diverse unicellular eukaryotes, as they utilize alternative genetic code (see below). Assuming that the broad host assignment (plants, animals, or fungi) of viruses can be extended over minor sequence dissimilarity (less than 10%), we identified only 1,038 metatranscriptomic contigs that belonged to the same RvANI90 cluster as viruses from VirusHostDB (Mihara et al., 2016) assigned to plant or animal hosts, indicating low prevalence of viruses infecting these hosts in the analyzed datasets. Additionally, specific host assignment to plants can be made for 1,038 metatranscriptomic contigs (in 6 families: *Tombusviridae, Virgaviridae, Betaflexiviridae, Alphaflexiviridae, Benyviridae,* and *Mayoviridae*), encoding movement proteins (MPs), which enable viruses to pass through plasmodesmata.

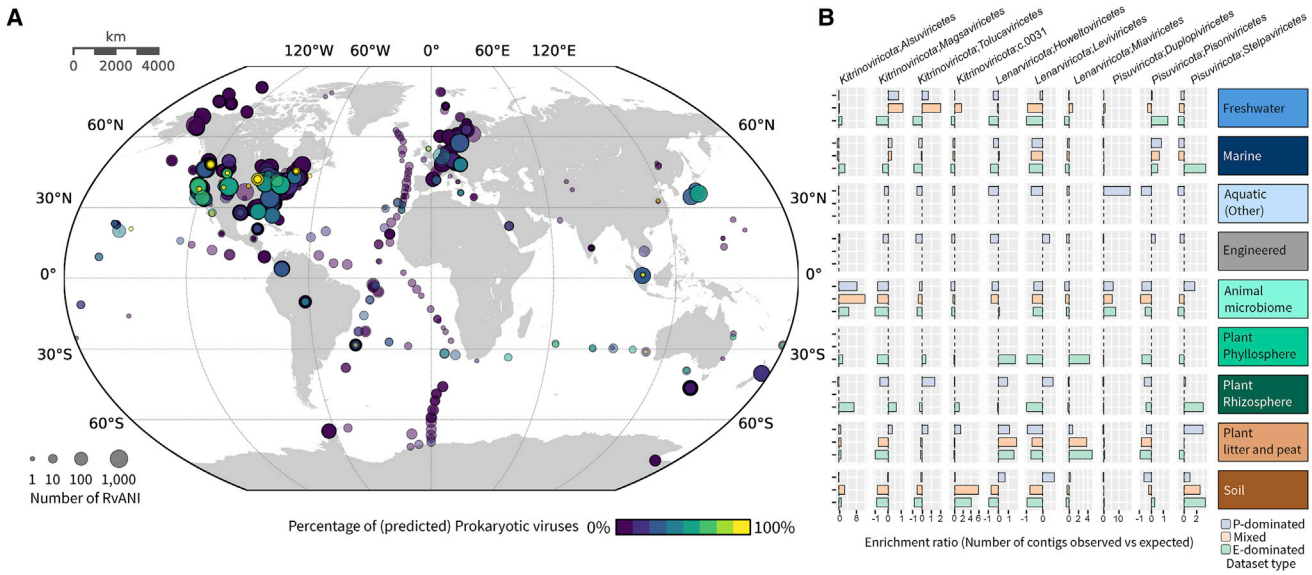


Figure 4. Global distribution of RNA viruses

(A) locations of analyzed samples containing RNA viruses. For each sample, the circle size reflects the number of distinct RvANI90, and the circle color indicates the proportion of sequences predicted as phages.

(B) Relative proportion of (proposed) RNA virus classes (x axis) detected across ecosystem types (y axis). To take into account the total number of genomes detected for each class and the total number of samples for each ecosystem type, the counts are represented as enrichments compared with the expected number of genomes assuming even distribution of all classes across all ecosystems. The datasets were divided into "E-dominated" (mostly composed of eukaryotic transcripts), "P-dominated" (mostly composed of prokaryotic transcripts), and "Mixed" (see Figure S4). Enrichments are shown only for combinations of ecosystem and dataset type (e.g., "Marine P-dominated") for which at least 20 metatranscriptomes with $\geq$1 RNA virus were detected.

See also Figure S4.

Modular evolution of RNA virus genomes

Here, we performed a comparative analysis of viral genomes from related clades, identifying instances of genomic modularity, such as fusion of genome segments, rearrangement of proteins, and segmentation of polyproteins. Common genomic rearrangements involving the structural module were observed in *Picornavirales*, where CPs were encoded both downstream or upstream of the genome replication module, as part of the same polyprotein or as separate proteins (Figure S5, Genome maps). Known viruses of *Benyviridae*, *Picobirnaviridae*, and *Botourmiaviridae* typically encode the CP and RdRP on different segments. Here, we identified members of these families where the RdRP and CP are on the same segment.

We detected multiple cases of structural gene module displacement by non-homologous counterparts. For instance, although members of *Potyviridae*, *Benyviridae*, and *Matonaviridae* encode 3 unrelated CPs and form helical filamentous, rod-shaped, or enveloped virus particles, respectively, some of the lineages branching near these viruses encode single jelly roll (SJR) CPs expected to form non-enveloped icosahedral virions. Given this lineage basal position, SJR CPs were likely ancestral in all three virus groups. In the *f.0226.base-Beny* group, several viruses encode both SJR and tobacco mosaic virus (TMV)-like CPs that can be predicted to form icosahedral and helical capsids, respectively (Figure S5), suggesting these viruses probably acquired the second CP yet retained the ancestral one. Exaptation of one of the CPs appears likely, as previously described for closteroviruses (Dolja et al., 2006). Non-homologous CPs were also identified in lineages basal to *Togaviridae* (*f.0271.base-Toga* and *f.0273.base-Toga*), where the typical *Togaviridae* icosahedral forming CPs were replaced by TMV-like CPs, likely forming rod-shaped helical virions, suggesting TMV-like CPs emerged in a common ancestor of *Hepelivirales* and *Martellivirales*. Conversely, in two identified *Virgaviridae* contigs (ND_191857 and ND_019381), the TMV-like CP was replaced by structural proteins of *Kitaviridae*. In *f.0268.base-Toga*, the typical *Togaviridae* structural module (including genes for CP and class II fusion [CIIF] protein) was replaced by a class I fusion protein and M protein of nidoviruses (ND_164660; Figure 5). Similar replacement of a membrane fusion glycoprotein was also identified in *Xinmoviridae* contigs, where CIIF protein replaced the typical class III fusion protein, yet retaining the typical mononegaviral nucleocapsid protein.

We identified several virus groups basal to *Hypoviridae* (capsidless mycoviruses) encoding CPs homologous to those of flexible helical viruses (*f.0066.base-Hypo*) or SJR CPs of icosahedral viruses (*f.0067.base-Hypo*, *f.0068.base-Hypo*, *f.0069.base-Hypo*), suggesting these families ancestor is likely capsid encoding. Similarly, we identified *Deltaflexiviridae* relatives encoding SJR CPs (ND_196199 and ND_246366 from *f.0215.base-Deltaflexi*) similar to those of tymoviruses, suggesting that *Deltaflexiviridae* evolved from a member of *Tymoviridae* following the switch to fungal hosts. The recurrent appearance of the SJR CP in base lineages of several groups of structurally diverse viruses is compatible with the proposed origin of most RNA viruses of eukaryotes from a simple ancestor that encoded RdRP and SJR CP (Koonin et al., 2020).

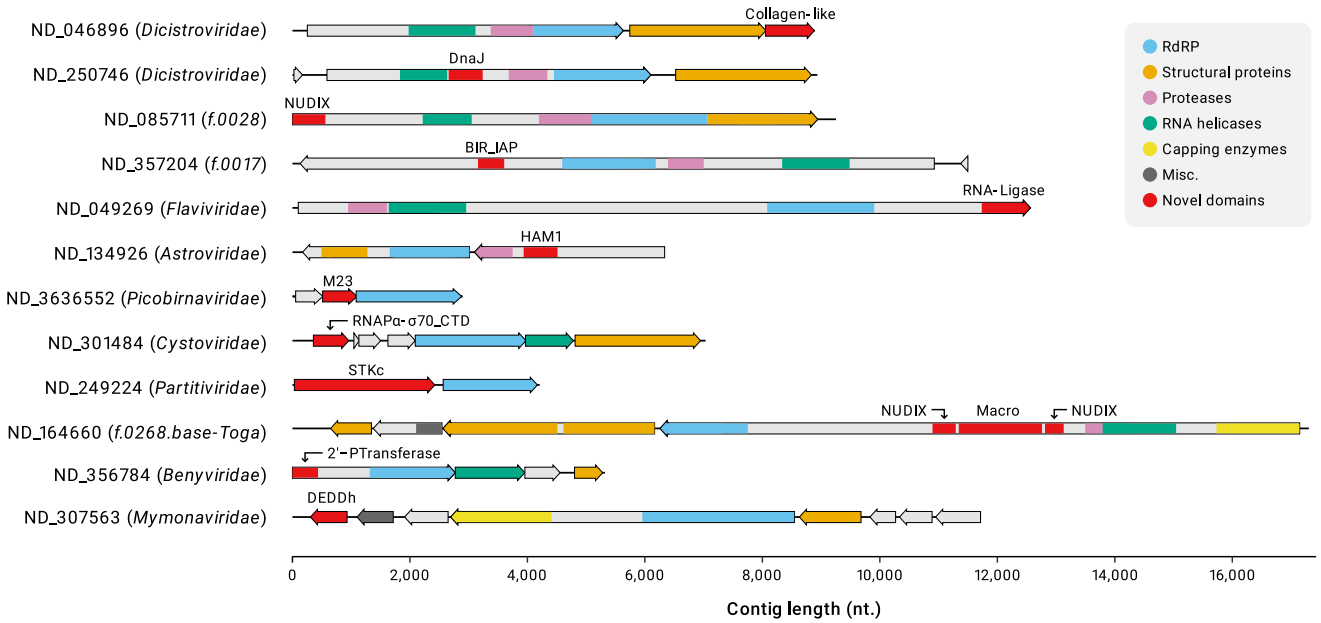


Figure 5. Diversity of protein domains in RNA viruses

Homologous domains are shown as boxes of the same color (see key on the right). Domains not commonly found in RNA viruses are shown in red and are labeled above the corresponding boxes. Virus taxa and contig identifiers are noted to the left of each virus genome. At the bottom, scale indicating the length in nucleotides. Abbreviations: NUDIX, nucleoside diphosphate-X hydrolase; BIR_IAP, baculoviral IAP repeat (BIR) domains of the inhibitor of apoptosis (IAP); HAM1, inosine triphosphate pyrophosphatase; M15, M23, and M34, peptidoglycan-digesting peptidases of MEROPS families M15, M23, and M34, respectively; RNAPα-σ70_CTD, a fusion of C-terminal domain of the α subunit of bacterial DNA-dependent RNA polymerase and C-terminal domain of σ70 factor; STKc, serine/threonine protein kinase; 2′-PTransferase, tRNA 2′-phosphotransferase; DEDDh, DEDDh-superfamily 3′-5′ exonuclease; RdRP, RNA-dependent RNA pol polymerase; Misc, miscellaneous.
See also Figures S5, S6, and S7.

Originally described in *Permutotetraviridae* and *Birnaviridae*, a distinct rearrangement (known as "domain permutation") occurs within the RdRP domain, where the order of the motifs (A, B, C.) differs from the canonical form. Here, ~2.9% of the RCR90 RdRP set (2,241) were identified as permuted. Our analysis suggests that motif swapping was ancestral in two classes (Figure 2), candidate class *c.0017* in *Pisuviricota* (which includes *Permutotetraviridae*, *Birnaviridae*, and 14 other tentative families [*f.0088-f.0101*]) and candidate class *c.0032* in *Kitrinoviricota* (covering 8 putative families [*f.0167–f.0174*], including many viruses from the Yangshan assemblage [Wolf et al., 2020]). Outside of *Pisuviricota* and *Kitrinoviricota*, we detected only a small clade consisting of 2 permuted RCR90 RdRPs within *Botourmiaviridae* (*Lenarviricota*).

Expansion of the protein domain repertoire of RNA viruses

Here, we annotated the identified viruses via an extensive search for protein domains (see STAR Methods and Figure S3). In line with previous studies (Wolf et al., 2020), the frequencies of the detected domains followed a power law-like distribution, where most domains only occurred in specific viral groups (Figure S7). Of the few hallmark domains that were widespread across the RNA viral tree, the most ubiquitous was the RdRP, followed by different types of CPs (CP_SJR, CP_levi), RNA helicases (SF1, SF2, SF3), and serine/cysteine proteases (Figure S7). Apart from the aforementioned lysis domains, we identified several domains predicted to modulate virus-host interactions and suppress the host antiviral response.

Several *Tobaniviridae* members, which primarily infect vertebrates, encoded homologs (HHpred p = 100%) (Zimmermann et al., 2018) of the cytokine receptor-associated Janus kinase (JAK) TYK2, which upon activation triggers host immune responses (Haan et al., 2006). These viral JAKs lacked the FERM and SH2 domains of typical TYK2 and may function as dominant negative inhibitors of the cellular JAKs via their pseudokinase domain. The only other RNA viruses predicted to encode a serine/threonine kinase are partitiviruses (Figure 5), although that kinase is unrelated to JAKs. Members of *f.0059*.base-Poty and *f.0167* families encode homologs of cytokine receptors of the tumor necrosis factor receptor SF, known to be involved in apoptosis and inflammation (Gravestein and Borst, 1998). The viral homologs may act as decoys of the host counterparts, sequestering the cytokines.

Some *Dicistroviridae* members, and several lineages basal to *Solinviviridae* (*f.0024.base-Solinvi*, *f.0014.base-Solinvi*, *f.0017.base-Solinvi*, *f.0018.base-Solinvi*) and *Polycipiviridae* (*f.0008.base-Polycipi*), contained homologs of baculoviral IAP repeat (BIR) domain (baculovirus inhibitor of apoptosis), known to function in cell cycle control and death (Clem, 2015).

Nucleoside diphosphate-X hydrolase (NUDIX) SF hydrolases are common in all domains of life and in dsDNA viruses (Vasudevan and Ryoo, 2015). Here, we identified NUDIX hydrolases in 13 different RNA virus families (*Flaviviridae*, *Nodaviridae*,

Cystoviridae, and several candidate families). Apart from the bacteria-infecting *Cystoviridae*, we suspect these RNA virus-encoded NUDIX hydrolases function like those of dsDNA viruses, acting as decapping enzymes promoting shutoff of host protein synthesis (Kago and Parrish, 2021).

In 11 diverse RNA virus families from the phyla *Kitrinoviricota* and *Pisuviricota*, we identified the J domain, the active moiety of DnaJ (Hsp40) co-chaperone (Laudenbach et al., 2021). In these viruses, the J domain is part of the virus polyproteins and might facilitate polyprotein folding and processing and/or virion assembly.

We also identified several enzymatic domains implicated in RNA repair and metabolism, including RtcB-like 3′-phosphate RNA ligase (Hughes et al., 2020), HAM1-like pyrophosphatase (Simone et al., 2013), DEDD-SF 3′-5′ exonuclease that could be involved in immune suppression, as in arenaviruses (Hastie et al., 2011), and tRNA 2′-phosphotransferase implicated in tRNA splicing (Sawaya et al., 2005). In cellular organisms, the latter enzyme is often encoded with NAD and ADP-ribose (NADAR) domain proteins implicated in NAD metabolism, in the context of RNA processing (de Souza and Aravind, 2012). NADAR domains have been originally detected in *Roniviridae* (+ssRNA viruses) and giant dsDNA viruses (de Souza and Aravind, 2012). We identified NADAR domains in RNA viruses from 12 families, emphasizing the potential importance of this domain for RNA virus replication.

In certain cystoviruses, we detected a protein with an N-terminal domain homologous to the C-terminal domain (CTD) of sigma70 factors (a subunit of the bacterial RNA polymerase holoenzyme, that directs the RNA polymerase to specific promoters; Paget and Helmann, 2003). The CTD of this cystoviral protein is similar to the C-terminal region of bacterial RNA polymerase alpha subunit. The CTDs of sigma70 and RNA polymerase alpha are known to interact (Chen et al., 2003), suggesting that this cystoviral protein reconstitutes this interaction interface and may participate in transcriptional takeover during infection, potentially overcoming the host antiviral defenses.

The identification of these diverse domains in RNA viruses of one or several lineages implies multiple mechanisms of virus-host interaction and, in particular, counter-defense, which remain to be investigated.

Alternative genetic codes in RNA viruses

Previous surveys identified several RNA virus groups utilizing non-standard genetic codes, suggesting they infect hosts with matching codes, such as ciliates (Wolf et al., 2020). Here, of the 77,510 RCR90 representatives, 5,843 (~7.5%) showed evidence of alternative genetic codes, indicated by the presence of canonical STOP codons within the RdRP core domain coding region (see STAR Methods). Although in most cases, it is impossible to identify the specific alternative code, of the cases where it was feasible, the most common codes were 6 (UAA and UAG coding Gln) and 14 (UAA and UGA coding for Tyr and Trp, respectively, along with recoding of three sense codons) that have been identified in ciliates (Ring and Cavalcanti, 2008) and flatworm mitochondria (Ross et al., 2016), respectively. Unlike many DNA viruses that use alternative genetic codes which actively reprogram the host cell's translation machinery for their benefit (Ivanova et al., 2014; Yutin et al., 2021), in RNA viruses, such codes are likely to represent adaptation to the host translation machinery. This phenomenon is well-known for multiple isolated mitoviruses that use the mitochondrial genetic code (UGA recoded from stop to Trp) and replicate inside mitochondria (Nibert, 2017), and indeed, ~51% of the viruses within the much expanded (2,553 of 5,006 RCR90) *Mitoviridae* use code 4 that is common in fungal mitochondria. Apart from mitoviruses, contigs with alternative genetic codes were detected in most of the large RNA virus groups, typically at frequencies of a few percent. We identified virus lineages enriched (>50%) in such codes, throughout the phylogenetic tree of the RdRPs (Table S5, green arcs in Figure 2). No coherent phylogenetic signal of alternative code was detected in *Duplornaviricota* and *Negarnoviricota*. Contrastingly, we detected 19 families of *Pisuviricota that* typically contained one or two small branches (8–30 RCR90 members) with apparent protist codes (UAA and/or UAG code for an amino acid). *Dicistroviridae* (monopartite +ssRNA arthropod viruses) stood out with 12 such branches, suggesting some of these dicistroviruses may be protists infecting, potentially arthropod-associated ones. Finally, in *Kitrinoviricota*, we observed a surprising distribution of alternative codes: 7 families included small branches with alternative codes, whereas 7 other families consisted exclusively (*f.0150, f.0177–f.181*) or primarily (*f.0176*) of viruses using alternative protist-like codes. In line with previous findings (Wolf et al., 2020), the present analysis suggests *Kitrinoviricota* includes a substantial, previously unsuspected, diversity of protist viruses.

DISCUSSION

Metagenomes and metatranscriptomes have become the principal sources of DNA and RNA virus discovery, respectively (Call et al., 2021; Simmonds et al., 2017). Here, we analyzed more than 2.5 million RNA virus contigs recovered from 3,598 diverse metatranscriptomes. Metatranscriptome analysis is prone to artifacts that stem, in particular, from chimeric RNA assemblies. Therefore, it is important to emphasize that all conclusions of this work are based on analysis of evolutionary conserved groups of RNA virus sequences, and not singletons, under the assumption that appearance of the same chimera in different assemblies is highly unlikely; several other safeguards against chimeric assemblies were implemented (see STAR Methods).

Our analysis resulted in a 9-fold-increase in the number of 90% RvANI clusters (between the species and genus ranks), a 5-fold increase in the total phylogenetic depth, an almost 6-fold increase in the number of representative RdRP sequences (RCR90), and a 5-fold increase in the number of putative taxa at the levels from family to class. In contrast, at the phylum level, the RNA virus taxonomy remained essentially stable, with the exception of adding two candidate phyla to the previously established 5.

Most of the previous assignments of RNA virus families to phyla remained stable, albeit with notable exceptions. Thus, *Cystoviridae* expanded by an order of magnitude and relocated from *Kitrinoviricota* to *Pisuviricota*, where it now forms a strongly supported clade with other dsRNA viruses, picobirnaviruses,

and partitiviruses. Given the greater reliability of phylogenetic analysis with the expanded family and the plausibility of the monophyly of these three groups of dsRNA viruses, the current position of *Cystoviridae* is likely to be valid. However, the placement of several other families, notably, *Flaviviridae*, was unstable. Although this family also moved from *Kitrinoviricota* to *Pisuviricota*, in this case, the actual affiliation remains uncertain.

Classification of the kingdom *Orthornavirae* into phyla appears to be robust, but the resolution of the phylogeny of the RdRP near the root might be insufficient to decipher the relationship among the phyla. The previously proposed scenario of the origin of dsRNA viruses from within positive-sense RNA viruses on multiple independent occasions and of negative-sense RNA viruses (*Negarnaviricota*) from the *Duplornaviricota* (Wolf et al., 2018) remains biologically plausible. However, phylogenetic analysis of the expanded set of RdRPs failed to vindicate this scenario in its entirety, although multiple origins of dsRNA viruses were supported. The basal position of *Negarnaviricota* observed here, albeit robust to the performed tests, most likely, is an artifact of deep phylogenetic analysis. In contrast, the basal position of *Lenarviricota* in the tree rooted with RT likely reflects the origin of the rest of the RNA viruses from a common ancestor with this phylum within the bacterial domain. This scenario appears particularly plausible considering the major expansion of the bacterial RNA virome in this work. Considering the size and diversity of the analyzed dataset, it appears likely that the information contained in the RdRP sequences is indeed insufficient to resolve the deepest relationships among RNA viruses. This problem will merit revisiting once sufficient diversity of RdRP structures accumulates, possibly, providing for a better phylogenetic resolution.

The present analysis eliminates the long-standing bias in the RNA virome toward eukaryote-infecting viruses (Koonin et al., 2015). Apart from the major expansion of the diversity of levi-like viruses, we obtained indications that multiple additional groups of viruses infect bacteria—in particular, picobirnaviruses and several clades of partitiviruses. A key line of evidence supporting this possibility is the discovery of numerous CRISPR spacers targeting RNA viruses, both members of *Leviviricetes* and a group of candidate RNA phages within partitiviruses.

The present results strongly suggest that drastic host shifts, known as horizontal virus transfer (HVT), between distantly related hosts, even crossing the prokaryote-eukaryote divide, is a major route of RNA virus evolution (Dolja and Koonin, 2018). The HVT events likely occurred on multiple, independent occasions within different phyla, classes, and possibly even orders of RNA viruses. In that regard, the small group of viruses, for which multiple CRISPR spacer matches were detected and that therefore was tentatively assigned to the *Roseiflexus* bacterium as the host, is notable. This narrow virus group from a unique habitat, likely, a genus, is lodged deeply within partitiviruses, many of which are known to infect fungi, plants, and invertebrates (Cross et al., 2020; Shi et al., 2016; Vainio et al., 2018).

In addition to the major expansion of the global RNA virome, this work also substantially expands the catalog of protein domains encoded in RNA virus genomes. The common theme among these domains that are each represented in narrow lineages of RNA viruses appears to be counter-defense via diverse molecular mechanisms. These findings imply that, despite their typically smaller genomes, RNA viruses are more similar to DNA viruses with respect to the exaptation of host genes than previously appreciated (Koonin *et al.*, 2022).

In summary, the results greatly expand the diversity of the kingdom *Orthornavira*, in particular that of RNA viruses associated with bacteria, while introducing relatively minor changes into the latest taxonomic scheme, supporting its general robustness. Additionally, multiple protein functionalities were predicted in RNA viruses. The large amount of sequence and derivative data generated in this work is available through the companion website (riboviria.org) or via the Zenodo deposit. We expect this resource to enable researchers to gain meaningful and comprehensive context when describing new RNA viruses in future studies, for example, by offering insights into the ecological distribution of specific viral lineages or via the clade-specific protein domain annotations. Furthermore, this resource can help researchers identify key RNA virus genomes to be characterized experimentally.

Limitations of the study

Our approach to the detection of RNA viruses relied heavily on the presence of an RdRP via profile searches that can miss extremely distant homologs with altered canonical sequence motifs. Furthermore, several RNA viruses possess "split" RdRPs, where the motifs are encoded in different ORFs or even genomic segments (Sutela et al., 2020; Chiba et al., 2021). Another drawback of our RdRP-based discovery is the lack of a systematic identification effort for segmented RNA virus genomes (as the non-RdRP coding segments would be unreported). Presently, genomic segments other than that encoding the RdRP were identified by co-occurrence analysis only for the group of bacteria-infecting partiti-like viruses targeted by CRISPR. Comprehensive detection of segmented RNA virus genomes is a task for future analyses, as is the assignment of different segments to each other/specific viral genome.

Two studies conducted concurrently with this work generated related insights. A large-scale survey of RNA-sequencing archives reporting numerous novel RNA viruses has been published by the Serratus team (Edgar et al., 2022), and a large-scale metatranscriptome analysis of oceanic RNA viruses has been published by the *Tara* Oceans project (Zayed et al., 2022). A comprehensive comparison of the results of the three studies that differed in many methodological aspects, including the scope of the analyzed metatranscriptomes, remains a major task for the future. However, to quantify the overlap among the results of the three projects and accordingly assess the novelty of each, we performed an automated comparison of RdRP clusters obtained with two clustering thresholds, 0.9 for fine grain and 0.5 for coarse grain classification (see STAR Methods).

The results of this comparison (Table S8 "cluster intersections") detected relatively small numbers of clusters shared by all three projects and showed that thousands of clusters were unique to each. At fine grain (threshold of 0.9), the greatest number of unique clusters was identified in the Serratus data, as could be expected given that this project included a considerably larger data set than the other two. However, at coarse grain

(threshold of 0.5), our present results included more unique clusters than the other two studies taken together, indicating that our work covers a substantially greater phylogenetic depth of RNA viruses. This comparison supports our conclusion that the current sampling of the global RNA virome is far from reaching saturation. Thus, the three studies are complementary, and incorporation of the results into a single phylogenomic framework and synthesis of the conclusions should substantially advance our knowledge of the RNA virosphere.

CONSORTIA

The RNA Virus Discovery Consortium members are Adrienne B. Narrowe, Alexander J. Probst, Alexander Sczyrba, Annegret Kohler, Armand Séguin, Ashley Shade, Barbara J. Campbell, Björn D. Lindahl, Brandi Kiel Reese, Breanna M. Roque, Chris DeRito, Colin Averill, Daniel Cullen, David A.C. Beck, David A. Walsh, David M. Ward, Dongying Wu, Emiley Eloe-Fadrosh, Eoin L. Brodie, Erica B. Young, Erik A. Lilleskov, Federico J. Castillo, Francis M. Martin, Gary R. LeCleir, Graeme T. Attwood, Hinsby Cadillo-Quiroz, Holly M. Simon, Ian Hewson, Igor V. Grigoriev, James M. Tiedje, Janet K. Jansson, Janey Lee, Jean S. VanderGheynst, Jeff Dangl, Jeff S. Bowman, Jeffrey L. Blanchard, Jennifer L. Bowen, Jiangbing Xu, Jillian F. Banfield, Jody W. Deming, Joel E. Kostka, John M. Gladden, Josephine Z. Rapp, Joshua Sharpe, Katherine D. McMahon, Kathleen K. Treseder, Kay D. Bidle, Kelly C. Wrighton, Kimberlee Thamatrakoln, Klaus Nusslein, Laura K. Meredith, Lucia Ramirez, Marc Buee, Marcel Huntemann, Marina G. Kalyuzhnaya, Mark P. Waldrop, Matthew B. Sullivan, Matthew O. Schrenk, Matthias Hess, Michael A. Vega, Michelle A. O'Malley, Monica Medina, Naomi E. Gilbert, Nathalie Delherbe, Olivia U. Mason, Paul Dijkstra, Peter F. Chuckran, Petr Baldrian, Philippe Constant, Ramunas Stepanauskas, Rebecca A. Daly, Regina Lamendella, Robert J. Gruninger, Robert M. McKay, Samuel Hylander, Sarah L. Lebeis, Sarah P. Esser, Silvia G. Acinas, Steven S. Wilhelm, Steven W. Singer, Susannah S. Tringe, Tanja Woyke, T.B.K. Reddy, Terrence H. Bell, Thomas Mock, Tim McAllister, Vera Thiel, Vincent J. Denef, Wen-Tso Liu, Willm Martens-Habbena, Xiao-Jun Allen Liu, Zachary S. Cooper, and Zhong Wang

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - Metatranscriptome acquisition
  - Primary Filtering process
  - Secondary Filtering process
  - Estimation of DNA remnants in intermediate sets
  - RdRP identification
  - Identification of the RdRP catalytic motifs A–D
  - Correction of putative frameshifts
  - Contig set augmentation with published genomes
  - Comprehensive identification of RNA virus contigs across metatranscriptomes
  - Phylogenetic reconstruction
  - Taxonomic affiliation of clades
  - Robustness of deep phylogeny
  - Assignment of individual contigs to RCR90 clusters
  - Identification of reliable CRISPR spacer hits
  - Habitat distribution and relative abundance estimation
  - Genetic code assignment and ORF calling
  - RBS identification and quantification
  - Domain annotation
  - Quality control and reliability of metatranscriptomic assemblies
  - Quantitative comparison with recently published RNA virus discovery endeavors
- QUANTIFICATION AND STATISTICAL ANALYSIS
- ADDITIONAL RESOURCES

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cell.2022.08.023.

ACKNOWLEDGMENTS

The authors would like to thank Shai Zilberzwige-Tal, David Burstein, Adi Stern, Leah Reshef, and Omry Lieber for helpful discussions. U.G. and U.N. are supported by the European Research Council (ERC-AdG 787514). U.N. is supported by a fellowship from the Edmond J. Safra Center for Bioinformatics at Tel Aviv University. Y.I.W. and E.V.K. are supported through the Intramural Research Program of the US National Institutes of Health (National Library of Medicine). V.V.D. was partially supported by NIH/NLM/NCBI Visiting Scientist Fellowship. The work of the U.S. Department of Energy Joint Genome Institute (S.R., A.P.C., I.M.C., N.I., D.P.-E., N.C.K., and all JGI co-authors), a DOE Office of Science User Facility, is supported by the Office of Science of the U.S. Department of Energy under contract no. DE-AC02-05CH11231. M.K. was supported by l'Agence Nationale de la Recherche grants ANR-20-CE20-009-02 and ANR-21-CE11-0001-01. D.K. was funded by the European Social Fund under no. 09.3.3-LMT-K-712-14-0027. D.A.B. is supported by grant NNX16SJ62G from the NASA Exobiology program, and by grant DE-FG02-94ER20137 from the Photosynthetic Systems Program, Division of Chemical Sciences, Geosciences, and Biosciences (CSGB), Office of Basic Energy Sciences of the U.S. Department of Energy. We gratefully acknowledge the contributions of many scientists and principal investigators, who sent extracted genetic material for isolate genomes, environmental metagenomes, and metatranscriptomes, or sequencing results as part of the Department of Energy Joint Genome Institute Community Science Program and allowed us to include in our study the RNA virus sequences detected in these publicly available data sets regardless of publication status.

AUTHOR CONTRIBUTIONS

U.G., E.V.K., V.V.D., and N.C.K. conceptualized and supervised the project. U.N., U.G., Y.I.W., and S.R. designed the discovery pipeline. Y.I.W. performed the phylogenetic analyses. U.N., Y.I.W., and S.R. performed the host assignment predictions. S.R. performed the habitat and ecological distribution analyses. A.P.C., U.N., D.K., and M.K. performed the protein domain analyses. U.N., S.R., Y.I.W., and A.P.C. performed the sequence clustering. S.R., D.A.B., and D.B. analyzed the Yellowstone hot springs assemblies and the Roseiflexus samples. U.N. and B.L. constructed the companion website. I.M.C., N.I., D.P.-E., and N.C.K. contributed to data and metadata gathering and curation in the IMG database. L.Z.A. contributed to the ecological and

protist analysis. U.N., S.R., V.V.D., N.C.K., U.G., Y.I.W., A.P.C., E.V.K., and M.K. wrote the manuscript, which was edited and approved by all authors.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: February 15, 2022
Revised: May 16, 2022
Accepted: August 24, 2022
Published: September 28, 2022

REFERENCES

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. *25*, 3389–3402. https://doi.org/10.1093/nar/25.17.3389.

Andreeva, A., Howorth, D., Chothia, C., Kulesha, E., and Murzin, A.G. (2014). SCOP2 prototype: a new approach to protein structure mining. Nucleic Acids Res. *42*, D310–D314. https://doi.org/10.1093/nar/gkt1242.

Andreeva, A., Kulesha, E., Gough, J., and Murzin, A.G. (2020). The SCOP database in 2020: expanded classification of representative family and superfamily domains of known protein structures. Nucleic Acids Res. *48*, D376–D382. https://doi.org/10.1093/nar/gkz1064.

Arroyo Mühr, L.S., Lagheden, C., Hassan, S.S., Kleppe, S.N., Hultin, E., and Dillner, J. (2020). De novo sequence assembly requires bioinformatic checking of chimeric sequences. PLoS One *15*, e0237455. https://doi.org/10.1371/journal.pone.0237455.

Attwood, T.K., Coletta, A., Muirhead, G., Pavlopoulou, A., Philippou, P.B., Popov, I., Romá-Mateo, C., Theodosiou, A., and Mitchell, A.L. (2012). The PRINTS database: a fine-grained protein sequence annotation and analysis resource–its status in 2012. Database (Oxford) *2012*, bas019. https://doi.org/10.1093/database/bas019.

Bahiri Elitzur, S., Cohen-Kupiec, R., Yacobi, D., Fine, L., Boaz, A., Diament, A., and Tuller, T. (2021). Prokaryotic rRNA-mRNA interactions are involved in all translation steps and shape bacterial transcripts. RNA Biol *18*, 684–698. https://doi.org/10.1080/15476286.2021.1978767.

Bickhart, D.M., Kolmogorov, M., Tseng, E., Portik, D.M., Korobeynikov, A., Tolstoganov, I., Uritskiy, G., Liachko, I., Sullivan, S.T., Shin, S.B., et al. (2022). Generating lineage-resolved, complete metagenome-assembled genomes from complex microbial communities. Nat Biotechnol *40*, 711–719. https://doi.org/10.1038/s41587-021-01130-z.

Boros, Á., Polgár, B., Pankovics, P., Fenyvesi, H., Engelmann, P., Phan, T.G., Delwart, E., and Reuter, G. (2018). Multiple divergent picobirnaviruses with functional prokaryotic Shine-Dalgarno ribosome binding sites present in cloacal sample of a diarrheic chicken. Virology *525*, 62–72. https://doi.org/10.1016/j.virol.2018.09.008.

Buchfink, B., Xie, C., and Huson, D.H. (2015). Fast and sensitive protein alignment using DIAMOND. Nat. Methods *12*, 59–60. https://doi.org/10.1038/nmeth.3176.

Bushnell, B. (2014). BBTools software package. https://sourceforge.net/projects/bbmap/.

Cahill, J., and Young, R. (2019). Phage lysis: multiple genes for multiple barriers. Adv. Virus Res. *103*, 33–70. https://doi.org/10.1016/bs.aivir.2018.09.003.

Call, L., Nayfach, S., and Kyrpides, N.C. (2021). Illuminating the virosphere Through global metagenomics. Annu. Rev. Biomed. Data Sci. *4*, 369–391. https://doi.org/10.1146/annurev-biodatasci-012221-095114.

Callanan, J., Stockdale, S.R., Shkoporov, A., Draper, L.A., Ross, R.P., and Hill, C. (2020). Expansion of known ssRNA phage genomes: From tens to over a thousand. Sci. Adv. *6*, eaay5981. https://doi.org/10.1126/sciadv.aay5981.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). BLAST+: architecture and applications. BMC Bioinformatics *10*, 421. https://doi.org/10.1186/1471-2105-10-421.

Carradec, Q., Pelletier, E., Da Silva, C., Alberti, A., Seeleuthner, Y., Blanc-Mathieu, R., Lima-Mendez, G., Rocha, F., Tirichine, L., Labadie, K., et al. (2018). A global ocean atlas of eukaryotic genes. Nat. Commun. *9*, 373. https://doi.org/10.1038/s41467-017-02342-1.

Chamakura, K.R., Tran, J.S., O'Leary, C., Lisciandro, H.G., Antillon, S.F., Garza, K.D., Tran, E., Min, L., and Young, R. (2020). Rapid de novo evolution of lysis genes in single-stranded RNA phages. Nat. Commun. *11*, 6009. https://doi.org/10.1038/s41467-020-19860-0.

Chamakura, K.R., and Young, R. (2020). Single-gene lysis in the metagenomic era. Curr. Opin. Microbiol. *56*, 109–117. https://doi.org/10.1016/j.mib.2020.09.015.

Chan, P.P., Lin, B.Y., Mak, A.J., and Lowe, T.M. (2021). TRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. Nucleic Acids Res. *49*, 9077–9096. https://doi.org/10.1093/nar/gkab688.

Chen, H., Tang, H., and Ebright, R.H. (2003). Functional interaction between RNA polymerase alpha subunit C-terminal domain and sigma70 in UP-element- and activator-dependent transcription. Mol. Cell *11*, 1621–1633. https://doi.org/10.1016/s1097-2765(03)00201-6.

Chen, I.-M.A., Chu, K., Palaniappan, K., Ratner, A., Huang, J., Huntemann, M., Hajek, P., Ritter, S., Varghese, N., Seshadri, R., et al. (2021). The IMG/M data management and analysis system v.6.0: new tools and advanced capabilities. Nucleic Acids Res. *49*, D751–D763. https://doi.org/10.1093/nar/gkaa939.

Cheng, H., Liao, Y., Schaeffer, R.D., and Grishin, N.V. (2015). Manual classification strategies in the ECOD database. Proteins *83*, 1238–1251. https://doi.org/10.1002/prot.24818.

Chiba, Y., Oiki, S., Yaguchi, T., Urayama, S.I., and Hagiwara, D. (2021). Discovery of divided RdRp sequences and a hitherto unknown genomic complexity in fungal viruses. Virus Evol. *7*, veaa101. https://doi.org/10.1093/ve/veaa101.

Clem, R.J. (2015). Viral IAPs, then and now. Semin. Cell Dev. Biol. *39*, 72–79. https://doi.org/10.1016/j.semcdb.2015.01.011.

Clum, A., Huntemann, M., Bushnell, B., Foster, B., Foster, B., Roux, S., Hajek, P.P., Varghese, N., Mukherjee, S., Reddy, T.B.K., et al. (2021). DOE JGI metagenome workflow. mSystems *6*, e00804-20. https://doi.org/10.1128/mSystems.00804-20.

Cross, S.T., Maertens, B.L., Dunham, T.J., Rodgers, C.P., Brehm, A.L., Miller, M.R., Williams, A.M., Foy, B.D., and Stenglein, M.D. (2020). Partitiviruses infecting Drosophila melanogaster and Aedes aegypti exhibit efficient biparental vertical transmission. J. Virol. *94*, e01070-20. https://doi.org/10.1128/JVI.01070-20.

Csardi, G., and Nepusz, T. (2006). The igraph software package for complex network research. Int.J. Complex Syst. *1695*, 1–9.

Davison, M., Treangen, T.J., Koren, S., Pop, M., and Bhaya, D. (2016). Diversity in a polymicrobial community revealed by analysis of viromes, endolysins and CRISPR spacers. PLoS One *11*, e0160574. https://doi.org/10.1371/journal.pone.0160574.

de Souza, R.F., and Aravind, L. (2012). Identification of novel components of NAD-utilizing metabolic pathways and prediction of their biochemical functions. Mol. Biosyst. *8*, 1661–1677. https://doi.org/10.1039/c2mb05487f.

Dessau, M., Goldhill, D., McBride, R.L., Turner, P.E., and Modis, Y. (2012). Selective pressure causes an RNA virus to trade reproductive fitness for increased structural and thermal stability of a viral enzyme. PLoS Genet. *8*, e1003102. https://doi.org/10.1371/journal.pgen.1003102.

Dolja, V.V., and Koonin, E.V. (2018). Metagenomics reshapes the concepts of RNA virus evolution by revealing extensive horizontal virus transfer. Virus Res. *244*, 36–52. https://doi.org/10.1016/j.virusres.2017.10.020.

Dolja, V.V., Kreuze, J.F., and Valkonen, J.P.T. (2006). Comparative and functional genomics of closteroviruses. Virus Res. *117*, 38–51. https://doi.org/10.1016/j.virusres.2006.02.002.

Edgar, R.C. (2021). MUSCLE v5 Enables Improved Estimates of Phylogenetic Tree Confidence by Ensemble Bootstrapping (Bioinformatics).

Edgar, R.C., Taylor, J., Lin, V., Altman, T., Barbera, P., Meleshko, D., Lohr, D., Novakovsky, G., Buchfink, B., Al-Shayeb, B., et al. (2022). Petabase-scale

sequence alignment catalyses viral discovery. Nature *602*, 142–147. https://doi.org/10.1038/s41586-021-04332-2.

Enright, A.J., Van Dongen, S., and Ouzounis, C.A. (2002). An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. *30*, 1575–1584. https://doi.org/10.1093/nar/30.7.1575.

Finn, R.D., Clements, J., and Eddy, S.R. (2011). HMMER web server: interactive sequence similarity searching. Nucleic Acids Res. *39*, W29–W37. https://doi.org/10.1093/nar/gkr367.

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinform. Oxf. Engl. *28*, 3150–3152. https://doi.org/10.1093/bioinformatics/bts565.

Galperin, M.Y., Wolf, Y.I., Makarova, K.S., Vera Alvarez, R., Landsman, D., and Koonin, E.V. (2021). COG database update: focus on microbial diversity, model organisms, and widespread pathogens. Nucleic Acids Res. *49*, D274–D281. https://doi.org/10.1093/nar/gkaa1018.

Gann, E.R., Kang, Y., Dyhrman, S.T., Gobler, C.J., and Wilhelm, S.W. (2021). Metatranscriptome library preparation influences analyses of viral community activity During a brown tide bloom. Front. Microbiol. *12*, 664189. https://doi.org/10.3389/fmicb.2021.664189.

Gravestein, L.A., and Borst, J. (1998). Tumor necrosis factor receptor family members in the immune system. Semin. Immunol. *10*, 423–434. https://doi.org/10.1006/smim.1998.0144.

Gregory, A.C., Zayed, A.A., Conceição-Neto, N., Temperton, B., Bolduc, B., Alberti, A., Ardyna, M., Arkhipova, K., Carmichael, M., Cruaud, C., et al. (2019). Marine DNA viral macro- and microdiversity from pole to pole. Cell *177*, 1109–1123.e14. https://doi.org/10.1016/j.cell.2019.03.040.

Haan, C., Kreis, S., Margue, C., and Behrmann, I. (2006). Jaks and cytokine receptors–an intimate relationship. Biochem. Pharmacol. *72*, 1538–1546. https://doi.org/10.1016/j.bcp.2006.04.013.

Hastie, K.M., Kimberlin, C.R., Zandonatti, M.A., MacRae, I.J., and Saphire, E.O. (2011). Structure of the Lassa virus nucleoprotein reveals a dsRNA-specific 3′ to 5′ exonuclease activity essential for immune suppression. Proc. Natl. Acad. Sci. USA *108*, 2396–2401. https://doi.org/10.1073/pnas.1016404108.

Hauser, M., Steinegger, M., and Söding, J. (2016). MMseqs software suite for fast and deep clustering and searching of large protein sequence sets. Bioinformatics *32*, 1323–1330. https://doi.org/10.1093/bioinformatics/btw006.

Hockenberry, A.J., Jewett, M.C., Amaral, L.A., and Wilke, C.O. (2018). Withingene Shine–Dalgarno sequences are not selected for function. Mol Biol Evol *35*, 2487–2498. https://doi.org/10.1093/molbev/msy150.

Holmes, E.C., and Duchêne, S. (2019). Can sequence phylogenies safely infer the origin of the global virome? mBio *10*. e00289-00219. https://doi.org/10.1128/mBio.00289-19.

Hughes, K.J., Chen, X., Burroughs, A.M., Aravind, L., and Wolin, S.L. (2020). An RNA repair operon regulated by damaged tRNAs. Cell Rep. *33*, 108527. https://doi.org/10.1016/j.celrep.2020.108527.

Hunter, J.D. (2007). Matplotlib: A 2D graphics environment. Comput. Sci. Eng. *9*, 90–95. https://doi.org/10.1109/MCSE.2007.55.

Hyatt, D., Chen, G.-L., Locascio, P.F., Land, M.L., Larimer, F.W., and Hauser, L.J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics *11*, 119. https://doi.org/10.1186/1471-2105-11-119.

International Committee on Taxonomy of Viruses Executive Committee (2020). The new scope of virus taxonomy: partitioning the virosphere into 15 hierarchical ranks. Nat. Microbiol. *5*, 668–674. https://doi.org/10.1038/s41564-020-0709-x.

Ivanova, N.N., Schwientek, P., Tripp, H.J., Rinke, C., Pati, A., Huntemann, M., Visel, A., Woyke, T., Kyrpides, N.C., and Rubin, E.M. (2014). Stop codon reassignments in the wild. Science *344*, 909–913. https://doi.org/10.1126/science.1250691.

Johnson, M., Zaretskaya, I., Raytselis, Y., Merezhuk, Y., McGinnis, S., and Madden, T.L. (2008). NCBI BLAST: a better web interface. Nucleic Acids Res. *36*, W5–W9. https://doi.org/10.1093/nar/gkn201.

Jones, P., Binns, D., Chang, H.Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., et al. (2014). InterProScan 5: Genome-scale protein function classification. Bioinformatics *30*, 1236–1240. https://doi.org/10.1093/bioinformatics/btu031.

Kago, G., and Parrish, S. (2021). The Mimivirus L375 Nudix enzyme hydrolyzes the 5′ mRNA cap. PloS One *16*, e0245820. https://doi.org/10.1371/journal.pone.0245820.

Käll, L., Krogh, A., and Sonnhammer, E.L.L. (2004). A combined transmembrane topology and signal peptide prediction method. J. Mol. Biol. *338*, 1027–1036. https://doi.org/10.1016/j.jmb.2004.03.016.

Käll, L., Krogh, A., and Sonnhammer, E.L.L. (2007). Advantages of combined transmembrane topology and signal peptide prediction–the Phobius web server. Nucleic Acids Res. *35*, W429–W432. https://doi.org/10.1093/nar/gkm256.

Katoh, K., and Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. *30*, 772–780. https://doi.org/10.1093/molbev/mst010.

Koonin, E.V., Dolja, V.V., and Krupovic, M. (2015). Origins and evolution of viruses of eukaryotes: the ultimate modularity. Virology *479–480*, 2–25. https://doi.org/10.1016/j.virol.2015.02.039.

Koonin, E.V., Dolja, V.V., and Krupovic, M. (2022). The logic of virus evolution. Cell Host Microbe *30*, 917–929. https://doi.org/10.1016/j.chom.2022.06.008.

Koonin, E.V., Dolja, V.V., Krupovic, M., Varsani, A., Wolf, Y.I., Yutin, N., Zerbini, F.M., and Kuhn, J.H. (2020). Global organization and proposed megataxonomy of the virus world. Microbiol. Mol. Biol. Rev. *84*, e00061-19. https://doi.org/10.1128/MMBR.00061-19.

Krishnamurthy, S.R., Janowski, A.B., Zhao, G., Barouch, D., and Wang, D. (2016). Hyperexpansion of RNA bacteriophage diversity. PLoS Biol. *14*, e1002409. https://doi.org/10.1371/journal.pbio.1002409.

Krishnamurthy, S.R., and Wang, D. (2018). Extensive conservation of prokaryotic ribosomal binding sites in known and novel picobirnaviruses. Virology *516*, 108–114.

Krogh, A., Larsson, B., von Heijne, G., and Sonnhammer, E.L.L. (2001). Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. J. Mol. Biol. *305*, 567–580. https://doi.org/10.1006/jmbi.2000.4315.

Kutyshenko, V.P., Prokhorov, D.A., Mikoulinskaia, G.V., Molochkov, N.V., Yegorov, A.Y., Paskevich, S.I., and Uversky, V.N. (2021). Comparative analysis of the active sites of orthologous endolysins of the Escherichia lytic bacteriophages T5, RB43, and RB49. Int. J. Biol. Macromol. *166*, 1096–1105. https://doi.org/10.1016/j.ijbiomac.2020.10.264.

Lauber, C., Seifert, M., Bartenschlager, R., and Seitz, S. (2019). Discovery of highly divergent lineages of plant-associated astro-like viruses sheds light on the emergence of potyviruses. Virus Res. *260*, 38–48. https://doi.org/10.1016/j.virusres.2018.11.009.

Laudenbach, B.T., Krey, K., Emslander, Q., Andersen, L.L., Reim, A., Scaturro, P., Mundigl, S., Dächert, C., Manske, K., Moser, M., et al. (2021). NUDT2 initiates viral RNA degradation by removal of 5′-phosphates. Nat. Commun. *12*, 6918. https://doi.org/10.1038/s41467-021-27239-y.

Li, D., Luo, R., Liu, C.M., Leung, C.M., Ting, H.F., Sadakane, K., Yamashita, H., and Lam, T.W. (2016). MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. Methods *102*, 3–11.

Lu, S., Wang, J., Chitsaz, F., Derbyshire, M.K., Geer, R.C., Gonzales, N.R., Gwadz, M., Hurwitz, D.I., Marchler, G.H., Song, J.S., et al. (2020). CDD/SPARCLE: the conserved domain database in 2020. Nucleic Acids Res. *48*, D265–D268. https://doi.org/10.1093/nar/gkz991.

Makarova, K.S., Wolf, Y.I., Iranzo, J., Shmakov, S.A., Alkhnbashi, O.S., Brouns, S.J.J., Charpentier, E., Cheng, D., Haft, D.H., Horvath, P., et al. (2020). Evolutionary classification of CRISPR-Cas systems: a burst of class 2 and derived variants. Nat. Rev. Microbiol. *18*, 67–83. https://doi.org/10.1038/s41579-019-0299-x.

Martinez-Hernandez, F., Fornas, O., Lluesma Gomez, M., Bolduc, B., de la Cruz Peña, M.J., Martínez, J.M., Anton, J., Gasol, J.M., Rosselli, R., Rodriguez-Valera, F., et al. (2017). Single-virus genomics reveals hidden cosmopolitan and abundant viruses. Nat. Commun. *8*, 15892. https://doi.org/10.1038/ncomms15892.

Mihara, T., Nishimura, Y., Shimizu, Y., Nishiyama, H., Yoshikawa, G., Uehara, H., Hingamp, P., Goto, S., and Ogata, H. (2016). Linking virus genomes with Host Taxonomy. Viruses *8*, 66. https://www.mdpi.com/1999-4915/8/3/66.

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G.A., Sonnhammer, E.L.L., Tosatto, S.C.E., Paladin, L., Raj, S., Richardson, L.J., et al. (2021). Pfam: the protein families database in 2021. Nucleic Acids Res. *49*, D412–D419. https://doi.org/10.1093/nar/gkaa913.

Morgulis, A., Gertz, E.M., Schäffer, A.A., and Agarwala, R. (2006). A fast and symmetric DUST implementation to mask low-complexity DNA sequences. J. Comput. Biol. *13*, 1028–1040. https://doi.org/10.1089/cmb.2006.13.1028.

Mukherjee, S., Stamatis, D., Bertsch, J., Ovchinnikova, G., Sundaramurthi, J.C., Lee, J., Kandimalla, M., Chen, I.-M.A., Kyrpides, N.C., and Reddy, T.B.K. (2021). Genomes OnLine database (GOLD) v.8: overview and updates. Nucleic Acids Res. *49*, D723–D733. https://doi.org/10.1093/nar/gkaa983.

Mushegian, A.R. (2020). Are there 10 31 virus particles on earth, or more, or fewer? J. Bacteriol. *202*, e00052-20. https://doi.org/10.1128/JB.00052-20.

Nayfach, S., Camargo, A.P., Schulz, F., Eloe-Fadrosh, E., Roux, S., and Kyrpides, N.C. (2021). CheckV assesses the quality and completeness of metagenome-assembled viral genomes. Nat. Biotechnol. *39*, 578–585. https://doi.org/10.1038/s41587-020-00774-7.

NCBI Resource Coordinators (2018). Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. *46*, D8–D13. https://doi.org/10.1093/nar/gkx1095.

Nguyen, L.-T., Schmidt, H.A., von Haeseler, A., and Minh, B.Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. *32*, 268–274. https://doi.org/10.1093/molbev/msu300.

Nibert, M.L. (2017). Mitovirus UGA(Trp) codon usage parallels that of host mitochondria. Virology *507*, 96–100. https://doi.org/10.1016/j.virol.2017.04.010.

Oliveira, H., Melo, L.D.R., Santos, S.B., Nóbrega, F.L., Ferreira, E.C., Cerca, N., Azeredo, J., and Kluskens, L.D. (2013). Molecular aspects and comparative genomics of bacteriophage endolysins. J. Virol. *87*, 4558–4570. https://doi.org/10.1128/JVI.03277-12.

Paget, M.S.B., and Helmann, J.D. (2003). The sigma70 family of sigma factors. Genome Biol. *4*, 203. https://doi.org/10.1186/gb-2003-4-1-203.

Potenza, E., Di Domenico, T., Walsh, I., and Tosatto, S.C.E. (2015). MobiDB 2.0: an improved database of intrinsically disordered and mobile proteins. Nucleic Acids Res. *43*, D315–D320. https://doi.org/10.1093/nar/gku982.

Potter, S.C., Luciani, A., Eddy, S.R., Park, Y., Lopez, R., and Finn, R.D. (2018). HMMER web server: 2018 update. Nucleic Acids Res *46*, W200–W204. https://doi.org/10.1093/nar/gky448.

Price, M.N., Dehal, P.S., and Arkin, A.P. (2010). FastTree 2 – approximately maximum-likelihood trees for large alignments. PLoS One *5*, e9490. https://doi.org/10.1371/journal.pone.0009490.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., and Glöckner, F.O. (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. *41*, D590–D596.

Quinlan, A.R. (2014). BEDTools: the Swiss-army tool for genome feature analysis. Curr. Protoc. Bioinform. *47*, 11.12.1–11.12.34. https://doi.org/10.1002/0471250953.bi1112s47.

Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS: the European molecular biology open software suite. Trends Genet. *16*, 276–277. https://doi.org/10.1016/s0168-9525(00)02024-2.

Richter, M., and Rosselló-Móra, R. (2009). Shifting the genomic gold standard for the prokaryotic species definition. Proc. Natl. Acad. Sci. USA *106*, 19126–19131. https://doi.org/10.1073/pnas.0906412106.

Ring, K.L., and Cavalcanti, A.R.O. (2008). Consequences of stop codon reassignment on protein evolution in ciliates with alternative genetic codes. Mol. Biol. Evol. *25*, 179–186. https://doi.org/10.1093/molbev/msm237.

Ross, E., Blair, D., Guerrero-Hernández, C., and Sánchez Alvarado, A. (2016). Comparative and transcriptome analyses uncover key aspects of coding- and long noncoding RNAs in flatworm mitochondrial genomes. G3 (Bethesda) *6*, 1191–1200. https://doi.org/10.1534/g3.116.028175.

Roux, S., Brum, J.R., Dutilh, B.E., Sunagawa, S., Duhaime, M.B., Loy, A., Poulos, B.T., Solonenko, N., Lara, E., Poulain, J., et al. (2016). Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. Nature *537*, 689–693. https://doi.org/10.1038/nature19366.

Roux, S., Krupovic, M., Poulet, A., Debroas, D., and Enault, F. (2012). Evolution and diversity of the Microviridae viral family through a collection of 81 new complete genomes assembled from virome reads. PloS One *7*, e40418. https://doi.org/10.1371/journal.pone.0040418.

Roux, S., Páez-Espino, D., Chen, I.-M.A., Palaniappan, K., Ratner, A., Chu, K., Reddy, T.B.K., Nayfach, S., Schulz, F., Call, L., et al. (2021). IMG/VR v3: an integrated ecological and evolutionary framework for interrogating genomes of uncultivated viruses. Nucleic Acids Res. *49*, D764–D775. https://doi.org/10.1093/nar/gkaa946.

Salazar, G., Paoli, L., Alberti, A., Huerta-Cepas, J., Ruscheweyh, H.J., Cuenca, M., Field, C.M., Coelho, L.P., Cruaud, C., Engelen, S., et al. (2019). Gene expression changes and community turnover differentially shape the global ocean metatranscriptome. Cell *179*, 1068–1083.e21. https://doi.org/10.1016/j.cell.2019.10.014.

Sawaya, R., Schwer, B., and Shuman, S. (2005). Structure-function analysis of the yeast NAD^+-dependent tRNA 2′-phosphotransferase Tpt1. RNA N.Y. *11*, 107–113. https://doi.org/10.1261/rna.7193705.

Schulz, F., Roux, S., Paez-Espino, D., Jungbluth, S., Walsh, D.A., Denef, V.J., McMahon, K.D., Konstantinidis, K.T., Eloe-Fadrosh, E.A., Kyrpides, N.C., et al. (2020). Giant virus diversity and host interactions through global metagenomics. Nature *578*, 432–436. https://doi.org/10.1038/s41586-020-1957-x.

Shi, M., Lin, X.-D., Tian, J.-H., Chen, L.-J., Chen, X., Li, C.-X., Qin, X.-C., Li, J., Cao, J.-P., Eden, J.-S., et al. (2016). Redefining the invertebrate RNA virosphere. Nature *540*, 539–543. https://doi.org/10.1038/nature20167.

Silas, S., Makarova, K.S., Shmakov, S., Páez-Espino, D., Mohr, G., Liu, Y., Davison, M., Roux, S., Krishnamurthy, S.R., Fu, B.X.H., et al. (2017). On the origin of reverse transcriptase-using CRISPR-Cas systems and their hyperdiverse, enigmatic spacer repertoires. mBio *8*. e00897–e00817. https://doi.org/10.1128/mBio.00897-17.

Sillitoe, I., Bordin, N., Dawson, N., Waman, V.P., Ashford, P., Scholes, H.M., Pang, C.S.M., Woodridge, L., Rauer, C., Sen, N., et al. (2021). CATH: increased structural coverage of functional space. Nucleic Acids Res. *49*, D266–D273. https://doi.org/10.1093/nar/gkaa1079.

Simmonds, P., Adams, M.J., Benkő, M., Breitbart, M., Brister, J.R., Carstens, E.B., Davison, A.J., Delwart, E., Gorbalenya, A.E., Harrach, B., et al. (2017). Consensus statement: virus taxonomy in the age of metagenomics. Nat. Rev. Microbiol. *15*, 161–168. https://doi.org/10.1038/nrmicro.2016.177.

Simone, P.D., Pavlov, Y.I., and Borgstahl, G.E.O. (2013). ITPA (inosine triphosphate pyrophosphatase): from surveillance of nucleotide pools to human disease and pharmacogenetics. Mutat. Res. *753*, 131–146. https://doi.org/10.1016/j.mrrev.2013.08.001.

Skennerton, C.T., Imelfort, M., and Tyson, G.W. (2013). Crass: identification and reconstruction of CRISPR from unassembled metagenomic data. Nucleic Acids Res. *41*, e105. https://doi.org/10.1093/nar/gkt183.

Söding, J. (2005). Protein homology detection by HMM-HMM comparison. Bioinform. Oxf. Engl. *21*, 951–960. https://doi.org/10.1093/bioinformatics/bti125.

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

Steinegger, M., Meier, M., Mirdita, M., Vöhringer, H., Haunsberger, S.J., and Söding, J. (2019). HH-suite3 for fast remote homology detection and deep protein annotation. BMC Bioinformatics *20*, 473. https://doi.org/10.1186/s12859-019-3019-7.

Steinegger, M., and Söding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat. Biotechnol. *35*, 1026–1028. https://doi.org/10.1038/nbt.3988.

Sutela, S., Forgia, M., Vainio, E.J., Chiapello, M., Daghino, S., Vallino, M., Martino, E., Girlanda, M., Perotto, S., and Turina, M. (2020). The virome from a collection of endomycorrhizal fungi reveals new viral taxa with unprecedented genome organization. Virus Evol. *6*, veaa076. https://doi.org/10.1093/ve/veaa076.

Vainio, E.J., Chiba, S., Ghabrial, S.A., Maiss, E., Roossinck, M., Sabanadzovic, S., Suzuki, N., Xie, J., and Nibert, M.; ICTV Report Consortium (2018). ICTV virus taxonomy profile: Partitiviridae. J. Gen. Virol. *99*, 17–18. https://doi.org/10.1099/jgv.0.000985.

van der Meer, M.T.J., Klatt, C.G., Wood, J., Bryant, D.A., Bateson, M.M., Lammerts, L., Schouten, S., Damsté, J.S.S., Madigan, M.T., and Ward, D.M. (2010). Cultivation and genomic, nutritional, and lipid biomarker characterization of Roseiflexus strains closely related to predominant in situ populations inhabiting Yellowstone hot spring microbial mats. J. Bacteriol. *192*, 3033–3042. https://doi.org/10.1128/JB.01610-09.

Vasudevan, D., and Ryoo, H.D. (2015). Regulation of cell death by IAPs and their antagonists. Curr. Top. Dev. Biol. *114*, 185–208. https://doi.org/10.1016/bs.ctdb.2015.07.026.

Wheeler, T.J., and Eddy, S.R. (2013). nhmmer: DNA homology search with profile HMMs. Bioinformatics *29*, 2487–2489. https://doi.org/10.1093/bioinformatics/btt403.

Wolf, Y.I., Kazlauskas, D., Iranzo, J., Lucía-Sanz, A., Kuhn, J.H., Krupovic, M., Dolja, V.V., and Koonin, E.V. (2018). Origins and evolution of the global RNA virome. mBio *9*. e02329–e02318. https://doi.org/10.1128/mBio.02329-18.

Wolf, Y.I., Silas, S., Wang, Y., Wu, S., Bocek, M., Kazlauskas, D., Krupovic, M., Fire, A., Dolja, V.V., and Koonin, E.V. (2020). Doubling of the known set of RNA viruses by metagenomic analysis of an aquatic virome. Nat. Microbiol. *5*, 1262–1270. https://doi.org/10.1038/s41564-020-0755-4.

Wu, R., Davison, M.R., Gao, Y., Nicora, C.D., Mcdermott, J.E., Burnum-Johnson, K.E., Hofmockel, K.S., and Jansson, J.K. (2021). Moisture modulates soil reservoirs of active DNA and RNA viruses. Commun. Biol. *4*, 992. https://doi.org/10.1038/s42003-021-02514-2.

Xu, S., Dai, Z., Guo, P., Fu, X., Liu, S., Zhou, L., Tang, W., Feng, T., Chen, M., Zhan, L., et al. (2021). ggtreeExtra: compact visualization of richly annotated phylogenetic data. Mol. Biol. Evol. *38*, 4039–4042. https://doi.org/10.1093/molbev/msab166.

Yu, G., Lam, T.T.-Y., Zhu, H., and Guan, Y. (2018). Two methods for mapping and visualizing associated data on phylogeny using ggtree. Mol. Biol. Evol. *35*, 3041–3043. https://doi.org/10.1093/molbev/msy194.

Yutin, N., Benler, S., Shmakov, S.A., Wolf, Y.I., Tolstoy, I., Rayko, M., Antipov, D., Pevzner, P.A., and Koonin, E.V. (2021). Analysis of metagenome-assembled viral genomes from the human gut reveals diverse putative CrAss-like phages with unique genomic features. Nat. Commun. *12*, 1044. https://doi.org/10.1038/s41467-021-21350-w.

Zayed, A.A., Wainaina, J.M., Dominguez-Huerta, G., Pelletier, E., Guo, J., Mohssen, M., Tian, F., Pratama, A.A., Bolduc, B., Zablocki, O., et al. (2022). Cryptic and abundant marine viruses at the evolutionary origins of Earth's RNA virome. Science *376*, 156–162. https://doi.org/10.1126/science.abm5847.

Zeigler Allen, L., McCrow, J.P., Ininbergs, K., Dupont, C.L., Badger, J.H., Hoffman, J.M., Ekman, M., Allen, A.E., Bergman, B., and Venter, J.C. (2017). The Baltic Sea virome: diversity and transcriptional activity of DNA and RNA viruses. mSystems *2*, e00125-16. https://doi.org/10.1128/mSystems.00125-16.

Zimmermann, L., Stephens, A., Nam, S.-Z., Rau, D., Kübler, J., Lozajic, M., Gabler, F., Söding, J., Lupas, A.N., and Alva, V. (2018). A completely reimplemented MPI bioinformatics toolkit with a new HHpred server at its core. J. Mol. Biol. *430*, 2237–2243. https://doi.org/10.1016/j.jmb.2017.12.007.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
All original data and code produced in this work	This paper	https://doi.org/10.5281/zenodo.6553771
Original code produced in this study	This paper	https://github.com/UriNeri/RVMT
accompanying interactive web portal	This paper	https://riboviria.org
Software and algorithms		
MMseqs2	Steinegger and Söding, 2017	https://github.com/soedinglab/MMseqs2
NCBI BLAST+ suite	Altschul et al., 1997; Johnson et al., 2008	https://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/
DIAMOND	Buchfink et al., 2015	https://github.com/bbuchfink/diamond
bbmap v38.81	Bushnell, 2014	https://sourceforge.net/projects/bbmap/
MUSCLE v.5	Edgar, 2021	https://www.drive5.com/muscle/downloads.htm
Mafft v7.407	Katoh and Standley, 2013	https://mafft.cbrc.jp/alignment/software/
HH-Suite	Steinegger et al., 2019	https://github.com/soedinglab/hh-suite
HMMER	Söding, 2005; Potter et al., 2018	http://hmmer.org/
CD-HIT	Fu et al., 2012	https://github.com/weizhongli/cdhit
MCL	Enright et al., 2002	https://micans.org/mcl/index.html
ggtree	Yu et al., 2018	https://bioconductor.org/packages/release/bioc/html/ggtree.html
ggtreeExtra	Xu et al., 2021	https://bioconductor.org/packages/release/bioc/html/ggtreeExtra.html
IQ-Tree	Nguyen et al., 2015	http://www.iqtree.org/
dustmasker (v1.0.0)	Morgulis et al., 2006	https://www.ncbi.nlm.nih.gov/IEB/ToolBox/CPP_DOC/lxr/source/src/app/dustmasker/
etandem (v6.6.0.0)	Rice et al., 2000	http://emboss.open-bio.org/rel/rel6/apps/etandem.html
R	The R Project for Statistical Computing	https://cran.r-project.org/
Python	Python Software Foundation	https://www.python.org
Igraph	Csardi and Nepusz, 2006	https://igraph.org/
Prodigal (v2.6.3)	Hyatt et al., 2010	https://github.com/hyattpd/Prodigal
tRNAscanME2	Chan et al., 2021	https://github.com/UCSC-LoweLab/tRNAscan-SE

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and additional data should be directed to and will be fulfilled by the lead contact, urineri@mail.tau.ac.il (U.N.).

Materials availability

This study did not generate new unique reagents, physical samples, or specific biological material. As a computational project, the input for this study is publicly available as detailed below in "metatranscriptome acquisition". All results and output of this study are described below in the "data and code availability" section.

Data and code availability

All original data and code produced in this work is freely and fully available through several venues (DOIs also listed in the key resources table):

- All the data, code, results produced in the course of this project, as well as the latest release of the accompanying interactive web portal (https://riboviria.org), are available via CERN's Zenodo repository (https://doi.org/10.5281/zenodo.6553771). This project is intended to serve as a community wide resource. As such, the Zenodo repository includes the additional information and various intermediary results and secondary analyses, such the predicted coding sequences, host assignments, phylogeny and taxonomic affiliation, raw domain hidden markov model (HMM) search matches, additional domain profile databases generated in this work (e.g. alignments, HMMs, original seed sequences and predicted function) as well the nucleic sequences for both the expanded (2.6M metatranscriptome derived) contig set and the manually consolidated "Reference Set" (see STAR Methods).
- As noted above, the Zenodo deposit includes the original code produced in this study, which corresponds to the latest version of the project's GitHub repository, which is available under the open-source MIT License at https://github.com/UriNeri/RVMT.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

Metatranscriptome acquisition

The identification of RNA viruses was performed on a total of 5,150 publicly available, pre-assembled metatranscriptomes, that were retrieved from IMG/M in January 2020 (Chen et al., 2021; Mukherjee et al., 2021). As previously described, the majority of these were assembled using MEGAHIT (Li et al., 2016) (see Table S4 for information referring to the assembler used in the different samples, and when available, reference to the study where the samples were originally published).

Primary Filtering process

For convenience, we summarized the final tools and cutoffs of the Primary and secondary filtration process in Table S6 - Discovery pipeline search and filtration thresholds.

Our initial criteria for contigs acquired from the IMG/M portal discarded sequences shorter than 1,000 nt or encoding rRNA genes (the remaining contigs were dereplicated at 99% sequence identity via mmseqs easy-linclust) (Steinegger and Söding, 2017).

To filter out sequences that were highly unlikely to represent RNA viruses, we compared the obtained metatranscriptome contigs to a compendium of DNA sequences built from 1,831 metagenomes originated from the same studies as 1,306 of the metatranscriptomes. We selected metagenomes that shared the metadata attribute of "Study_ID" with the 5,510 metatranscriptomes in the Genomes OnLine Database (GOLD) portal (Mukherjee et al., 2021) as these DNA assemblies would cover a similar range of habitats as the analyzed metatranscriptomes. Using multiple sequence search tools (specifically, MMseqs2 (nucleic - nucleic (search type 3) (Hauser et al., 2016; Steinegger and Söding, 2017), DIAMOND (translated nucleotide versus the IMG sourced DNA metagenomic predicted ORFs (diamond blastx) (Buchfink et al., 2015), and NCBI BLAST (nucleic - nucleic - blastn) (Altschul et al., 1997; Camacho et al., 2009; Johnson et al., 2008)) in an iterative manner, we identified and excluded metatranscriptomic contigs that matched sequences in the DNA sequence dataset (Figure 1A), based on the assumption that RNA viruses would not be present in DNA assemblies — which would be comprised of cellular organisms, DNA-based mobile elements, and integrated retroviruses. The iterative search was performed such that each iteration gradually increased the search sensitivity (e.g., through decreased word length (BLASTn) and higher sensitivity value (MMseqs2 "–sensitivity")), while discarding all sequences from the metatranscriptomes collection that produced reliable matches to sequences in the "DNAome", before advancing with the filtered output to the next iteration. This process was repeated for a total of five iterations, though we should note the initial iterations were mainly exploratory (used for crude tuning of the procedure).

Secondary Filtering process

To further filter the contig set, we supplemented the above filtering process output with 5,954 RNA viral sequences from reference databases and performed an additional iterative filtering procedure using public databases (NCBI NT/NR and IMG/VR) as the DNA set. To prevent the exclusion of bona fide RNA virus sequences, we masked entries of the public databases that matched reference RNA viruses from subsequent iterations. All discarded contigs were aggregated and supplemented with manually identified DNA encoded contigs, creating a database of "false positives", that was used to further filter the metatranscriptome dataset through exclusion of sequences with producing passable matches to the "false positive" set. The procedure of collecting the discarded matches to further refine the working set was repeated three times.

Estimation of DNA remnants in intermediate sets

To evaluate remnants of DNA sequences in the working set through the filtration process, we routinely analyzed random contig subsets by (1) computing the RdRP to reverse-transcriptase domain ratio as a proxy to the RNA virus to DNA-encoded contigs; (2) manually inspecting the presence of the most frequent non-RNA virus-related domains. Of note, several specific domains

recurred frequently during this performance evaluation, and manual examination revealed these to be domains of known repeats. Mostly, these contigs were fully populated with matches to such repeat domains, and that these had cellular matches in the public DBs, whose alignment values were just below our reporting or acceptance criteria. Hence, we decided to discard these contigs if they were completely coding for multiple repeats, as there would be no sufficient coding space for these to encode an identifiable RdRP.

Following the below RdRP identification step (described in the section below) approximately 130 reverse-transcriptases had passed the various filtration processes and were manually removed. MMseqs2, the PFamA Database (Mistry et al., 2021) and the RdRP and RT collection from Wolf et al. (2018), were used in all the profile searches performed in this evaluation.

RdRP identification

Previously published multiple sequence alignments of RdRPs and reverse-transcriptases (Wolf et al., 2018, 2020) were formatted as tool-specific subject databases, and employed as queries to search a sequence database consisting of the 6-frame end-to-end translations of contigs passing the above-described filtering processes, using PSI-BLAST, hmmsearch, DIAMOND and MMseqs2. To estimate the desired search cutoffs, we supplemented the query set with non-RdRP sequences likely to produce false matches (termed ''true-negative'' set), constructed as follows: (1) using a large set of RdRPs as queries for an hhsearch (from the HH-Suite) against the PDB70 database (2019), collecting all matches of bitscore $\geq$ 20 that were not from RNA viruses, that aligned with at least 2 RdRPs; (2) fetching PDB entries clustered with those at 70% identity, (via ftp://resources.rcsb.org/sequence/clusters/bc-70.out); (3) fetching Pfam entries relating the resulting PDB IDs, and sequences linked to the Pfam entries; (4) collapsing highly similar sequences to a single representative (MMseqs2 minimum coverage: 100%, minimum ident.: 90%). Subject RdRP profiles capable of producing alignments to any sequence from the ''true-negative'' set were discarded. Otherwise, acceptance criteria for the RdRP profiles searches were: profile coverage $\geq$ 50%, E-value $\leq$ 1e-10 and score $\geq$ 70. These stringent parameters were then fine tuned to represent the best possible value a non-RdRP sequence was able to achieve.

Subsequently, reliable RdRP matches were trimmed to the approximate core domain, which we operationally defined as motif A–D (see ''Motif A–D identification'' below). The extracted RdRP core sequences were pre-clustered (CD-HIT, coverage $\geq$ 75%, % ID $\geq$ 90) (Fu et al., 2012), passed to an all vs. all (DIAMOND BLASTp) run, formatted for use with MCL using mcxload (–stream-mirror –stream-neg-log$_{10}$ -stream-tf "ceil(200)"), clustered (MCL, Inflation value between 3.6 and 2.8), aligned (MUSCLE), and formatted as profile databases as described above (Altschul et al., 1997; Buchfink et al., 2015; Edgar, 2021; Enright et al., 2002; Steinegger and Söding, 2017). This process was repeated twice. Subsequently, contigs with putative RdRPs were used to recover additional contigs from the entire metatranscriptomic collection, which were highly similar yet shorter than the initial search length criteria (see below ''Comprehensive identification'' for details). Of the resulting collection, sequences covering $\geq$ 75% of an RdRP profile, or with identifiable motifs A–D, were considered sufficiently complete for downstream phylogenetic analysis.

Identification of the RdRP catalytic motifs A–D

A custom motif library (available in the project Zenodo archive, see data and code availability) was built by semi-manual partitioning of the previously published RdRP MSAs noted in the ''RdRP identification'' section. To identify the motifs along the individual RdRP sequences, a similar iterative search as described above for the full length RdRP domain was performed.

Correction of putative frameshifts

A set of 1,656 contigs contained a clear RdRP domain signature on more than one frame, commonly separated by < 20 nucleotides (n=1,118). In order to avoid the omission of these signatures as simple incomplete, we addressed these in two manners: (1) if any of one of the signatures covered $\geq$75% of the subject RdRP profile, or coding for the desired catalytic motifs A–C, that signature would be used; or (2) by concatenation of the two signatures into a single amino acid sequence.

Contig set augmentation with published genomes

To assess the novelty of our findings in terms of the number and diversity of newly predicted viral genomes, and in order to avoid the exclusion of established viral lineages that may be underrepresented in environmental metatranscriptomes, we aggregated and compiled a collection of ''previously published'' viral genomes termed ''Reference Set''. These include RdRP-carrying sequences identified in NCBI's NT database (NCBI Resource Coordinators, 2018), as well as sequences not indexed (at the time of writing) in such public databases, that were identified in several previous large scale and notable RNA virus surveys and transcriptomic atlases. Our criteria for addition of these supplementary sequences required that they originate from peer reviewed publications, and that all underlying sequences were entirely publicly available, with no restrictions. The NCBI NT sequences were identified via an RdRP scan procedure similar to the procedure described above (see RdRP identification). The previously published set was made from an expansive set of *Leviviricetes* described by Callanan et al. (2020), the ''Yangshan-assemblage'' and other described by Wolf et al. (2020), and the proposed Plastroviruses group described by Lauber et al. (2019), as well as several RdRPs identified in the ocean atlas of genes (Carradec et al., 2018; Salazar et al., 2019). Following their aggregation, these sequences underwent a similar procedure described for the metatranscriptomic sequences identified in this work (i.e. length filtration, clustering, and RdRP core domain extraction). The eventual sequence set was labeled as ''Known'' (i.e. not novel), and noted as such in the data generated by this work (e.g. branch colour in Figure 1). The processed ''supplemental sequence set'' was merged into the main sequence set (those

identified in this study) and the combined set (termed "VR1507") was used in all downstream analysis (phylogenetic reconstruction, domain analysis etc).

Comprehensive identification of RNA virus contigs across metatranscriptomes

Because metatranscriptome assemblies can often yield incomplete genomes that would not fulfil the criteria for *de novo* RdRP detection (see above), we used the "VR1507" contig set (see above), we initiated a secondary "sweeping" scan for additional RNA virus contigs from the non-clustered, non-filtered (length, DNA similarity, RdRP presence) "bulk-set" of metatranscriptomic contigs (Figure 1). To this end, the "VR1507" was used as bait for highly similar contigs in the "bulk-set", using an non-sensitive mmseqs search (mmseqs search –search-type 3 –min-aln-len 120 –min-seq-id 0.66 -s 1 -c 0.85 –cov-mode 1) followed by stringently filtering the recovered matches (E-value < 1e-9, Identity > 95%, target-Coverage $\geq$ 95%). These criteria were selected as a quality assurance measure, so that the recovered contigs would be mostly contained within the "VR1507" contig counterpart (this large expansive data set is available in project's Zenodo repository, see data and code availability). This envelopment criteria was added to avoid capture of chimeric, or otherwise uncertain, nucleic regions that extend over the "VR1507" query. The filtered bulk contig set was combined with "VR1507" and consisted of 2,658,344 contigs (termed "Add1507"). To ascertain that this procedure was adequately stringent in avoiding the capture of false positives, we verified if we carried it out on non-RNA virus containing DNA metagenome, no contigs would be captured. For this end, we used a recently published high quality bovine (Rumen) DNA metagenome (i.e. long-read, HiFi assemblies) (Bickhart et al., 2022), selected as it was not part of the DNA-sequence set used in the primary and secondary filtration steps used in the discovery pipeline, making it a reliable benchmark. In this search, not a single contig passed our alignment threshold of 95% identity (a single contig produced a short alignment of 72% ID).

Phylogenetic reconstruction

We selected a diverse set of representative RdRPs for the phylogenetic analysis by performing a preliminary MMseqs2 clustering run (see Table S6, sheet "Clustering information), on a subset of the sequences which contained complete or near-complete RdRPs. These representatives were termed RCR90, and went through several iterations of clustering (MMseqs2 with sequence identity threshold of 0.5), alignment (MUSCLE5) (Edgar, 2021) and profile-profile comparison (HHsearch) (Steinegger et al., 2019), as described below. "Permuted" RdRPs (sequences with transposed motif C, following the C-A-B-D configuration) were identified and "de-permuted" (i.e. the loop, containing motif C, was cut from the sequence and reinserted downstream from the motif B). Once all identified sequences with transposed motifs were brought into the canonical A-B-C-D configuration, the following procedure was employed to produce a multiple sequence alignment consisting of all RCR90 set:

- Sequences were clustered using MMseqs2 with sequence identity threshold of 0.3; sequences in the resulting 4,514 clusters were aligned using MUSCLE5; profile-profile comparison of the cluster alignments using HHSEARCH produced a 4,514x4,514 distance matrix (the distances were estimated as $d_{AB} = -\ln(S_{AB}/\min(S_{AA}, S_{BB}))$, where S_{AB} is the HHSEARCH score for comparison of the profiles A And B); a maximum-linkage tree was produced from the distance matrix using the R function hclust();
- The tree was cut at the depth threshold of 1.5, producing 1,360 subtrees;
- Each of the subtrees was used as a guide to hierarchical alignment of the corresponding profiles using HHALIGN, producing 1,360 alignments;
- 1,360 consensus sequences (excluding sites with more than 2/3 of gap characters) were extracted from these alignments and aligned using MUSCLE5;
- Each position in the alignment of consensus sequences was expanded to the corresponding column of the original alignment, producing an alignment of 77,510 RdRps (where the original RdRp sequences were reduced to a set of positions, matching their local consensus);
- Sites with >90% of gap characters were removed from this alignment; the resulting alignment was aligned with the alignment of ten RTs (five group II intron sequences and five non-LTR retrotransposon sequences) using HHALIGN.

The alignment of RdRps and RTs was used to reconstruct an approximate maximum likelihood tree using the FastTree (V.2.1.4 SSE3, Price et al., 2010) program (WAG evolutionary model, gamma-distributed site rates) and rooted between RTs and RdRps.

Taxonomic affiliation of clades

Tree leaves with existing taxonomic information were identified by mapping (MEGA-BLAST, E-value < 1e-30, query coverage $\geq$ 95%, subject coverage $\geq$ 95%, Alignment length > 200, Identity $\geq$ 98%, (Alignment_length)/Query_length > 0.95) VR1507 sequence set to the latest ICTV data at the time of analysis (July 20, 2021 release of the Virus Metadata Repository (VMR) file, corresponding to MSL36, and available at https://talk.ictvonline.org/taxonomy/vmr/m/vmr-file-repository/13175). Overall, 2,765 contigs were mapped, and the ICTV taxonomic information was cloned to the VR1507 queries based on the highest score. For the reminder of VR1507 contigs, we performed a similar procedure using the NCBI's NR database (these amount to an additional 6,878 mapped contigs, though a non-negligible amount of those lacked taxonomic information or matched abolished taxonyms).

The procedure to establish the taxonomic affiliation of internal nodes on the tree (i.e. clades) relies on the above taxonomic assignment of reference tree leaves, as well all on two principles:

- All sequences, descending from the last common ancestor of reference leaves, assigned to a taxon *T*, also belong to taxon *T*; sequences descending from deeper tree nodes, do not belong to taxon *T* and, therefore, and should be assigned to a new taxon (taxa) of the same rank;
- The depth, at which a tree clade splits into taxa of the given rank, is defined by existing taxa of the same rank and is locality-dependent (e.g. the characteristic depths of families could be different for different phyla);

Application of these principles assumes that the existing taxonomy is non-contradictory with respect to the tree, i.e. the reference sequences, assigned to taxa, form monophyletic clades that are non-overlapping and non-nested within the same rank (e.g. a family clade can't be embedded into another family). An inspection of the taxonomic affiliation of reference leaves showed that this assumption, while typically satisfied, is violated in multiple places. This necessitates disentangling the conflicting relationships first. To this end, the following procedure was applied to all taxa of the given rank (i.e. separately for phyla, classes, etc):

- The tree was pruned to contain only leaves with this rank defined (e.g. all leaves without a family assignment are stripped); leaf weights (w_i) were derived from the pruned tree;
- For each taxon *T*, present in the tree, the total weight of leaves in this taxon was calculated ($W_T = \Sigma w_i$ across the leaves, assigned to *T*);
- For any tree clade in the tree, the total weight of leaves in this clade was calculated ($W_C = \Sigma wi$ across the leaves, belonging to *C*);
- For each combination of clade *C* and taxon *T*, the clade-taxon weight was calculated ($W_{CT} = \Sigma w_i$ across the leaves, belonging to *C* and assigned to *T*); then a precision-like and recall-like measures can be calculated ($P_{CT} = W_{CT} / W_C$ and $R_{CT} = W_{CT} / W_T$) and combined into a quality index $Q_{CT} = P_{CT} * R_{CT}$.
- For each taxon *T*, present in the tree, the clade $C_T = \text{argmax}\, Q_{CT}$ was identified as the "native" location of the taxon *T* (the clade, where the maximum weight of taxon *T* is concentrated with the minimal intrusion of other taxa); leaves, belonging to clade C_T, but not assigned to *T*, and leaves, assigned to *T*, but not belonging to clade C_T, were labeled as "intruding" or "outlying" respectively;

All tree-incompatible taxonomic assignments were examined and resolved. In most cases the most agnostic way to resolve the conflict was used (i.e. stripping the taxonomic labels from the corresponding leaves). In one case, most of the families within *Timlovirales* order of *Lenarviricota*, were found to be nested inside a very deep-branching family of *Blumeviridae*. For the purpose of this work, we retained the *Blumeviridae* label on the largest clade of *Timlovirales* that didn't have conflicting family assignments and removed the *Blumeviridae* label from the rest of *Timlovirales*. In a few other cases where small families were wholly nested into larger ones (e.g. a solo leaf classified as *Sunviridae* inside a large *Paramyxoviridae* clade) the embedded family label was removed for the purpose of subsequent analysis and restored post hoc. Once the taxonomic labels of all leaves were brought into compatibility with the tree, the following procedure was performed to assign new taxonomic labels to unlabeled leaves for each taxonomic rank separately:

- All nodes of the tree were assigned depth, defined as the longest node-to-leaf path across all leaves, descending from this node;
- In the full tree of 77,510 leaves the last common ancestor node of each taxon was determined; depths of the taxa, defined as the depth of the LCA node plus the length of the incoming tree edge, was recorded; all unlabeled leaves, descending from the taxon LCA, were assigned to this taxon;
- All clades outside of existing taxa were isolated; for each such clade the depths of all existing sister taxa were determined; if a clade has only one sister taxon, the search for the closest relatives was extended toward the root until at least another related taxon was identified; the threshold depth was calculated as the average for the set of related taxa;
- Clades outside of existing taxa were dissected at the threshold depth; each resulting (sub)clade was assigned to a new taxon of the given rank;
- New taxa that have a single existing taxon as a sister are labeled as associated with this taxon.

The novel taxa were given names, indicating rank (i.e. prefixed by *p*, *c*, *o, f* and *g* for phylum, class, order, family and genus respectively), followed by an ordinal number for new taxa of this rank, and optionally, terminated with a label for taxa that are associated with a previously described taxon (e.g. *f.0127.base-Noda* is the 127th new family that is basal to *Nodaviridae* in the RdRP tree).

Robustness of deep phylogeny

To assess the robustness of deep phylogenetic reconstruction, the following procedure was performed:

- a list of 201 families with at least 20 RCR90 sequences was collected
- a random representative of each family and from RT set was sampled
- a sub-alignment of 202 sequences for the sample was extracted from the master alignment
- a phylogenetic tree was reconstructed using the IQ-Tree program (Nguyen et al., 2015) with an automatically selected best fitting model

100 independent samples were analyzed in the following manner:

First, clades with the highest quality index (QI, described above in the Taxonomic affiliation of clades section) were identified for each of the five known phyla; the quality index values were used as a measure of the phylum monophyly under the subsampling. Families, involved in breaking the monophyly of the respective phyla (note that a leaf can be both an outlier with respect to its own phylum and an intruder into another phylum), were recorded.

Second, the subsampled trees were collapsed to the phylum level; 15 (out of 100) trees with paraphyletic phyla were excluded (those, where e.g. the highest-quality clade for *Pisuviricota* was embedded within the highest-quality clade for *Kitrinoviricota*). An extended majority-rule consensus tree was constructed for the remaining 85 trees with (largely) monophyletic phyla using the IQ-Tree program; branch support values were multiplied by 0.85 (the fraction of such trees among the whole sample).

Assignment of individual contigs to RCR90 clusters

Once the novel areas of the RCR90 megatree described above were fully populated by the major taxonomic ranks (Phylum → Genus), we proceeded to affiliate contigs from the larger VR1507 set (see above - contig sets). Contig affiliation was performed in a gradual manner by separation into the following 4 levels:

Level A. are contigs encoding the RdRPs used to create the tree. Level B. consists of contigs encoding RdRPs with exceptionally high amino acid identity to RdRPs from level A, (via best BLASTp match with Identity $\geq$90%, Query-Coverage $\geq$75%, and E-value < 1e-3). Level.C consisted of contigs from the same RvANI90 cluster (see definition below) as contigs from levels {A, B}, and Level D. consists of contigs sharing high nucleic similarity to those from levels {A - C}, (via best dc-MEGABLAST hit at Identity $\geq$90%, Query-Coverage $\geq$75% OR Nident $\geq$ 900nt and E-value < 1e-3). Based on the distribution of ICTV-labeled RdRPS in the above noted levels, we estimate that the majority contigs affiliated in this manner, would roughly share the same taxonomic ranks down to genus level.

Of note, for level C., we devised custom measurement unit, RvANI, which is an extension of standard average nucleic identity (ANI) clustering, designed to accommodate the fragmented nature of metatranscriptomic assemblies, thus avoiding an overestimation of novelty caused by the relatively low pairwise coverage of related sequences. Briefly, RvANI is calculated as follows: Initially, mmseqs is used to calculate all pairwise sequence alignments in the contig set, which are then used for the traditional ANI and alignment fraction (AF) calculations, where:

$$ANI = (\%ID \times Alignmentlength) \div Min(lengthofcontigm., lengthofcontign.)$$

$$AF = Min(Alignmentcoverageofcontigm., Alignmentcoverageofcontign.)$$

Given all pairs of ANI and AF (for prokaryotes 95-96% ANI is the commonly accepted species boundary, with similarly granular definitions for certain viruses (Nayfach et al., 2021; Richter and Rosselló-Móra, 2009) clusters are defined as connected components in a nucleic similarity graph pruned for pairwise alignments with ANI $\geq$90% and AF $\geq$90%. RvANI corrects for uneven genome coverage in metatranscriptomes by reinserting specific pairwise alignments to the pruned nucleic similarity graph, even if their AF is below the required cutoff, as long as the underlying pairwise alignment fulfill these criteria: %ID $\geq$ 99, Alignment Length $\geq$ 150 [bp], and the alignment occurs between the edge of the contigs, i.e. the alignment covers the 5' or 3' termini of each contig).

Subsequently, we defined RvANI90 clusters as the different connected components (using R-igraph package) in the nucleic similarity graph processed as described above (Csardi and Nepusz, 2006).

Identification of reliable CRISPR spacer hits

RNA virus sequences were compared to predicted bacteria and archaea CRISPR spacer sequences to (i) identify which viruses may infect a prokaryotic host, and (ii) possibly predict a specific host taxon for these viruses. First, non-redundant RNA virus sequences were compared to 1,568,535 CRISPR spacers predicted from whole genomes of bacteria and archaea in the IMG database (Chen et al., 2021) using blastn v2.9.0 with options "-dust no -word_size 7". To minimise the number of false-positive hits due to low-complexity and/or repeat sequences, CRISPR spacers were excluded from this analysis if (i) they were encoded in a predicted CRISPR array including 2 spacers or less, (ii) they were $\leq$ 20bp, or (iii) they included a low-complexity or repeat sequence as detected by dustmasker (v1.0.0) (Morgulis et al., 2006) (options "-window 20 -level 10") or a direct repeat of $\geq$ 4bp detected with etandem (v6.6.0.0) (Rice et al., 2000) (options "-minrepeat 4 -maxrepeat 15 -threshold 2"). To link RNA viruses to CRISPR spacers, only blastn hits with 0 or 1 mismatch over the whole spacer length were considered. The spacer and array with hits were further inspected to check (i) whether the spacers were of consistent length throughout the array, and (ii) whether Cas and/or RT genes were found in the putative host genome, and if so whether these were adjacent to the CRISPR array with the hit. To expand the search for CRISPR link beyond bacteria and archaea for which a draft genome is available, we next used the same approach to compare non-redundant RNA virus sequences to 53,372,161 CRISPR spacers predicted from metagenome assemblies available in the IMG database. Spurious spacers were filtered out using the same methods as for the genome-derived CRISPR arrays (see above), and only hits for which the RNA virus and the CRISPR spacers originated from the same ecosystem (as defined in the GOLD database) were retained. Since CRISPR spacer arrays are often assembled on short contigs without any other gene, we used the repeat sequence of the arrays to link them to a putative host. Repeat sequences from metagenome-derived CRISPR arrays with at least 1 hit

to an RNA virus sequence were compared to all IMG Bacteria and Archaea genomes using blastn (v2.9.0) with options "-perc_identity 90 -dust no -word_size 7". The location of these hits in the putative host genome was then checked for the presence of a predicted CRISPR spacer array, Cas genes, and RT genes. When individual RNA virus sequences or spacers were putatively linked to multiple host genomes, these were prioritized based on the following criteria: (i) the spacer array is identified next to an RT-encoding CRISPR array, (ii) an RT-encoding CRISPR array is identified elsewhere in the genome, (iii) the spacer array is identified next to a Type III CRISPR array, (iv) a Type III CRISPR array is identified elsewhere in the genome, (v) another type of CRISPR array is identified in the genome, and (vi) no identifiable Cas gene can be identified in the genome.

The spacer content of CRISPR arrays encoded by *Roseiflexus sp*. RS-1 in Mushroom Spring was further studied as follows. First, the CRISPR arrays of 17 metagenomes sampled from microbial mats in Mushroom Spring (Table S3) were specifically assembled using the dedicated tool Crass v1.0.1 with default parameters (Skennerton et al., 2013). Next, all arrays based on repeats corresponding to known CRISPR arrays in *Roseiflexus sp.* RS-1 (Table S3) were identified and the corresponding spacers collected and filtered as previously described. RNA virus sequences as well as DNA virus sequences from the IMG/VR v3 database (Roux et al., 2021) were compared to this database of *Roseiflexus sp*. *RS-1* spacer arrays using blastn (v2.9.0) with options "-dust no -word_size 7". Sequences from putative RNA phages infecting *Roseiflexus sp*. *RS-1* were first identified based on hits to $\geq$ 1 RS-1 spacer with $\leq$ 1 mismatch across the whole spacer length. For these selected phages, hits with up to 4 mismatches across the spacer length were then collected to enable the detection of more distant virus-spacer hits.

Candidate capsid segments of *Roseiflexus sp*. *RS-1* clade *genPartiti.0019 viruses* were identified based on 3 criteria: spacer match to the RNA-targeting CRISPR array, no corresponding DNA sequence, and high coverage correlation to $\geq$ 1 RdRP contig across the metatranscriptome time series. First, a similar blastn comparison to Crass-assembled spacers (blastn with options "-dust no -word_size 7" and $\leq$ 1 mismatch allowed) was used to identify putative capsid-encoding contigs i.e., excluding all contigs encoding an RdRP or a CRISPR array, in the same metatranscriptomes targeted by the *Roseiflexus sp.* RS-1 Type III-RT CRISPR array (n=3,958). Next, candidates with $\geq$ 1 spacer match were compared to all contigs from Mushroom Spring DNA metagenomes (blastn (v2.9.0) with options "-task megablast -max_target_seqs 500 -perc_identity 90"), and all candidates with a matching DNA contig ($\geq$ 90% identity) were considered to be likely DNA phages and excluded (n=3,650). Finally, the coverage of all *genPartiti.0019* RdRP contigs and all candidate capsid segments was obtained using read mapping as described below (bbmap.sh (v.38.90) with options "vslow minid=0 indelfilter=2 inslenfilter=3 dellenfilter=3"), and candidates with a Pearson correlation of $\geq$ 0.9 across the 42 Mushroom Spring metatranscriptomes were retained as likely capsid segments (n=88). To evaluate the gene content of these capsid segments, cds were predicted de novo using Prodigal (v2.6.3) (Hyatt et al., 2010) (option "-p meta"), and clustered using a standard blast-mcl pipeline (blastp (v2.9.0) with default options, hits selected based on score $\geq$ 50, MCL clustering (v.14-137) with an inflation value of 2). For the three largest protein clusters, a sequence alignment was built using MAFFT v7.407, (Katoh and Standley, 2013) and used as input to an hhsearch against the virus-focused uniprot public database (uniprot_sprot_vir70), and a custom database made from capsids of known partitiviruses and picobirnaviruses (available in the project's Zenodo repository, see data and code availability "Partiti_Picob_CP.tar.gz" and PC1_PROMALS3D_new.hhr).

Habitat distribution and relative abundance estimation

For visualisation purposes, location, ecological, and taxonomic information for each metatranscriptome were obtained from the IMG and GOLD databases. Specifically, GPS coordinates and ecosystem classification were obtained from GOLD, with the ecosystem information further grouped in custom categories (Table S4). To roughly estimate the host diversity present in each metatranscriptome, the taxonomic information of all contigs as predicted by the IMG annotation pipeline (Clum et al., 2021) was queried at the domain level, i.e. Bacteria, Archaea, Eukarya, and Viruses. The ratio between the number of contigs assigned to Bacteria and Archaea and the number of contigs assigned to Eukarya was then used as a proxy to determine "Prokaryote-dominated" from "Eukaryote-dominated" datasets. Specifically, datasets with a ratio of Eukaryote-affiliated to Prokaryote-affiliated contigs $\leq$ 0.3 or $\geq$0.7 were considered as "Prokaryote-dominated" or "Eukaryote-dominated", respectively, while other datasets were considered as "Mixed". The map was drawn using the packages matplotlib v3.3.4 and basemap v1.2.2 for python 3.8.5 (Hunter, 2007).

For read mapping, a dereplicated set of RNA virus sequences (95% ANI over 95% AF, established using CheckV anicalc.py and aniclust.py scripts; Roux et al., 2021), was established, hereafter "NR-mapping" dataset. Quality-trimmed reads (sensu; Clum et al., 2021) from 3,998 metatranscriptomes (Table S4) were then mapped to this dataset as follows. First, contigs from each metatranscriptome were compared to the NR-mapping dataset using blastn v2.9.0+ (E-value $\leq$ 0.01). All contigs with cumulated blast hits of $\geq$ 90% average nucleotide identity covering $\geq$ 80% of the shortest sequence were considered as putative RNA viruses. All reads mapping to contigs identified as putative RNA viruses and all unmapped reads were extracted from the existing IMG read mapping information, and mapped de-novo on the NR-mapping dataset using bbmap v38.81 (Bushnell, 2014) with the following options: "vslow minid=0 indelfilter=2 inslenfilter=3 dellenfilter=3". This step was done to reduce the computing time and the risk of false-positive mapping by excluding all reads mapping to non-viral metatranscriptome contigs. The resulting bam files were then filtered with FilterBam (https://github.com/nextgenusfs/augustus/tree/master/auxprogs/filterBam) retaining only mapping at $\geq$50% identity and $\geq$50% coverage, and genomecov from bedtools v2.30.0 (Quinlan, 2014) was used to calculated the average coverage depth for each contig in each sample. The relative proportion of a taxon was then calculated as the cumulated coverage for the taxon members divided by the total accumulated coverage of all predicted RNA virus contigs in this dataset.

Genetic code assignment and ORF calling

Presently, ORF identification software designed for diverse metagenomic data are limited to the standard genetic code (11) or the Mold mitochondrial genetic code (4) (opted when the predicted ORFs are unnaturally short). To identify clades likely to use alternative genetic codes, we extracted the RdRp core footprints and scanned them for in-frame standard stop codons.

We first separated all RdRP-encoding contigs into two subsets: ''standard'' and ''non-standard'' if any canonical stop codons occurred within the narrow coordinates of the RdRP core. Then, the ''standard'' set was subjected to metaprodigal CDS prediction using default parameters (via Prodigal's (v2.6.3) metagenomic mode (''anonymous'')) (Hyatt et al., 2010). In the ''non-standard'' subset, the stop codon usage patterns were aggregated across the contigs, associated with each tree leaf, and classified into ''mitochondrial'' (using UGA as a sense codon), and ''protist'' (other patterns). Prevalence of patterns (relative frequency among the descendant leaves) was calculated for internal tree nodes; clades with high prevalence were noted and investigated. For practical purposes, the ORFs predictions of the ''non-standard'' subset were performed by using the first genetic code enabling the entire RdRP core to be translated. Cases for which none of the available genetic codes enabled the uninterrupted translation of the RdRP core were assigned the general ''non-standard'' value and were predicted using the mitochondrial genetic code (4).

To discard the possibility of active recoding of tRNAs by these predicted RNA viruses, the VR1507 set was subjected to a single pass of tRNAscanME2 (Chan et al., 2021), using the ''global'' flag (for non-specific domain of life tRNA prediction). No tRNAs were identified on any of the viral contigs predicted to use an alternative genetic code, suggesting these are most likely an adaptation to their host rather than an element of a virus-host arms race, as seen in some dsDNA phages (Ivanova et al., 2014).

RBS identification and quantification

Using VR1507 as input, the RBS quantification was performed as described in Schulz et al. (2020). Briefly, Prodigal (v2.6.3) was run as described above (see ''genetic code assignment'') (Hyatt et al., 2010; Schulz et al., 2020), we then sourced the ''rbs_motif'' field from Prodigal's GFF output files, and classified the different 5′ UTR sequences as either ''SD'' (for motifs similar to AGGAGG, the canonical Shine-Dalgarno), ''None'' and ''Other'' (for details, see data and code availability,

''RBS_Motif2Type.tsv''). Then, for each contig, we defined the ''%SD'' as the ratio between all ''SD'' ORFs, and all ORFs with a true start (i.e. not truncated by the contigs' edge, field ''start_type'' different from ''Edge'').

Domain annotation

To perform an initial domain annotation of the proteins encoded by RdRP-containing contigs, we used hmmsearch (from the HMMER V3.3.2 suite) (Finn et al., 2011; Wheeler and Eddy, 2013) to match these proteins to hidden markov models (HMMs) gathered from multiple protein profile databases using a maximal E-value of 0.001 (PFam 34, COG 2020 release, CDD v.3.19, CATH/Gene3D v4.3, RNAVirDB2020, ECOD 2020.07.17 release, SCOPe v.1.75) (Andreeva et al., 2014, 2020; Cheng et al., 2015; Galperin et al., 2021; Lu et al., 2020; Mistry et al., 2021; Sillitoe et al., 2021; Wolf et al., 2020). We supplemented this set of HMMs with a custom collection of profiles with bacteriolytic functions (termed ''LysDB'' - available in the project's Zenodo repository, see data and code availability). LysDB was built from (1) manually reviewed profile entries from public databases which we could link to GO terms related to cell lysis by viruses, or virus exit from host cell, and (2) custom profiles for ''Sgl'' proteins, which were experimentally demonstrated by Chamakura et al to induce cell lysis (Chamakura et al., 2020). Additionally, we used InterProScan (v.5.52-86.0) to scan the protein sequences using MobiDBLite (v2.0), Phobius (v.1.01), PRINTS (v. 42.0), TMHMM (v.2.0c) (Attwood et al., 2012; Jones et al., 2014; Käll et al., 2004; Käll et al., 2007; Krogh et al., 2001; Potenza et al., 2015).

Because the public protein profile databases that were used for initial annotation might contain HMMs that represent polyproteins, which span multiple functional domains, we developed and employed a procedure to identify such profiles which were masked from the subsequent annotation process. For this procedure, we first used the hmmemit command to convert HMMER profiles into multiple sequence alignments, which were then used as input to an all-versus-all profile comparison performed using HH-Suite. Next, putative polyprotein profiles were identified by flagging the profiles that encompassed at least two other non-overlapping profiles (''get_polyproteins.ipynb'' script, see data and code availability). The unmatched regions between the polyprotein domains were extracted to create a set of conserved, yet unknown domains, termed ''InterDomains''. Additionally, profiles with over 1000 match states (defined as columns with less than 50% gaps) were manually examined using HHpred. Several of the identified polyprotein profiles were split into their constituent domains. Subsequently, all hmmsearch results were aggregated and profile matches were prioritized based on their classification level (uncurated profiles, or ones of unknown function (e.g. ''DUF'') were deprioritized) and by their relative alignment statistics. To improve the quality of the functional annotation of the domain profiles and to assign functions to unannotated profiles we identified clusters of similar profiles (clans, hereafter). First, profiles with at least one hit in the initial annotation pass were extracted from their original DB, reformatted as HH-Suite's HHMs (as described above) and used for an additional all-versus-all step. The output of this profile comparison was then used as input to a graph-based clustering process using the Leiden algorithm (''get_clan_membership.ipynb'' script, see data and code availability), which identifies clans as communities of highly similar domains. Clan membership was then used to improve the coverage of the functional annotation by transferring annotation from functionally annotated profiles to other clan members. Briefly, this procedure followed a consensus-based label assignment. For example, a clan with 12 profiles labeled as ''RdRP'', and 2 ''unclassified'' profiles, was set as an ''RdRP'' clan and the 2 unclassified members were reclassified as ''RdRP''. Cases of conflicts were either left unresolved, or by opting to the lowest denominator. For example, a clan with 4 ''unclassified'' profiles, that also had 12 member profiles labeled ''Super family 2 Helicase'' and an

additional 10 member profiles labeled ''Super family 1 Helicases'', was set to ''Helicase-uncertain'', and this label was extended to those 4 ''unclassified'' members.

All subsequent profile matches passing a predefined cutoff (E-value $\leq$ e-7, score $\geq$ 9, alignment length $\geq$ 8[AA]). were used to generate a new custom profile database, in a process similar to the one used for RdRPs (see above). Only clusters with $\geq$ 10 sequences, sharing the same functional classification, were used to generate HMMs. This profile set was then supplemented by most of the profiles from the above-mentioned RNAVirDB2020 database, as well as several dozen select profiles from the other databases (this final profile database termed ''NVPC'' is available via the projects Zenodo repository, see data and code availability). Finally, we queried the six-frame translations of the **330k** contig set using hmmsearch as described above, using the new profile database. (Figure S3 - Annotation pipeline). Subsequently, we generated tentative genome-maps for $\approx$4-20 representative contigs for each of the 400+ identified families (novel and established) using GGGenomes (https://github.com/thackl/gggenomes), which were then manually examined to identify novel domains as well as uncommon domain fusion and segmentations.

Quality control and reliability of metatranscriptomic assemblies

Metagenomic assemblies are prone to various types of artifacts that can result in apparent contigs in the assembly that do not represent any existing nucleic acid molecules in the original biological sample (Arroyo Mühr et al., 2020). Notoriously, chimeras (contigs mis-assembled from at least two different nucleic molecules) can be a major setback for novelty claims and can be difficult to identify and separate from real genetic entities. We addressed this concern by implementing several stringent procedures to avoid any misinterpretation that could stem from the analysis of potentially chimeric contigs:

1. Firstly, no claims in this work are based on singletons. Rather, we only report observations based on the analysis of evolutionarily conserved stemming groups of sequences (two or more alignable contigs, ideally, from multiple assemblies) or from features conserved at the coarse phylogenetic level (family-level and above). The likelihood of the chimera recurring across multiple assemblies appears negligible.
2. Secondly, when unexpected observations were made, such as those on genome rearrangement, gene fission and gene fusion, we manually inspected each case at the read level, that is, traced the original sequencing runs and mapped (via the procedure described above in the section ''Habitat distribution and relative abundance estimation'') the raw Illumina short reads to the contigs in question, and examined the distribution of reads along the assembled contigs, checking that the contigs (and not only the RdRP-coding region) were well covered. Contigs in which some portions showed abnormally low coverage or skewed GC% content were deemed unreliable and discarded.
3. We observed and removed several dozen contigs from the set we built by aggregating published sources as likely chimeras (mostly, part levivirus, part rRNA). Prompted by this observation, we searched the entire VR1507 contig set against the SILVA rRNA database (BLASTn against SILVA SSU & LSU Ref NR99, default parameters) (Quast et al., 2013), and manually examined 40 contigs encoding ribosomal proteins identified in the ''domain annotation'' section, to ribosomal protein profiles in the public databases (e.g. Ribosomal protein L3 PF00297.24) Over all, we flagged 75 potential chimeras of these types, (23 of which originate from the previously published sources, see Table S6, sheet ''rRNA_summary'' for details). Only the RdRPs of these suspect contigs were used in downstream analyses, whereas the rest of the contig was disregarded.
4. The DNA subtraction we performed drastically reduced the abundance of chimeras that consisted in part from RNA virus sequences and in part from DNA encoded ones, whether rRNA or mRNA. Obviously, however, this procedure cannot eliminate chimeras that consist of portions of different RNA virus genomes. Because such chimeras would be difficult to differentiate from bona fide recombinant virus genomes, we employed a heuristic to identify these using the domain annotations to detect contigs with duplicated full-length RdRP footprints. These were deemed chimaeric because RNA viruses normally encode a single (full length) RdRP. We found a single such case, ND_250651, a chimera that is part levivirus, part cystovirus.

Quantitative comparison with recently published RNA virus discovery endeavors

44,779 RdRPs from the Tara project were downloaded from https://datacommons.cyverse.org/browse/iplant/home/shared/iVirus/ZayedWainainaDominguez-Huerta_RNAevolution_Dec2021. Serratus project RdRPs were represented by 296,623 unique PalmDB sequences that were downloaded from https://github.com/rcedgar/palmdb repository. The Serratus sequences represent a tightly defined RdRP core (containing only motifs A, B and C) with a median length of 107 (compared to the 453 aa for the RCR90 set). Of note, our study, the Tara project, and the Serratus projects, each defined differently which regions of the RdRP could be used for MSA and subsequent phylogenetic analysis. Hence, we restricted our comparison to the RdRP region closest to a lowest common denominator between the studies, which is the region shared and defined by palmDB. We performed this by using all 329,202 unique RdRP sequences from this study and the 44,779 RdRPs from the Tara project, for a BLASTP search (e-value 0.0001) against the PalmDB set, using the best hit to trim the query (specifically, with the query of length K and the hit footprint of p1..p2 against the subject of length *L* and the hit footprint of *q1..q2*, the query was trimmed to max(1,*p1*-q1-1)..min(*K,p2+L-q2*) to account for missing parts of the subject). Queries without a significant hit to PalmDB were left untrimmed. The full set of sequences was pooled together and clustered using MMseqs2 with sequence identity thresholds of 0.9 and 0.5 (–min-seq-id 0.5/0.9 -c 0.333 -e 0.1 –cov-mode 1 –cluster-mode 2). All sequences were classified into four categories: i) "known" (GenBank and other published sources from the current

dataset (see STAR Methods "contig set augmentation with published genomes"), ii) "RVMT" (RNA Virus MetaTranscriptomes from the current dataset), iii) Serrarus and iv) Tara. Clusters were examined for the presence of members from each of the four sets, and the cluster set intersections are listed in Table S8.

QUANTIFICATION AND STATISTICAL ANALYSIS

Exact thresholds, including the expect value (E-values), for all analyses derived from sequence searches or alignments procedures (e.g domain prediction, CRISPR spacer matching, etc) are provided in the relevant main text or in method details, and in Table S6 (sheet "filtration thresholds" for E-values used in DNA filtration process, and sheet "Clustering_information" for clustering thresholds and associated quantification).

ADDITIONAL RESOURCES

In hope of providing a long lasting community resource, we created an accompanying interactive web portal (riboviria.org) that allows users to download portions of the data generated in this work based on phylogeny and data type (e.g., a subset of the domain annotations for all contigs affiliated with a certain family). Both programmatic and graphical access to the data are supported through the web portal. The website's code is also available under the MIT License at github.com/Benjamin-Lee/riboviria.org. For all taxonomic levels, this platform includes raw nucleic sequence, phylogenetic trees, metadata, and annotations.

CellPress
OPEN ACCESS

Supplemental figures

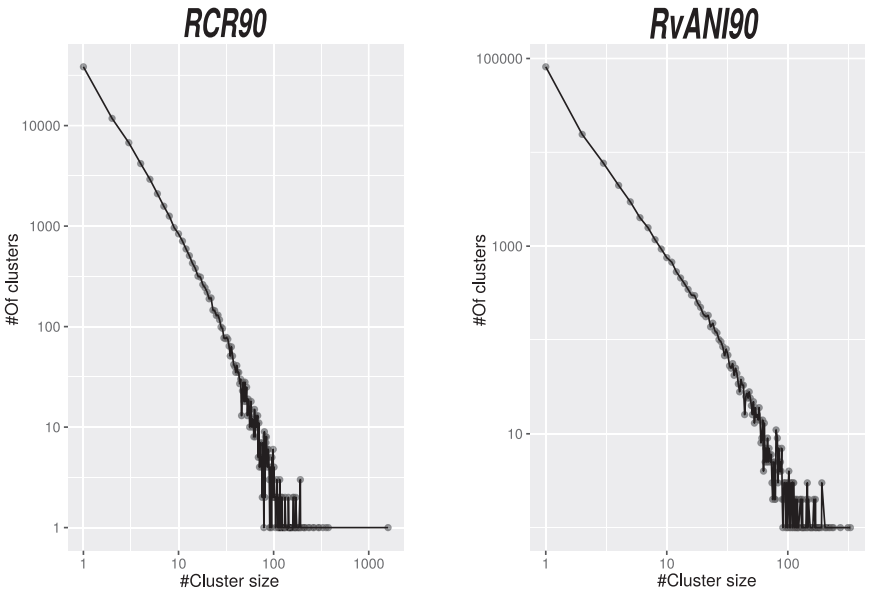


Figure S1. Distribution of contigs in RCR90/RvANI clusters, related to Figures 1B and 1C and Table 1

Each panel displays the total number of clusters (left panel RCR90, right panel RvANI90) on the horizontal axis (logarithmic scale) against their size (total number of membering contigs) on the vertical axis (logarithmic scale).

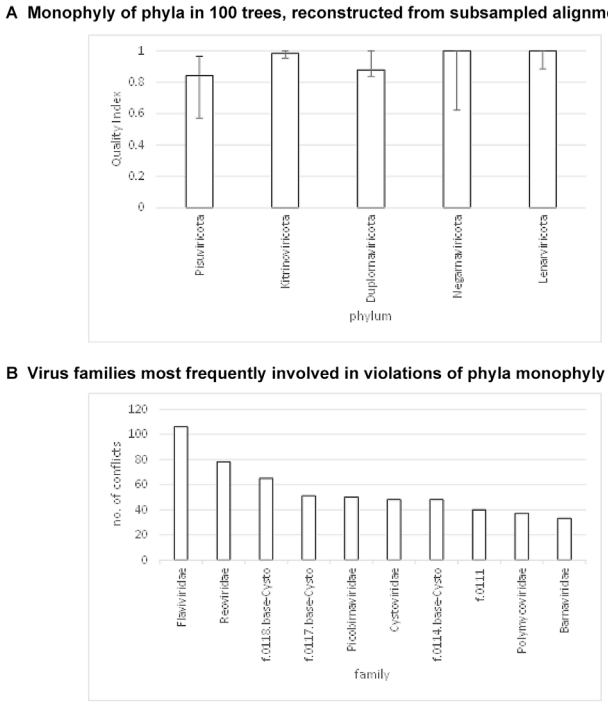


Figure S2. Robustness of deep phylogenetic reconstructions, related to Figure 2

(A) Quality index (the product of the fraction of phylum members that form a monophyletic clade and the fraction of other phyla members in this clade). The bar shows the median value across 100 independent samples of one member of a family with at least 20 members; the whiskers indicate the 5% and 95% percentiles.

(B) The virus families, most often involved in monophyly violations (where a leaf is either outside of the clade of its phylum or inside a clade of the other phylum). The number of violations is shown.

(C) The extended majority consensus tree of the five previously known phyla. The consensus tree was recovered from 85 (out of 100) samples that have non-embedded monophyletic phyla, and the support values were multiplied by 0.85.

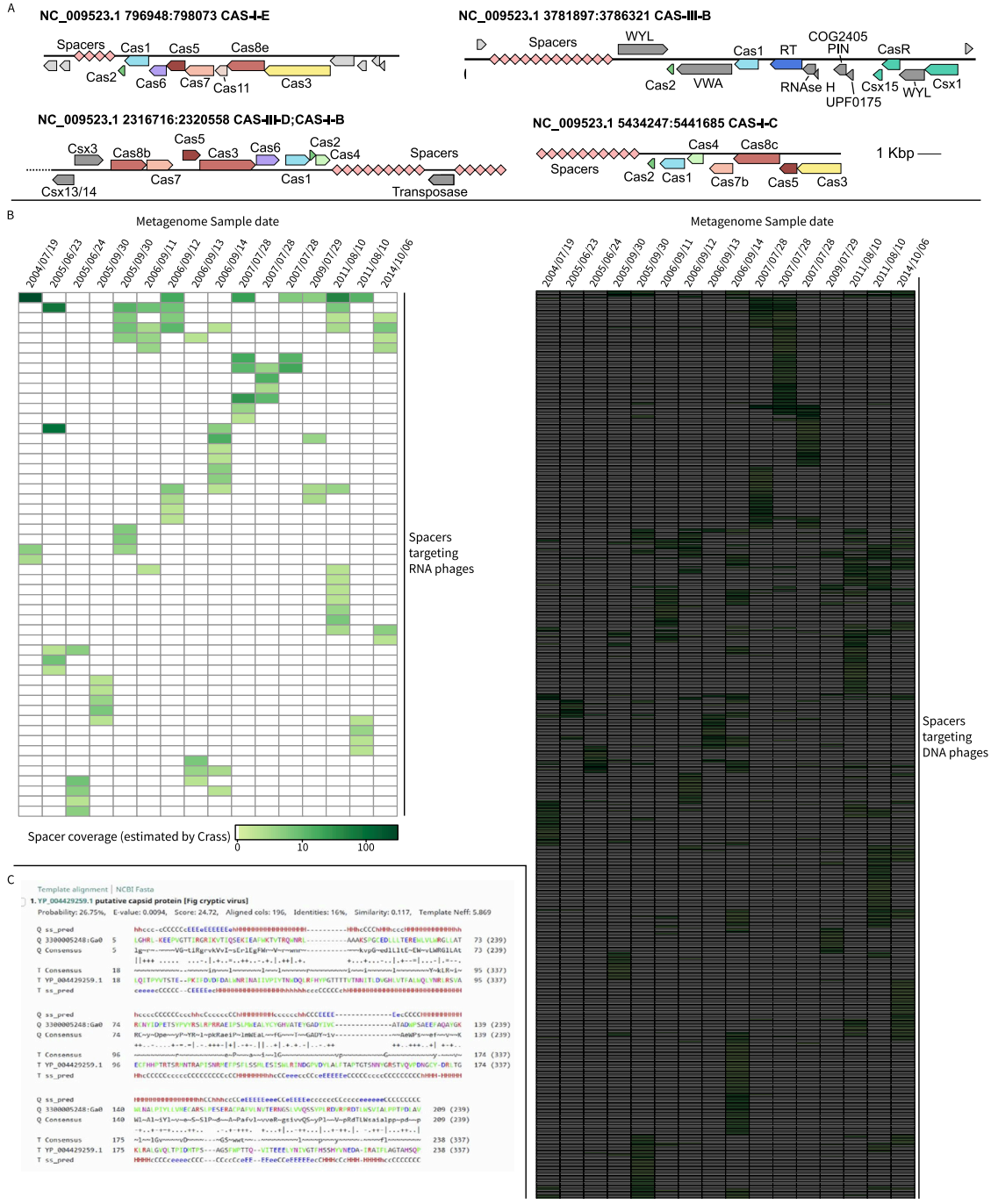


Figure S3. *Roseiflexus* sp. RS-1 CRISPR arrays and related viruses, related to Figure 3

(A) Map of the 4 CRISPR-Cas regions in *Roseiflexus* sp. RS-1 (NC_009523.1) including predicted CRISPR arrays (red diamonds) and Cas genes (colored genes).

(B) Coverage heatmaps across Mushroom Spring and Octopus Spring metagenomes, for spacers associated with *Roseiflexus* sp. RS-1 (see Table S3). Spacers matching predicted RNA phages are displayed on the left, and spacers matching DNA phages are displayed on the right for reference.

(C) Example of alignment obtained with hhpred for a putative capsid protein from a predicted novel RNA phage infecting *Roseiflexus* sp. RS-1 and the closest publicly available homolog: fig cryptic virus capsid protein.

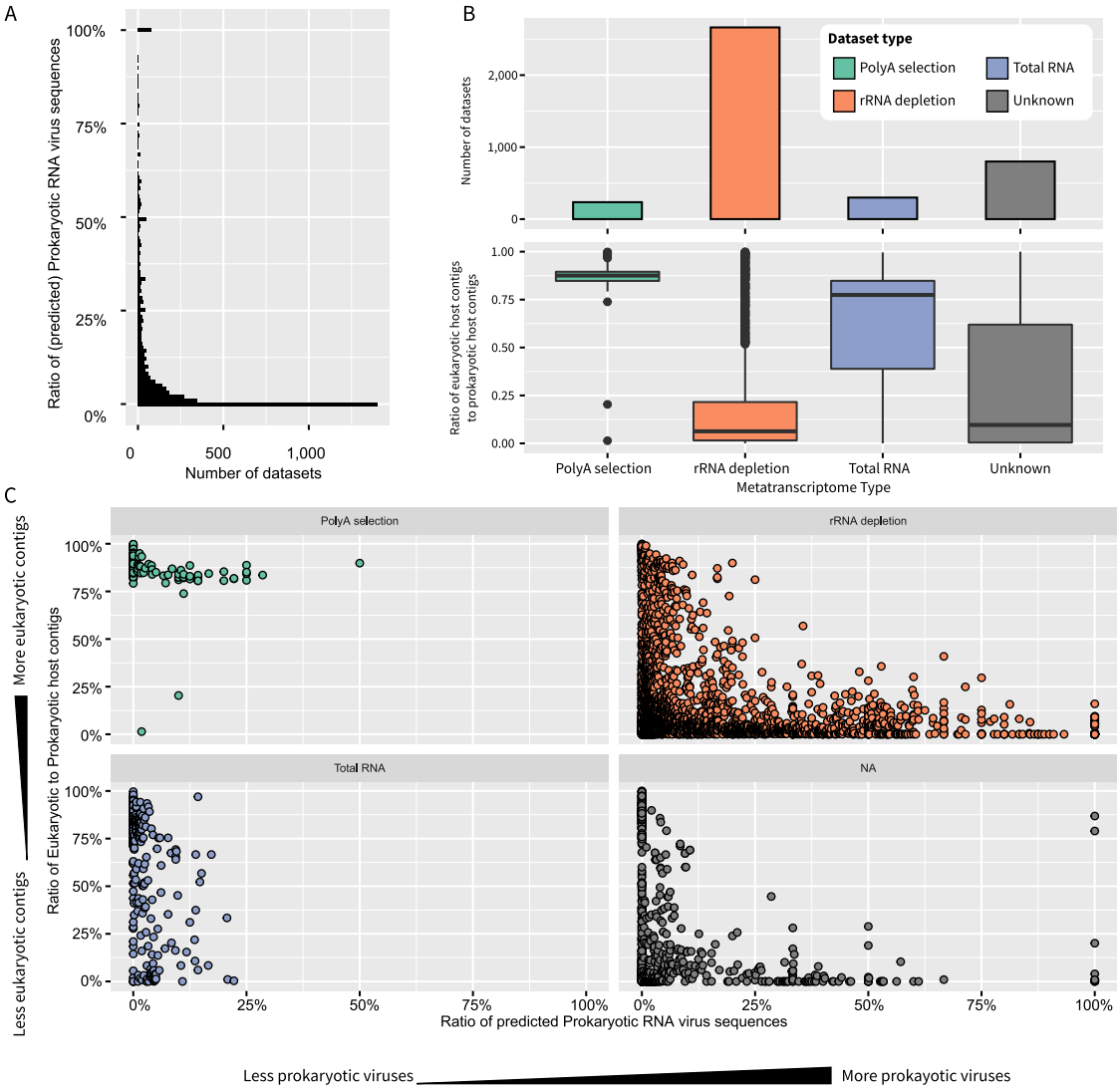


Figure S4. Identification of different metatranscriptome types and associated virus types, related to Figures 3 and 4

(A) Distribution of the ratio of viruses predicted to infect prokaryotic hosts across individual samples.

(B) Distribution of non-viral contigs affiliated as eukaryotes or prokaryotes (hosts) across samples, separated based on the protocol used to generate the metatranscriptome. The protocol information was obtained from the Gold, and summarized as follows: "poly(A) selection": transcript enrichment based on poly(A) tail, "rRNA depletion": use of a kit(s) and/or protocol(s) for depletion of rRNA templates, "total RNA": cDNA library prepared from the extracted RNA with no poly(A) selection or rRNA depletion step, "unknown": no information available.

(C) Relationship between the ratio of eukaryote/prokaryote RNA viruses (x axis) and the ratio of eukaryote/prokaryote host contigs (y axis). Each dataset type is presented in a separate panel.

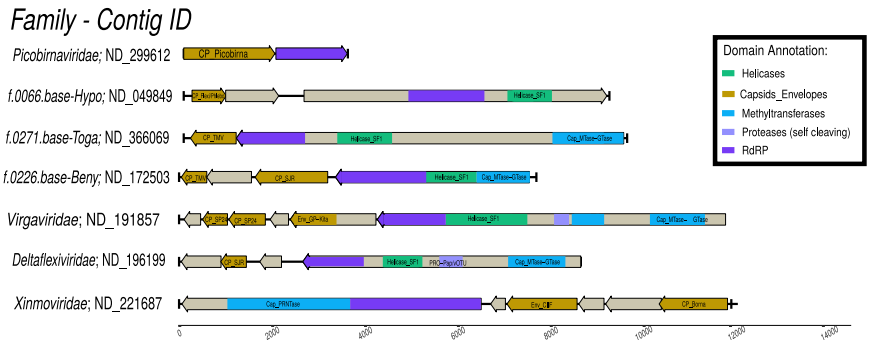


Figure S5. Acquisitions and replacements of structural modules in RNA viruses, related to Figure 5

"*Picobirnaviridae*; ND_299612" and "f.0226.base-Beny; ND_172503" exemplify fusions of genomic segments encoding capsid proteins (CPs) and RdRPs, which are encoded on separate segments in previously described picobirnaviruses and benyviruses. "f.0066.base-Hypo; ND_049849" and "*Deltaflexiviridae*; ND_196199" encode Flexi/Phlebo-like CP and single jelly roll (SJR) CPs, respectively, although other members of the respective families comprise capsid-less viruses. "*f.0271.base-Toga*; ND_366069" and "*Virgaviridae*; ND_191857" represent genomes with non-homologous replacements of the CP genes. In "*Xinmoviridae*; ND_221687," class III fusion glycoprotein gene, typical of xinmoviruses, has been replaced by a gene encoding a class II fusion glycoprotein (CIIF). Abbreviations: Env, envelope protein; GP, glycoprotein; PRO-Pap/vOTU, papain-like protease; SF1, superfamily 1; Cap_MTase-GTase, capping enzyme with methyltransferase-guanylyltransferase activities.

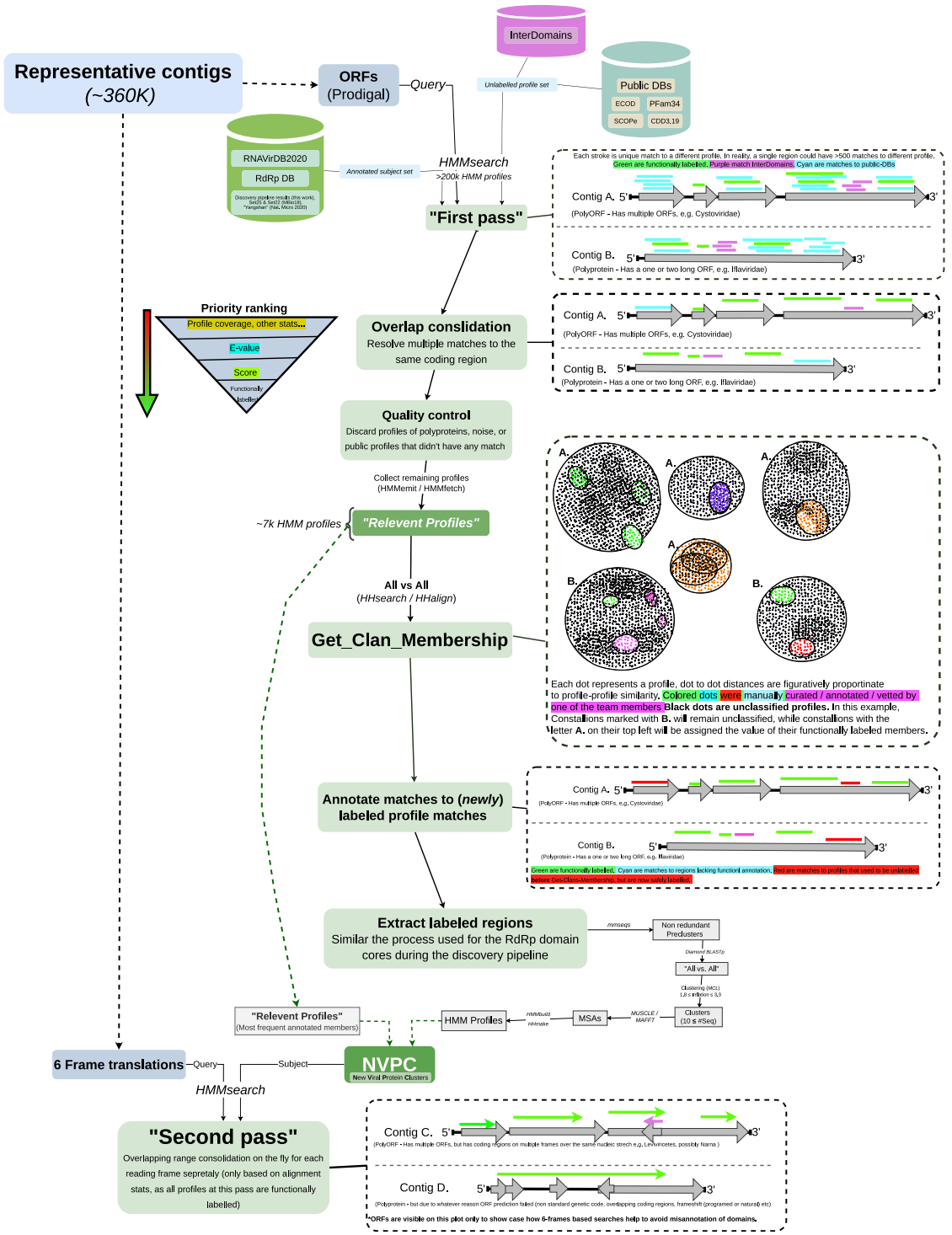
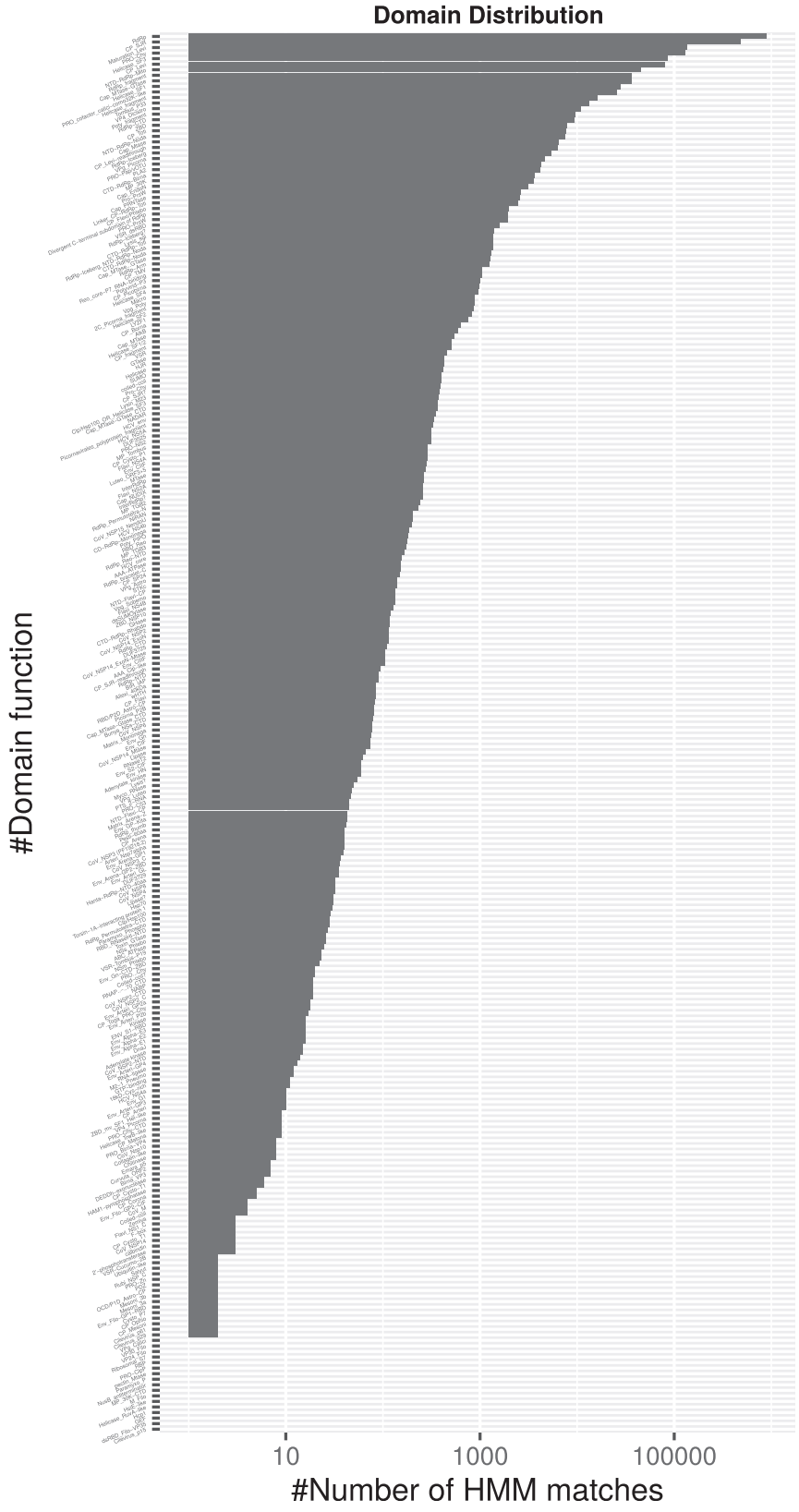


Figure S6. Extended annotation pipeline, related to Figure 5

Flowchart diagram visualizing the procedures used in the domain identification and functional annotation sections of the project.



(legend on next page)

Figure S7. Identified domain distribution, related to Figure 5

The predicted viral function or structure of the final domain hits (vertical axis, slanted text labels), against the total number of reliable observed HMM search matches (horizontal axis, logarithmic scale).