Identification of RNA Virus–Derived RdRp Sequences in Publicly Available Transcriptomic Data Sets

Ingrida Olendraite, Katherine Brown, and Andrew E. Firth*

Division of Virology, Department of Pathology, Addenbrookes Hospital, University of Cambridge, Cambridge, United Kingdom

***Corresponding author:** E-mail: aef24@cam.ac.uk.

Associate editor: Thomas Leitner

Abstract

RNA viruses are abundant and highly diverse and infect all or most eukaryotic organisms. However, only a tiny fraction of the number and diversity of RNA virus species have been catalogued. To cost-effectively expand the diversity of known RNA virus sequences, we mined publicly available transcriptomic data sets. We developed 77 family-level Hidden Markov Model profiles for the viral RNA-dependent RNA polymerase (RdRp)—the only universal "hallmark" gene of RNA viruses. By using these to search the National Center for Biotechnology Information Transcriptome Shotgun Assembly database, we identified 5,867 contigs encoding RNA virus RdRps or fragments thereof and analyzed their diversity, taxonomic classification, phylogeny, and host associations. Our study expands the known diversity of RNA viruses, and the 77 curated RdRp Profile Hidden Markov Models provide a useful resource for the virus discovery community.

Key words: **RNA virus, RdRp, Orthomyxoviridae, splicing, pHMM, virus discovery.**

Introduction

RNA viruses evolve rapidly and are extremely diverse, with an ancient evolutionary ancestry (Koonin et al. 2015). They are highly abundant in eukaryotic hosts, whereas the viromes of prokaryotes tend to be dominated by DNA viruses (Nasir et al. 2014, Koonin et al. 2015, though cf. Callanan et al. 2020 and Neri et al. 2022). Genomes range in size from ∼2 to ∼41 kb (Nibert 2017; Saberi et al. 2018). Over long evolutionary timescales, RNA viruses evolve via extensive recombination of protein-coding genes between different lineages, between virus and host, and via de novo gene formation (Keese and Gibbs 1992; Shi et al. 2016; Dolja and Koonin 2018; Wolf et al. 2018). The unique hallmark gene of RNA viruses—and indeed the only protein that is ubiquitously conserved throughout all RNA viruses—is the RNA-dependent RNA polymerase (RdRp). Despite immense divergence and, in some cases, almost imperceptible similarity at the primary sequence level, structural studies have confirmed homology—and therefore shared evolutionary ancestry—between the RdRps of all three Baltimore RNA virus groups (Mönttinen et al. 2021; Jácome et al. 2022). Besides being the only hallmark protein of RNA viruses, the RdRp is also relatively highly conserved compared with other RNA virus proteins and is therefore the most appropriate protein by which to identify new RNA viruses with homology search algorithms such as BLASTP (Altschul et al. 1990) or HMMER (Eddy 2011).

In the traditional Baltimore Classification system, RNA viruses are classified into three groups based on the nature of their genomic nucleic acid: single-stranded positive-sense RNA (+ssRNA) viruses, double-stranded RNA (dsRNA) viruses, and single-stranded negative-sense RNA (−ssRNA) viruses (Baltimore 1971). "Positive-sense" refers to the coding sense, although a small number of RNA viruses have also evolved additional genes on the opposite strand (Nguyen and Haenni 2003; Dinan et al. 2020). During replication in a host cell, all RNA viruses must obviously produce the complementary RNA. Thus, the virus genome is defined to be the nucleic acid that is packaged into virus particles for extracellular transmission. A small proportion of viruses lack capsids, and here, the appropriate Baltimore group is determined by phylogenetic relatedness to viruses with capsids. Depending on species, RNA viruses can have segmented or nonsegmented genomes that are almost always linear, but in a few cases (notably chuviruses and the recently discovered ambiviruses) are or appear to be circular (Li et al. 2015; Forgia, Navarro, et al. 2022; Lee et al. 2022).

Deep phylogenetic analysis of RNA virus RdRps has outlined five putative major branches of RNA viruses (Wolf et al. 2018; Koonin et al. 2020). The basal Branch 1 comprises +ssRNA narna-, mito-, and ourmiaviruses and their bacterial levivirus relatives. Branch 2 comprises the vast picornavirus-like supergroup including +ssRNA picorna-, calici-, polero-, nido-, astro-, and potyviruses and the dsRNA partiti-, picobirna-, amalga-, and hypoviruses. Branch 3 includes the +ssRNA alpha-, flavi-, tombus-, noda-, yan-, zhao-, and weiviruses and their relatives. Branch 4 comprises the dsRNA reo-, toti-, chryso-, megabirna-, quadri-, and giardiaviruses, besides their bacterial cystovirus relatives. Finally, Branch 5—predicted to have evolved out of Branch

4—comprises all known −ssRNA viruses, including bunya-, orthomyxo-, mononega-, chu-, qin-, and yueviruses. These branches were subsequently designated as phyla, with Branches 1–5 corresponding to phyla Lenarviricota, Pisuviricota, Kitrinoviricota, Duplornaviricota, and Negarnaviricota, respectively. Some recent studies have indicated the existence of additional phylum-level groups (Neri et al. 2022; Zayed et al. 2022), though, beyond sequence data, very little is known about the members of these new clades. Virus taxonomy is supervised by the International Committee on Taxonomy of Viruses (ICTV) and, over the past several years, has been in a state of flux as the system adapts to the "metagenomic era," adopts additional levels of classification, and moves to binomial genus/species names to increase consistency with the Latin binomials used for cellular organisms. In this study, we used the ICTV 2017 taxonomy (except when referring to more recent literature) (King et al. 2018).

The ~5,500 RNA virus (kingdom Orthornavirae) species currently represented in GenBank constitute just a tiny fraction of the estimated millions of RNA virus species on Earth (Geoghegan and Holmes 2017; Kuhn et al. 2019; Dance 2021; Harvey and Holmes 2022). Recent high-throughput RNA sequencing (RNA-Seq) studies—performed with the express purpose of identifying RNA viruses—have revealed vast numbers of novel RNA viruses, and many new family-level virus clades in diverse eukaryotic host organisms (Cook et al. 2013; Li et al. 2015; Shi et al. 2016, 2018; Olendraite et al. 2017; Charon et al. 2020; Chiapello et al. 2020; Sutela et al. 2020; Wolf et al. 2020; Batson et al. 2021; Chen et al. 2022; Forgia, Chiapello, et al. 2022; Kinsella et al. 2022; Rosario et al. 2022; reviewed in Dolja and Koonin 2018; Greninger 2018; Obbard 2018; Zhang et al. 2019; Cobbin et al. 2021; Harvey and Holmes 2022). However, a much larger number of RNA-Seq studies are performed for projects that are unrelated to virus discovery, but instead aim to study the transcriptomes of the targeted cellular organisms. The results of these studies are often deposited in the National Center for Biotechnology Information (NCBI) Short Read Archive (SRA; raw sequencing reads) and Transcriptome Shotgun Assembly (TSA; assembled contigs) databases (Sayers et al. 2022). It should be noted that only a small fraction of SRA data sets have currently been assembled into TSA data sets. Nonetheless the TSA database still holds assemblies for thousands of RNA-Seq data sets from diverse host organisms. By analyzing public RNA-Seq data in the TSA database, we aimed to expand the known diversity of RNA viruses and their host species.

Profile Hidden Markov Models (pHMMs) provide a more sensitive method than BLASTP for finding distant homologues of known protein sequences while (in contrast to structure-based approaches) maintaining computational speed (Eddy 2011). In this analysis, we used known RNA virus sequences to develop 77 family-level RdRp pHMMs. We used these to search with HMMER (Eddy 2011) for sequences encoding putative RNA virus RdRps in the TSA database, supplemented with virus sequences from the NCBI RefSeq (ref) and nonredundant nucleotide (nr/nt) databases. We identified 12,109 RdRp-encoding sequences and analyzed their diversity, taxonomic classification, phylogeny, and host associations. We provide a listing of all sequences found and our curated family-level pHMMs, both of which will be useful resources for evolutionary, taxonomic, host association, ecological, and comparative genomic studies.

Results

Enriching Viral RdRp Diversity

Profile Hidden Markov Models (pHMMs) provide a fast and sensitive method for identifying distant homologues of protein sequences (Eddy 2011). We began by constructing pHMMs for the RNA-dependent RNA polymerase (RdRp) proteins of 77 RNA virus family-level clades (see Materials and Methods). We used these pHMMs together with HMMER (Eddy 2011) to search for candidate RdRp-encoding sequences in the TSA database, supplemented with virus sequences from the NCBI RefSeq (ref) and nonredundant nucleotide (nr/nt) databases. Since some RNA viruses are segmented, a caveat with this strategy is that it is not always possible to retrieve entire virus genomes.

We first discarded nr/nt and ref sequences that had ≥80% nucleotide identity to longer sequences. Next, for the TSA sequences and remaining nr/nt and ref sequences, we extracted all ORFs of length ≥60 nucleotides and applied HMMER to each ORF (see Materials and Methods). A total of 15,044 putative virus RdRp sequence fragments with a significant match ($P \leq 10^{-6}$) against at least 1 pHMM were identified, corresponding to 12,136 translated ORFs in 12,109 unique accessions (supplementary data set 1, Supplementary Material online). A total of 5,867 accessions were from the TSA database. These sequences were sorted into "classified," "ambiguously classified," or "unclassified" groups, based on the HMMER bit score divided by the length in amino acids of the alignment between an ORF and a matched pHMM (referred to herein as "IDscore"). In brief, if the highest IDscore for a sequence:profile match was <0.25, a sequence was sorted into the unclassified group. Otherwise, if a sequence had statistically significant hits to more than one of the pHMMs and the top two IDscores were within 20% of each other, the sequence was sorted into the ambiguously classified group. Finally, sequences with an IDscore of 0.25 or higher, and at least 20% difference in IDscore between the first and second best hits, were sorted into the classified group and classified according to the pHMM of the top IDscore. Fifty-five percent of sequences were sorted into the classified group, 43% into the ambiguously classified group, and 2% into the unclassified group (fig. 1A).

In most cases, after trimming translated ORF sequences to the RdRp core (i.e., the best pHMM match positions), the identified RdRp sequences were between 400 and 500 amino acids (a.a.) in length (fig. 1A). Typically, actual

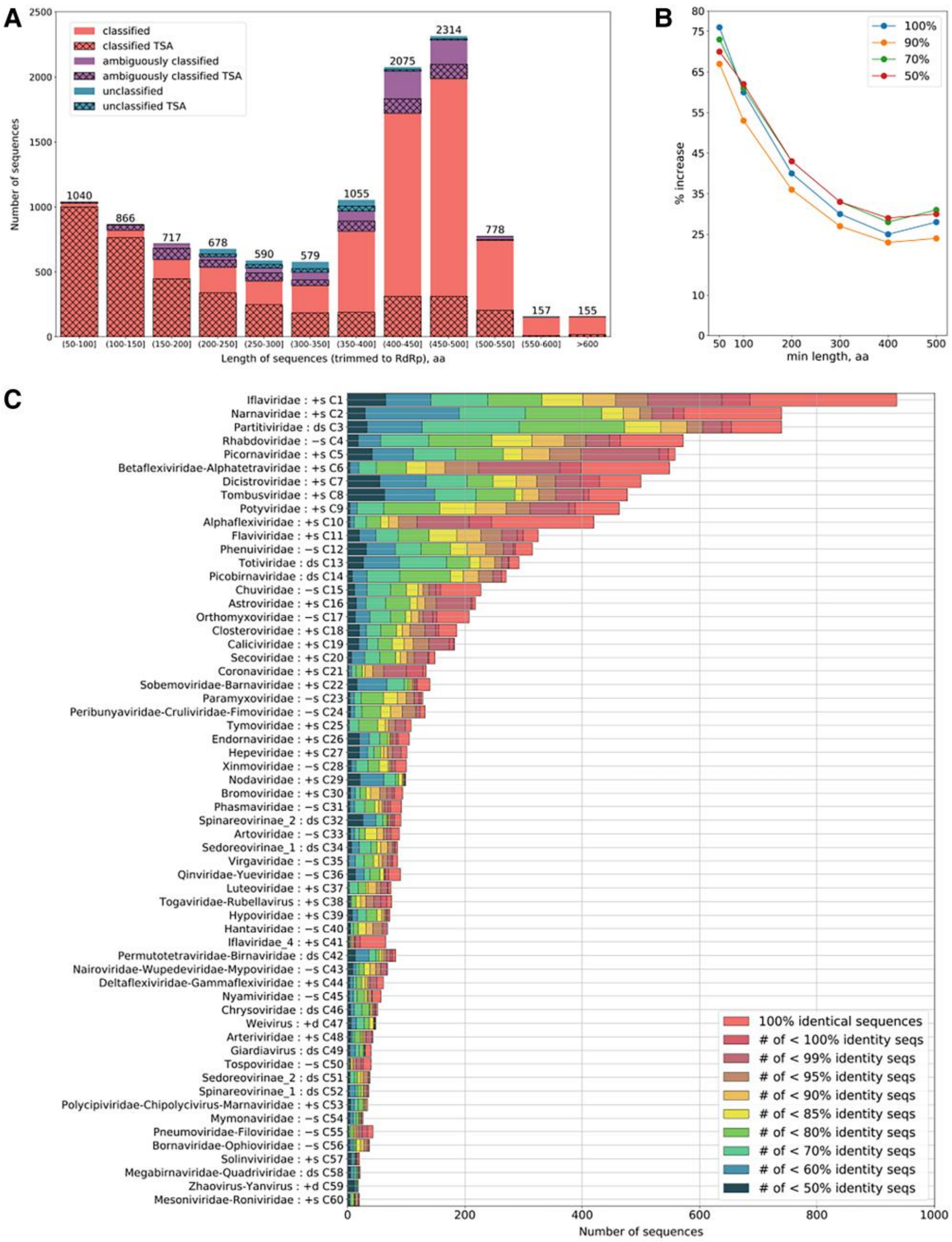

Fig. 1. Various metrics of identified sequences. (*A*) Numbers of identified RdRp-encoding ORFs (ref, nr/nt, and TSA) and their lengths after trimming to the RdRp core (see main text) and removing duplicate identical sequences. (*B*) Percentage increase in the number of RdRp clusters as a function of trimmed RdRp core fragment length (*x*-axis) and clustering identity threshold, upon adding the TSA-derived sequences to the nr/nt and ref sequences. The *y*-axis shows the percentage increase in clusters after using different CDHIT (Li and Godzik 2006; Fu et al. 2012) identity thresholds (50%, 70%, 90%, and 100%, as indicated) for nr/nt + TSA sequences compared with nr/nt sequences alone. (*C*) Numbers of sequences identified in each cluster at different pairwise amino acid identity thresholds. Duplicate identical sequences were removed. Identities were calculated via pairwise alignment in Biopython (Cock et al. 2009, see Materials and Methods) and dividing the number of identical aligned residues by the shorter sequence length.

full-length viral RdRp core sequences range approximately from 350 to 600 a.a. in length (depending on virus clade). Most of the shorter sequences found derived from the TSA database. This was expected, as transcriptome shotgun assembly often results in fragmented rather than full-length sequences (Bushmanova et al. 2019). Thus, many of the RdRp sequences found—but especially the shorter ones—are expected to be incomplete.

We verified that putative RNA viral ORFs were more similar to a known RNA virus than to any non–RNA viral sequence by using a low stringency BLASTP search (Altschul et al. 1990) against the nonredundant protein (nr) database (supplementary data set 2, Supplementary Material online). This allows us to identify any sequences potentially derived from other sources, such as cellular genes, retrotransposons, or DNA viruses. All sequences showed significant similarity ($E \leq 0.05$) to a known RNA virus sequence. Twenty-five had better scores ($E \leq 0.05$) against another sequence not derived from an RNA virus. In silico assembly can sometimes lead to artefactual virus–virus or virus–host chimeric sequences (reanalyzing the TSA source data is possible, but was not attempted in this study). It is also possible that some TSA sequences derive from transcribed genome–integrated virus fragments (known as endogenized virus elements or EVEs; Katzourakis and Gifford 2010; Aiewsakun and Katzourakis 2015; Gilbert and Belliardo 2022); in this work, we do not distinguish viral sequences from transcribed EVEs. Either of these scenarios could result in sequences with both RNA viral and non–RNA viral regions, and either could explain 19 of the 25 sequences which had higher BLAST bit scores against non–RNA viral proteins than against viral RdRps. These sequences have <25 nt of overlap between the best viral and best nonviral hit, so they are adjacent on the sequence, as expected for chimeric and EVE-derived sequences. For the remaining six sequences, which have a larger overlap, there appear to be errors in the GenBank records for the nonviral proteins with which they share similarity: four match GenBank accessions QHN95476.1, RDY14258.1, or PKA56654.1, which have regions labeled as mitovirus RdRp within cellular genes in their GenBank annotation; one matches AWS06671.1, which is named as a retrovirus but appears to be Darwin bee virus 2; and one matches KMQ91513.1, labeled as an ant glucuronate isomerase but with strong BLAST similarity to the RdRp from a number of insect viruses. It is likely that these five proteins are actually RNA viral in origin.

We also used HHSearch (Steinegger et al. 2019) against the Pfam database (Finn et al. 2014) to confirm the presence of RNA viral domains in our putative viral ORFs (supplementary data set 2, Supplementary Material online). In total, 98.9% of sequences contained recognizable viral domains (including RdRp in 98.4% of cases). The sequences not identified using this method were primarily (250 of 281 cases) members of the orders Tymovirales, Durnavirales, and Patatavirales, all of which form parts of phylum-level Pfam pHMM models, which may be less sensitive than other order- or class-level models.

Whereas the ambiguously classified and unclassified groups contain many novel and divergent viral sequences (discussed further below), sequences in the classified group share substantial similarity with previously known sequences from currently existing virus families. Nonetheless, we wanted to investigate the extent to which the addition of the TSA-derived classified RdRp sequences increases the diversity of RdRps within currently defined RNA virus groups. For each of the 77 (mostly family level) pHMMs, we took all the TSA, ref, and nr/nt RdRp sequences that were classified to that group and, using BLASTP, compared all versus all pairwise identities within each group to find the greatest pairwise divergences by group (supplementary fig. S1, Supplementary Material online). For several of the pHMM groups (e.g., roni-, fimo-, yue-, xinmo-, arto-, deltaflexi-, nyami-, and chu-like viruses), the diversity was substantially increased with the addition of the TSA-derived sequences. In some such cases, this may indicate the presence of novel sister clades (or sister families) within the group. Inspired by Wolf et al. (2020), we also plotted the percentage increase in number of RdRp clusters, as a function of trimmed RdRp core fragment length and clustering identity threshold, upon adding the TSA-derived sequences to the nr/nt sequences. For a length cutoff threshold of 400 a.a., we saw about a 30% increase in the number of RdRp clusters at CDHIT identity thresholds of 70% or 50% (fig. 1B).

After the sorting step, some RdRp pHMM groups had hundreds of sequences, whereas others had very few. For simplicity of presentation, we appended the groups containing very few sequences to larger RdRp groups, based (to a certain extent) on their RdRp similarities and the size of the larger groups. This resulted in 60 clusters, with sizes and amino acid sequence diversity as illustrated in figure 1C. PhyML trees of the different virus groups are available in supplementary data set 3, Supplementary Material online.

We also compared the number of RdRp sequences found per group in the TSA and in the nr/nt databases (fig. 2 and supplementary fig. S2, Supplementary Material online). Our TSA search revealed large numbers (>100 per group) of new ifla-, narna-, partiti-, betaflexi-, dicistro-, alphaflexi-, rhabdo-, tombus-, chu-, poty-, phenui-, orthomyxo-, toti-, and picobirna-like RdRps. New TSA sequences also represented >75% of some groups such as the ifla_4-, arto-, xinmo-, deltaflexi-, ifla-, qin-, giardia-, nyami-, yue-, chu-, alphaflexi-, phasma-, and narna-like viruses [note, ifla-4 denotes a partition of the iflaviruses in the Aiewsakun and Simmonds (2018) groupings that we used for pHMM production]. Many of these virus groups have previously been found to be associated with arthropods, fungi, plants, or protists. On the other hand, for virus groups that previously have been found to be mainly or strictly vertebrate-associated, such as the picorna-, astro-, calici-, corona-, paramyxo-, hanta-, and arteri-like viruses, we found relatively few new TSA sequences, and the TSA sequences we found comprised <7% (average 2.4%) of the total number of sequences found for each group.

For several virus groups, our analysis revealed large numbers of TSA-derived sequences. For example, 227 (186 TSA and 41 nr/nt) RdRp sequences were classified as best matching the chu-like virus pHMM, from which we obtained 158 nonidentical trimmed sequences for phylogenetic analysis (supplementary fig. S3, Supplementary Material online). Some of the sequences comprise fragments that, in some cases, might derive from the same genome and appear in slightly different places on the tree depending on which region of the RdRp core each fragment covers. A likely full-length RdRp core was present in 85 (45 TSA and 40 nr/nt or ref) of the 158 sequences. There were 14 TSA-derived sequences (from 14 different invertebrate NCBI BioProjects) with contig length >10,000 nt that likely correspond to substantially complete representatives of the nonsegmented form of the chu-like virus genome.

Host Associations of TSA-Derived RdRp Sequences

Next, we looked at which types of cellular life forms the classified TSA-derived sequences putatively infect. For this, we extracted the target organism species name from the metadata accompanying each TSA data set and downloaded the corresponding taxonomic information (family, order, class, phylum, etc.) from NCBI. We then grouped TSA data sets into the following categories: vertebrates (subphylum Vertebrata), arthropods (phylum Arthropoda), invertebrates (kingdom Metazoa/Animalia excluding Arthropoda and Vertebrata), plants (unranked

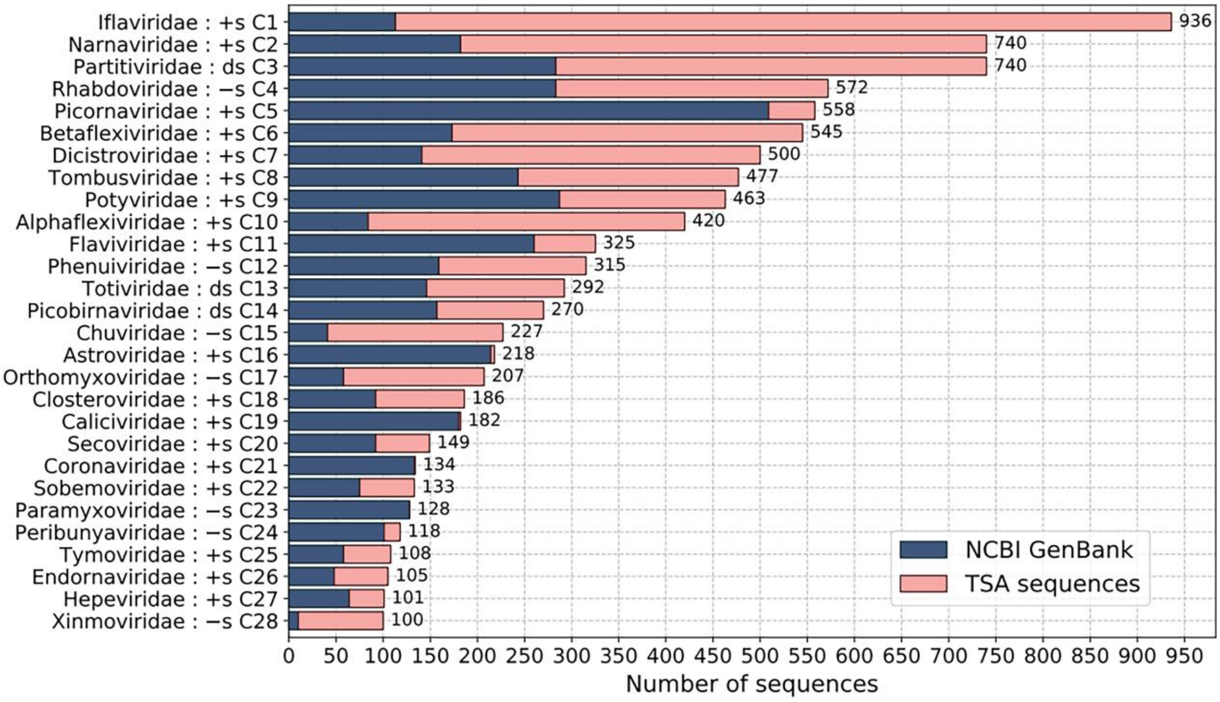

Fig. 2. Total number of classified sequences in each group of classified sequences which had 100 or more sequences (cluster numbers C1–C28; see supplementary fig. S2, Supplementary Material online for all clusters C1–C60). Blue (darker, on the left), nr/nt and ref sequences; pink (lighter, on the right), TSA sequences; +s, +ssRNA viruses; −s, −ssRNA viruses; and ds, dsRNA viruses.

clade Viridiplantae), fungi (kingdom Fungi), protists (domain Eukaryota excluding Metazoa, Fungi, and Viridiplantae), and metagenomic samples (data sets annotated as metagenomic, environmental, or bacterial, though note that the eight bacterial data sets did not produce any matches to our RdRp pHMMs). Next, we calculated the number of different RdRps, number of TSA data sets, and number of distinct host species represented within each host category (fig. 3 and supplementary table S1, Supplementary Material online). We identified relatively few RdRps in vertebrates (1.7% of total number), despite vertebrates representing almost 20% of the TSA data sets. Meanwhile, 48.6% of all identified RdRps were found in arthropods which represent 34.7% of the TSA data sets, equating to ∼2.2 RdRps per data set on average (1,976 RdRps/918 data sets). A similar ratio was observed for plants (1,440 RdRps/650 data sets). Not surprisingly, the highest ratio, ∼6.6 RdRps per data set, was observed for metagenomic samples. This could be due to more fragmented sequences in metagenomic samples and/or the multispecies nature of such samples.

We also checked which individual host species provided the most nonidentical RdRp core sequences (combined over multiple TSA data sets and excluding metagenome data sets). The plants *Saccharum* hybrid (with 78 RdRp sequences), *Tinospora cordifolia* (74), and *Agave tequilana* (53), followed by the insect *Bemisia tabaci* (38), provided the most RdRp sequences. Many other plant and invertebrate species also provided >10 RdRp sequences per species (supplementary fig. S4, Supplementary Material online). Factors contributing to RdRp richness may include the number of independent samples for a host species, sequencing depth, pooling strategy (e.g., pooling RNA from multiple individual organisms into one sample), and likelihood of contamination (e.g., fungi on plant leaves, gut contents in whole-insect samples, etc.). In any case, from these numbers, it is clear that there remains an enormous amount of unsampled RNA virus diversity in plants and arthropods.

We also calculated the number of unique host species for each classified pHMM cluster (fig. 4*A*). In five cases—ifla-, partiti-, narna-, dicistro-, and rhabdo-like viruses—the identified RdRps derived from >100 putative host species. Meanwhile, for Picornaviridae-like viruses, despite this family having numerous NCBI sequences, TSA sequences were found only in a few data sets with the majority being from vertebrate samples. As expected, groups such as ifla-, chu-, xinmo-, phasma-, arto-, and solinvi-like RdRps were largely arthropod associated. This is consistent with the diversity of arthropod viruses observed in previous studies, such as Shi et al. (2016), Batson et al. (2021), and Chang et al. (2021). Groups such as alphaflexi-, poty-, betaflexi-, and tospo-like RdRps were largely plant associated, again as observed previously, for example, by Mifsud et al. (2022).

However, there were also unexpected associations. For example, classical dicistroviruses are an arthropod-associated group, yet ∼45% of the dicistro-like RdRps derived from nonarthropod TSA data sets. Many of these dicistro-like RdRps originated from nonarthropod invertebrate (∼20%) and plant (∼22%) data sets (fig. 4*B*). In some

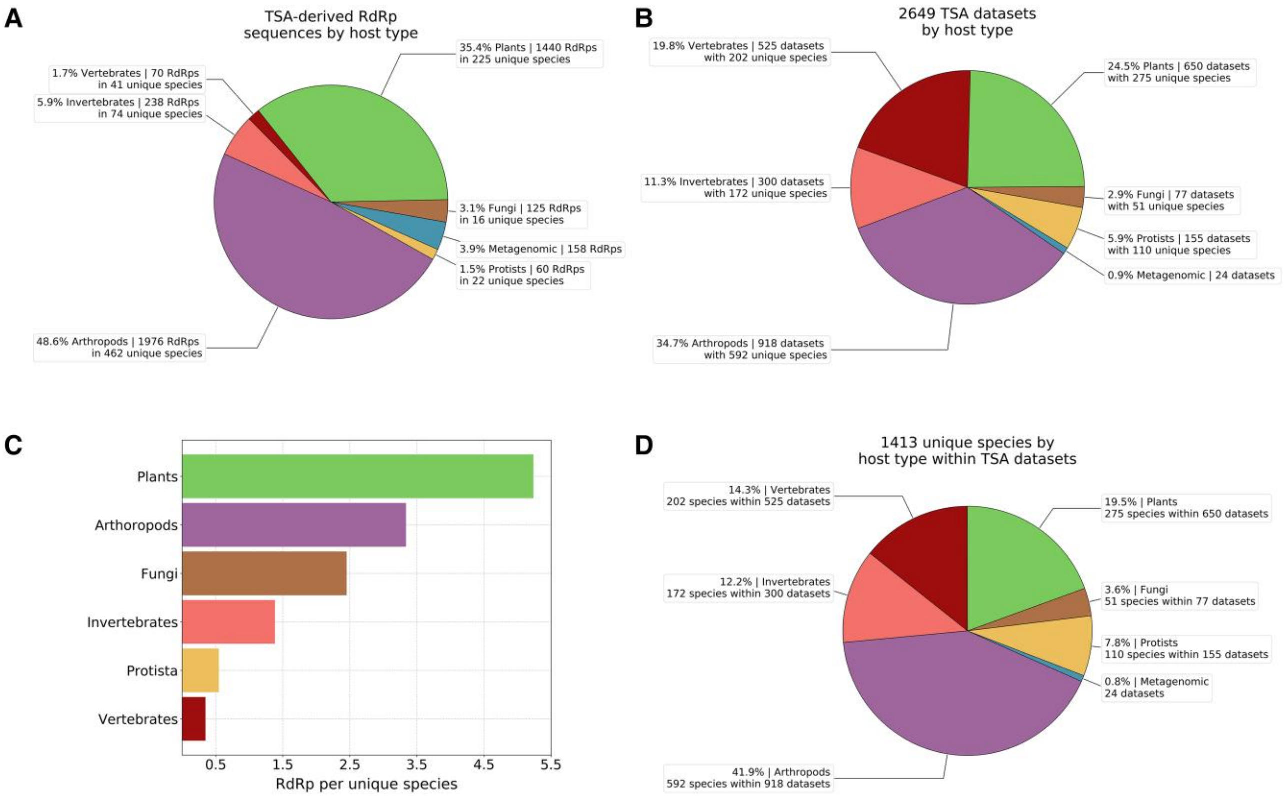

Fig. 3. Distributions of RdRps and host species across TSA data sets. Only nonidentical RdRp core sequences were used (i.e., discarding duplicate 100% identical RdRp sequences within each classified pHMM group, including any identical to nr/nt or ref sequences, leaving the longest representative). No RdRps were detected in the eight bacteria TSA data sets with our pHMMs. (*A*) RdRp counts per host type. (*B*) TSA data set counts per host type. (*C*) Mean number of RdRps per host species. Note that within metagenomics samples, the majority of "species" were named "gut metagenome." (*D*) Numbers of unique TSA data set host species, grouped by host type.

cases, this might result from contamination (e.g., arthropods accidentally sequenced along with plant leaves). Mifsud et al. (2022) quantified contamination in plant TSA data sets and found arthropod contamination in many libraries. Plant-grazing arthropods can also transiently introduce their viruses into plants (Gildow and D'Arcy 1988; Wamonje et al. 2017). In addition, some dicistro-like sequences found in data sets labeled as vertebrate likely derive from arthropod viruses present in vertebrate fecal samples.

Interestingly, the majority (>60%) of classified metagenomic-derived RdRps were placed within the picobirna-like group (fig. 4*B*). It has been suggested that picobirnaviruses may in fact be RNA bacteriophages rather than viruses of eukaryotes (Krishnamurthy and Wang 2018; Wang 2022), which may explain why they are largely absent from our eukaryote-derived TSA data sets but abundant in the metagenomic data sets which typically comprise gut metagenomes. This is also consistent with Chen et al. (2022) where the virome of metagenomic fecal samples was found to contain many picobirna-like viruses.

Evolution of Motif C of the RdRp

RNA virus RdRps contain a number of highly conserved motifs, labeled A–G (or I–VIII) (Koonin 1991; Koonin and Dolja 1993; Bruenn 2003; te Velthuis 2014). Of these, motif C (or motif VI) is the most distinctive and most highly conserved. We decided to leverage our collection of diverse RdRps to more fully understand the extent of variation within the core triplet (typically GDD) of motif C among classified viruses. We took all classified sequences within each pHMM group, aligned them, and manually located the conserved motif C. Consistent with previous work, we saw six possible variations of the core amino acid triplet: GDD, SDD, GDN, IDD, ADN, and ADD (in order of frequency; fig. 5). Interestingly, Forgia, Chiapello, et al. (2022) identified additional variations (NDD, GDQ, and HDD; besides SDD and ADD) in the ormycoviruses —a recently discovered group of fungi-infecting viruses that are highly divergent from other known RNA viruses.

In most +ssRNA virus families, GDD was found to be predominant. As expected, ADD was present in Chipolycipivirus-like sequences (Olendraite et al. 2017) and SDD in members of the order Nidovirales (Koonin and Dolja 1993). Within the group classified to the noda-like pHMM, we noticed four sequences (NC_033077.1, KX883125.1, KX883170.1, and GU976287.1) with SDD instead of GDD. Although highly divergent from each other (46–61% a.a. pairwise identities) these sequences form a distinct monophyletic clade with bootstrap support of 1.0 within the noda-like phylogenetic tree (supplementary

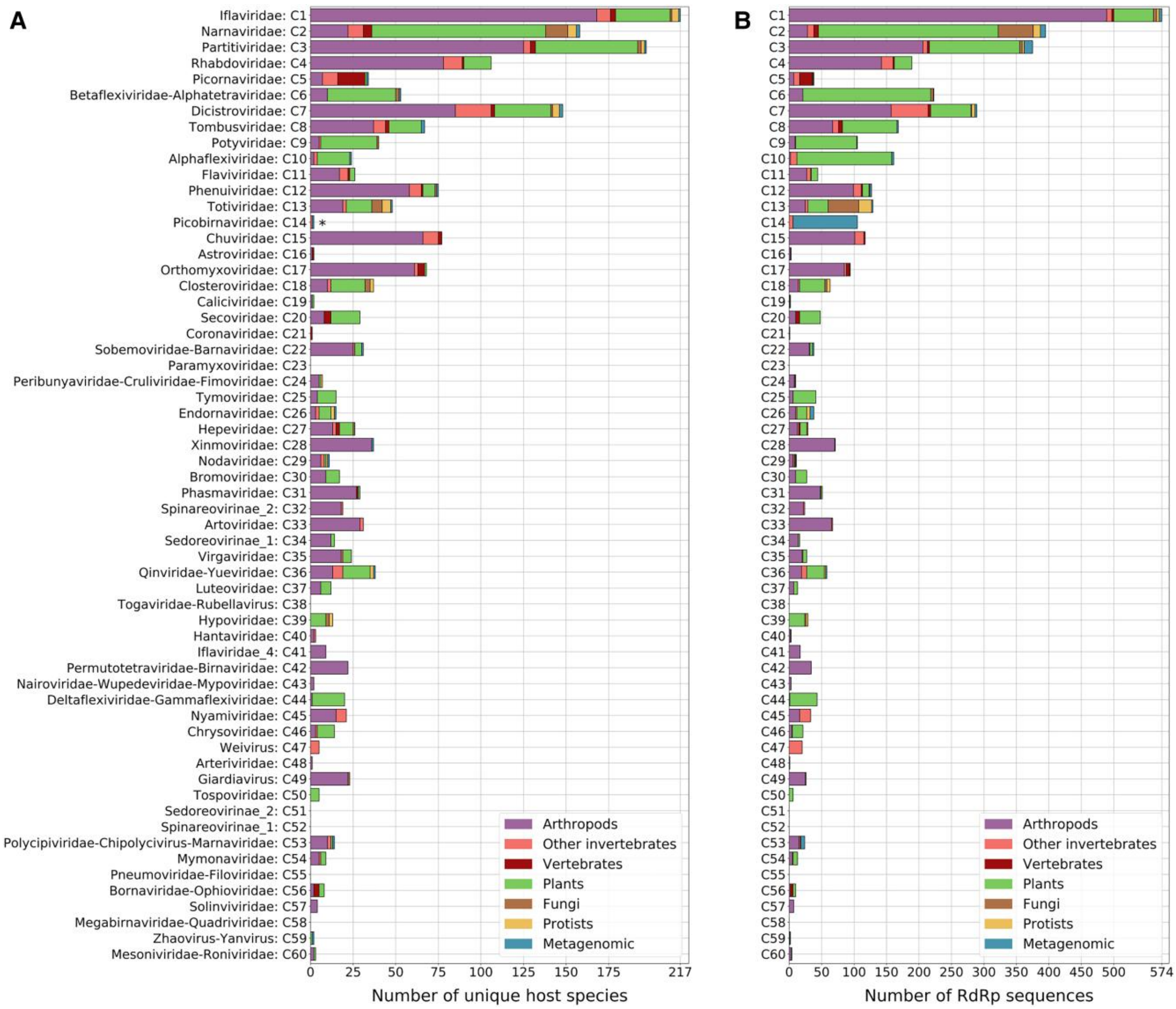

Fig. 4. Numbers of unique putative host species (*A*) and numbers of TSA-derived RdRps (*B*) for different classified pHMM clusters, separated by host species category as indicated in the key. Duplicate 100% identical RdRp core sequences were removed (as in fig. 5). Asterisk (*)—note that the majority of the metagenomic data sets are labeled as "gut metagenome" which is here counted as a single "species" name.

figs. S5*A* and S6*A*, Supplementary Material online). Thus, the SDD is likely to be a true variation and not the result of sequencing errors or sequencing of defective sequences. This is further supported by the fact that changing a glycine codon (GGN) to one of the serine codons (UCC, UCU, UCA) utilized in these four sequences would require changes in both the first and second positions of the codon. We saw a similar phenomenon among permutotetra-like sequences, where a small distinct clade (bootstrap support 1.0) has SDD instead of GDD (supplementary figs. S5*B* and S6*B*, Supplementary Material online). Here, the sequences comprise NC_028381.1, GBSU01004473.1, GBSU01004474.1, and NC_033140.1, where the former three are similar sequences (>90% nt identity) from *Aphis glycines* (soybean aphid), but the latter is more divergent (99% coverage and 83% a.a. identity to NC_028381.1 in the RdRp ORF). In this case, the utilized serine codons are AGC and AGU. In the tree of hypo-like sequences, GDD and SDD are both well represented, and in this case, there are multiple different clades of SDD- or GDD-containing sequences (supplementary figs. S5*C* and S6*C*, Supplementary Material online), indicating multiple switches from GDD to SDD or vice versa during the evolution of hypo-like viruses (although it is possible that this could partly be an artefact of poor sequence alignment and nonrobust phylogenetic inference). These various cases indicate that an ancestral GDD has mutated to SDD on several different occasions in different groups of +ssRNA viruses. Interestingly, we noticed some cases of tombus-like sequences with GDN, which is normally associated with members of the −ssRNA order *Mononegavirales*. Additional tombus-like sequences with GDN have been noted by Gilbert et al. (2019) who proposed a new family, Ambiguiviridae, to contain these viruses. Again, these GDN-containing sequences form a distinct monophyletic clade (bootstrap support 0.97; supplementary figs. S5*D* and S6*D*, Supplementary Material online).

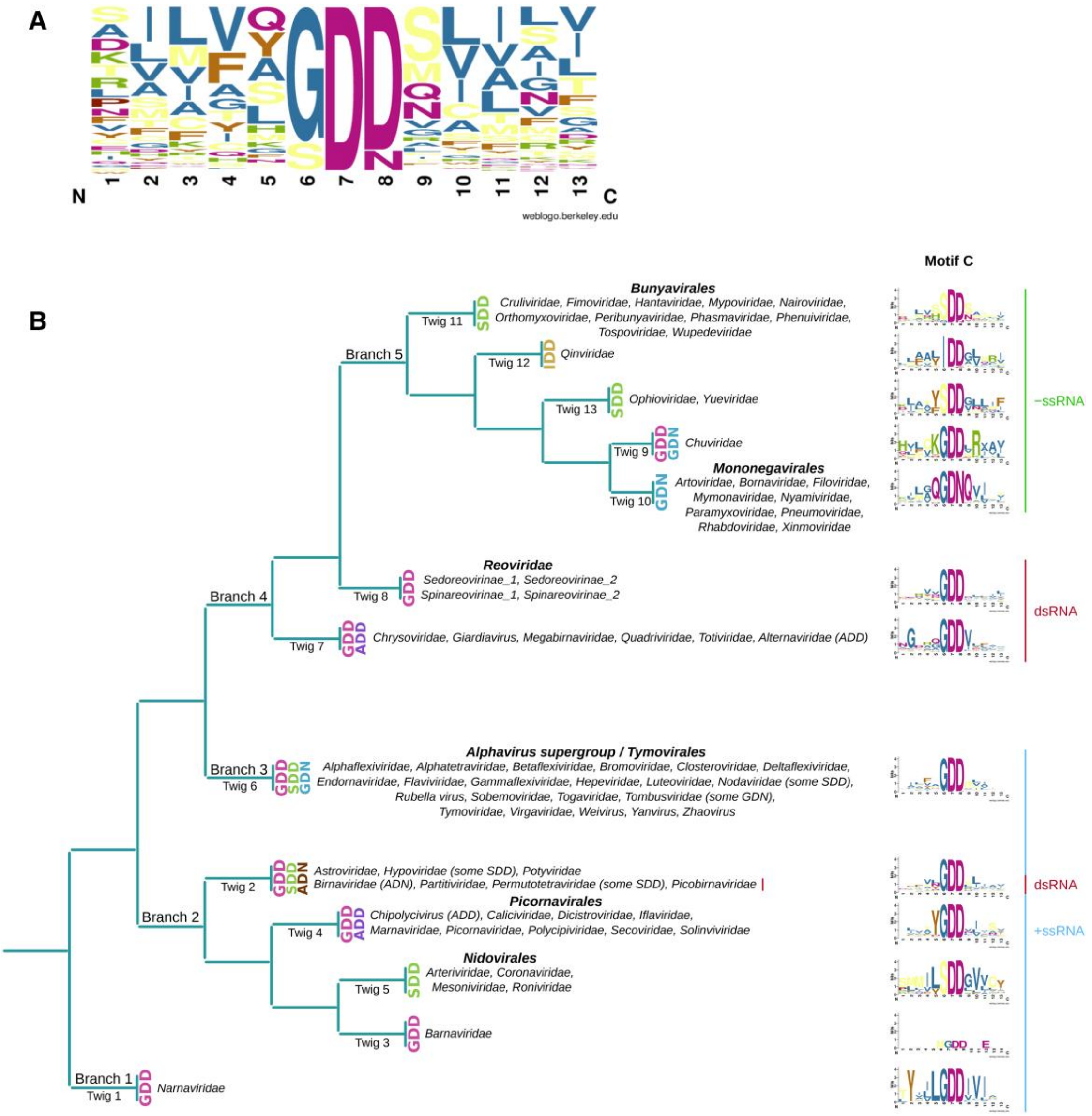
A
N
1 2 3 4 5 6 7 8 9 10 11 12 13
C
weblogo.berkeley.edu
B
Motif C
Bunyavirales
Twig 11
SDD
Cruliviridae, Fimoviridae, Hantaviridae, Mypoviridae, Nairoviridae,
Orthomyxoviridae, Peribunyaviridae, Phasmaviridae, Phenuiviridae,
Tospoviridae, Wupedeviridae
Branch 5
Twig 12
IDD
Qinviridae
Twig 13
SDD
Ophioviridae, Yueviridae
Twig 9
GDD GDN
Chuviridae
Mononegavirales
Twig 10
GDN
Artoviridae, Bornaviridae, Filoviridae,
Mymonaviridae, Nyamiviridae,
Paramyxoviridae, Pneumoviridae,
Rhabdoviridae, Xinmoviridae
−ssRNA
Reoviridae
Twig 8
GDD
Sedoreovirinae_1, Sedoreovirinae_2
Spinareovirinae_1, Spinareovirinae_2
Branch 4
Twig 7
GDD ADD
Chrysoviridae, Giardiavirus, Megabirnaviridae, Quadriviridae, Totiviridae, Alternaviridae (ADD)
dsRNA
Alphavirus supergroup / Tymovirales
Branch 3
Twig 6
GDD SDD GDN
Alphaflexiviridae, Alphatetraviridae, Betaflexiviridae, Bromoviridae, Closteroviridae, Deltaflexiviridae,
Endornaviridae, Flaviviridae, Gammaflexiviridae, Hepeviridae, Luteoviridae, Nodaviridae (some SDD),
Rubella virus, Sobemoviridae, Togaviridae, Tombusviridae (some GDN),
Tymoviridae, Virgaviridae, Weivirus, Yanvirus, Zhaovirus
Twig 2
GDD SDD ADN
Astroviridae, Hypoviridae (some SDD), Potyviridae
Birnaviridae (ADN), Partitiviridae, Permutotetraviridae (some SDD), Picobirnaviridae
dsRNA
Picornavirales
Branch 2
Twig 4
GDD ADD
Chipolycivirus (ADD), Caliciviridae, Dicistroviridae, Iflaviridae,
Marnaviridae, Picornaviridae, Polycipiviridae, Secoviridae, Solinviviridae
+ssRNA
Nidovirales
Twig 5
SDD
Arteriviridae, Coronaviridae,
Mesoniviridae, Roniviridae
Twig 3
GDD
Barnaviridae
Branch 1
Twig 1
GDD
Narnaviridae

FIG. 5. Conservation and diversity in RdRp motif C. (*A*) Sequence logo, produced with WebLogo (Crooks et al. 2004), showing overall amino acid frequencies in the core amino acid triplet and the five flanking amino acids on either side. All nonidentical trimmed RdRp sequences in our study were used (cdhit -c 1.0). (*B*) Schematic representation of motif C central triplet variability overlaid on the inferred evolutionary relationships of RNA viruses from Wolf et al. (2018, 2020).

The dsRNA viruses are the least diverse in motif C and mostly have GDD. However, members of the proposed family Alternaviridae (Gilbert et al. 2019) were found to have ADD. As already established for family Birnaviridae (Gorbalenya et al. 2002), a deviation from GDD to ADN was observed within birna-like sequences. We noticed a few partiti-like sequences with GDE, namely, GFDF01011954.1, GBMJ01010875.1, GBBP01108788.1, GBBP01108783.1, GEFG01022027.1, GEFD01014070.1, and GEEY01016471.1. However, they do not exclusively cluster together. Moreover, they appeared to have fragmented RdRp ORFs and therefore are likely defective sequences, perhaps corresponding to transcribed endogenized viral elements (EVEs) that might have mutated since the original integration into the host genome, leading to a broken ORF and mutation of the aspartic acid codon to a glutamic acid codon which only requires a single-nucleotide change. Notably, partitivirus sequences have been found to be particularly frequently integrated into their host genomes (Chiba et al. 2011). Similarly, a single betaflexi-like sequence, GAGH01076983, with GDE clearly contains a fragmented RdRp ORF and therefore is also presumably defective.

The −ssRNA viruses employ many different motif C variations and GDD is not the most frequently used central triplet. In line with previous knowledge (te Velthuis 2014), members of the order Bunyavirales and the families Orthomyxoviridae, Ophioviridae, and Yueviridae as well as the TSA sequences which best match the pHMMs of these families have SDD, whereas all members of the order

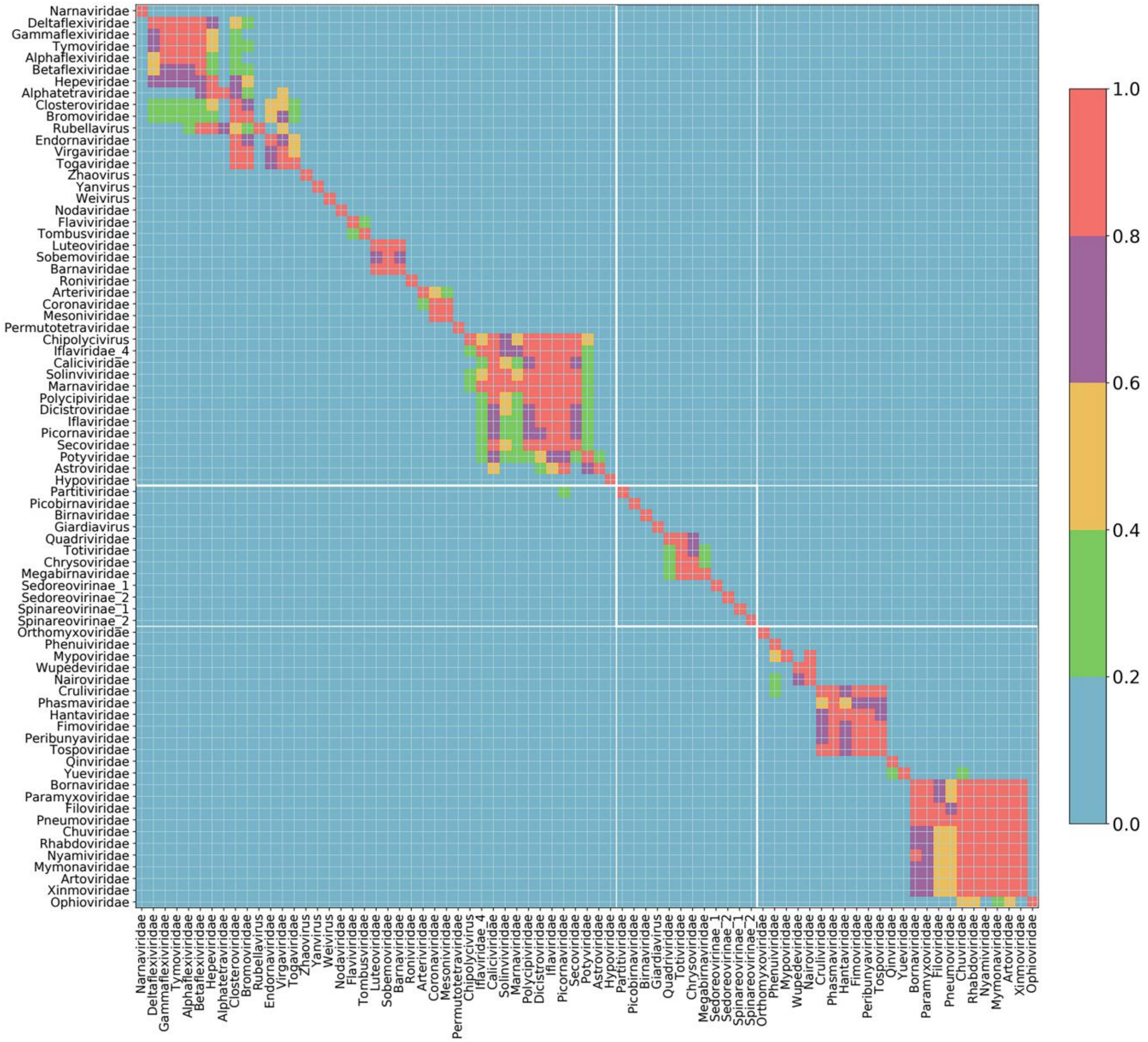

Fig. 6. Heatmap of pHMM match co-occurrences for each RdRp sequence. All classified group ref, nr/nt, and TSA RdRp ORFs were used (supplementary data set 1, Supplementary Material online). For each group on the *y*-axis (best match pHMM), the number of co-occurrences with each group on the *x*-axis (second best match pHMM) was determined, and the count was normalized by the maximum count for the group given on the *y*-axis. Thus 1.0 is the highest co-occurrence score, whereas 0.0 corresponds to pairs of pHMMs that were never matched by the same sequence.

Mononegavirales have GDN. Uniquely, IDD is represented within the sequences that matched the Qinviridae pHMM, as observed previously (e.g. by Charon et al. 2022). Members of the family Chuviridae predominantly have GDD though we also noticed three distinct sequences with GDN. Since these sequences (GDRW01001314.1, GDRW0122005.1, and NC_033704.1) are distinct and come from two unrelated projects, but form a monophyletic clade within the chu-like virus phylogeny (supplementary figs. S3, S5*E*, and S6*E*, Supplementary Material online), the GDN is likely to be a true variation and not the result of sequencing errors or sequencing of defective sequences.

Evolutionary Relationships of the RdRp above the Family Level

RNA viruses are extremely divergent at the level of primary sequence, and relationships above the level of family are often unclear due to difficulties in making robust sequence alignments for phylogenetic analysis. We sought to utilize our large number of RdRp sequences to further investigate higher-level evolutionary relationships. We noticed that each RdRp ORF usually had statistically significant matches to multiple pHMMs ($P < 10^{-6}$). Indeed, some ORFs matched as many as 11 different RdRp profiles. Therefore, we investigated which matched pHMMs tended to occur together, as this might uncover high-level evolutionary relationships. For this analysis, we used the classified group of RdRp sequences from nr/nt, ref, and TSA combined. All pHMM match scores were sorted, and the best and the second best matched pHMM for each ORF were taken to the next step. A heatmap and a network diagram were used to visualize the co-occurrence of different pHMMs (fig. 6 and supplementary fig. S7, Supplementary Material online).

Firstly, it was clear that endorna-like RdRp sequences match the RdRp profiles of +ssRNA rather than dsRNA viruses as their second best match. When we performed this analysis, Endornaviridae was considered to be a dsRNA virus family. However, a phylogenetic association with +ssRNA viruses was observed by Roossinck et al. (2011), and Wolf et al. (2018) inferred that endornaviruses, which in fact are capsidless, had been erroneously labeled as dsRNA viruses. In our analysis, the most closely related families to Endornaviridae were Virgaviridae, Togaviridae, Closteroviridae, and Bromoviridae which were recently grouped into the newly established order Martellivirales. These co-occurrences also connect families within other orders such as the Tymovirales (Alphaflexiviridae, Betaflexiviridae, Deltaflexiviridae, and Gammaflexiviridae) and the Hepelivirales (Hepeviridae and Alphatetraviridae; Koonin and Dolja 1993; Koonin et al. 2015; Wolf et al. 2018) and more loosely link these three orders together, consistent with the 2021 ICTV taxonomy which groups Martellivirales, Tymovirales, and Hepelivirales into the class Alsuviricetes (phylum Kitrinoviricota).

Co-occurrence outside of the same Baltimore group was also apparent for the dsRNA Partitiviridae and the +ssRNA Picornaviridae. For 20–40% of the RdRp ORFs that best matched the Partitiviridae pHMM, the only other match was the Picornaviridae pHMM. This result is consistent with the placement by Wolf et al. (2018) of Partitiviridae together with multiple groups of +ssRNA viruses in Branch 2, making the recently established phylum Pisuviricota. The Picornaviridae pHMM contained many divergent sequences, enabling it to accommodate and tolerate very high diversity. This may explain why partiti-like sequences more easily had second best matches to the Picornaviridae pHMM as opposed to the pHMMs of the other Picornavirales families. The members of the order Picornavirales formed a robustly connected group, with strong links also to the Potyviridae (order Patatavirales) and Astroviridae (order Stellavirales), whereas other members of class Pisoniviricetes clustered elsewhere.

For −ssRNA viruses, families within the class Monjiviricetes formed a very distinctive group, incorporating the Mononegavirales and the Chuviridae (a member of the order Jingchuvirales; Di Paola et al. 2022). The Chuviridae provides a tentative link between this class and two other classes of −ssRNA virus—the Yunchangviricetes (Yueviridae) and Chunqiuviricetes (Qinviridae). In contrast, the Bunyavirales group was split into two clusters. The first cluster comprised Phenuiviridae, Mypoviridae, Wupedeviridae, and Nairoviridae, whereas the second comprised Cruliviridae, Phasmaviridae, Hantaviridae, Fimoviridae, Peribunyaviridae, and Tospoviridae. Interestingly, sequences classified to the Ophioviridae pHMM sometimes had secondary matches to pHMMs of families within the Mononegavirales order, but not vice versa. For dsRNA viruses, there was a clear clustering of Quadriviridae, Totiviridae, Chrysoviridae, and Megabirnaviridae that form the recently established order Ghabrivirales. Interestingly, and again consistent with the phylogeny of Wolf et al. (2018) and the classification of Aiewsakun and Simmonds (2018), the Giardiavirus pHMM did not form part of this cluster even though the genus Giardiavirus is currently classified in the Totiviridae family. Another order-like clustering was observed for the Luteoviridae, Sobemoviridae, and Barnaviridae pHMMs (note that our Sobemoviridae pHMM corresponds to what is now designated genus *Sobemovirus* in family Solemoviridae along with *Polemovirus*, *Polerovirus*, and *Enamovirus* and our Luteoviridae pHMM—following the Aiewsakun and Simmonds groupings—contains only *Enamovirus* sequences).

Sensitive Detection of Divergent RdRps Using pHMMs

To search for novel and unusual sequences, the most interesting group of putative RdRps is the unclassified group. We placed 361 sequences within the unclassified group, of which 142 are TSA-derived (supplementary fig. S8, Supplementary Material online). First, we investigated the lengths of these contigs, their full RdRp-encoding ORFs, and the RdRp core (full ORF trimmed to the best pHMM match positions; supplementary fig. S9, Supplementary Material online). The RdRp-encoding contigs varied from 700 nt to 35,913 nt, with 51 sequences longer than 10,000 nt. The full RdRp-encoding ORFs varied from 240 to 8,398 codons, with 185 out of 361 longer than 1,000 codons. After trimming these ORFs to the pHMM match positions, the sequences were 300–400 a.a. in length, as expected for the core of an RdRp. The unclassified group contains particularly divergent viruses, with a mean of 26.1% identity to the most similar reference sequence, compared with 64.6% for the classified sequences (measured with BLASTP against sequences included in the input pHMMs, excluding self matches for the GenBank sequences; supplementary fig. S10, Supplementary Material online).

We also checked which classified family pHMM each of the 361 unclassified sequences best matched. By our definition of unclassified sequences, in these cases, the match score is very low though still statistically significant. We found that unclassified sequences best matched only 43 of the 77 pHMMs (supplementary fig. S11, Supplementary Material online). It is important to note however that some pHMMs were created using only very similar input sequences and therefore are less able than other pHMMs to match related but divergent sequences. The highest proportion of unclassified sequences matched the Hepeviridae pHMM. There were 32 such sequences and

18 of them were TSA derived. For the Picornaviridae pHMM, there were 18 sequences and 11 of them were TSA derived. In the case of the Qinviridae pHMM, all 11 matched unclassified sequences were from the TSA database.

Within the unclassified sequences, we found some clades—such as family Arenaviridae and genus *Sinaivirus*—that correspond to known virus taxa for which for various reasons we had not included pHMMs in our analysis. These provided a useful control, demonstrating the ability of our pipeline to find new family-level groups or divergent singletons which sometimes showed very remote similarity to existing taxonomic groups. There was also a group of five very long *Nidovirales* sequences—corresponding to a group of viruses with the largest known RNA genomes to date—which were published (Debat 2018; Saberi et al. 2018) but not yet classified when we performed our analysis. Sometimes, newly identified unclassified group sequences formed new sister clades to the clade of their best-match pHMM. Although there were multiple such instances, below we describe examples among toti- and giardia-like viruses. We also highlight a divergent mononegavirus with apparently splicing-dependent RdRp expression, besides new clades of orthomyxo-like viruses.

An Expansion of the Toti-Like Viruses

There were 22 unclassified sequences which had the best match to the Totiviridae pHMM. Among these sequences, 6 were TSA-derived and the shortest one was 2,697 nt in length. Based on genome organization, they appeared to resemble typical Totiviridae viruses with a capsid-encoding ORF followed by an RdRp-encoding ORF. To see if these unclassified sequences form a separate clade, and how phylogenetically different they are from sequences with a strong match to the Totiviridae pHMM, we compared the 22 sequences with classified sequences. There were 292 sequences classified to the Totiviridae pHMM; therefore, we discarded sequences which were shorter than 400 a.a. in length for the RdRp core region and that shared >70% similarity (CDHIT -c 0.7) to a longer sequence, leaving 80 sequences. The 22 unclassified and remaining 80 classified sequences were used to produce a PhyML phylogenetic tree (supplementary fig. S11, Supplementary Material online). Four main clades were observed, two of which contained only classified sequences and covered all ref and nr/nt sequences of the respective genera: *Totivirus* (clade 1) and *Victorivirus*, *Leishmaniavirus*, and *Trichomonavirus* (clade 2).

A third clade contained 16 ref or nr/nt and 5 TSA sequences, and, of these, 4 were classified sequences, and the remainder were unclassified sequences. A strongly supported subclade within clade 3 contained 10 sequences of 8.1–9.4 kb in length, substantially longer than most toti-like sequences. All ten sequences were found in fungal hosts—nine in Ascomycota and one in Basidiomycota. These sequences correspond to a previously proposed new family, Fusagraviridae (Wang et al. 2016; Lee et al. 2017; Arjona-Lopez et al. 2018). This family is not yet recognized by the ICTV but is supported by our analysis as a monophyletic group distinct from the Totiviridae. Another subclade within clade 3 contains *Ustilago maydis* virus H1, currently classified in ICTV as part of the *Totivirus* genus. All the other members of genus *Totivirus* fall firmly within clade 1, which suggests this species should be reclassified. In our phylogeny, *U. maydis* virus H1 clusters with a member of the *Botybirnavirus* genus (which is currently not classified above genus level), besides diatom colony–associated virus 17 types A and B (both of which have not been taxonomically classified beyond a provisional link to the Totiviridae). Our phylogeny would support the addition of the *Botybirnavirus* genus to the *Ghabrivirales*, the order which contains family Totiviridae, but not to the Totiviridae family itself. By incorporating the results of other recent studies, it is apparent that the subclade of clade 3 containing our three TSA sequences (from the stone fly *Perla marginata*, the red alga *Kappaphycus alvarezii*, and the orchid *Chiloglottis trapeziformis*) corresponds to the totivirus-like clade identified by Kartali et al. (2019) as containing their Umbelopsis ramanniana virus 3. This clade also incorporates the clade shown by Charon et al. (2021) as containing their Chrysaor toti-like virus and Laestrygon toti-like virus and the clade shown in Shi et al. (2016) containing diatom colony–associated dsRNA virus 17, Beihai barnacle virus 15, Hubei toti-like virus 5, Beihai sesarmid crab virus 7, and Beihai razor shell virus 4. Mifsud et al. (2022) identified Elkhorn sea moss toti-like virus—likely to be the same virus as our red alga TSA sequence—and a number of other plant-associated viruses which also fall into this clade. Thus, this grouping brings together our three novel sequences with a number of previously scattered known sequences.

The fourth clade contains five sequences from arthropods, all with an unclassified status in our analysis—two TSA sequences (GFQL01013152.1 from the moth *Carposina sasakii*, GBNZ01013113.1 from a *Heterodontonyx* sp. wasp) and three ref or nt/nt sequences (NC_007915.3, penaeid shrimp infectious myonecrosis virus; NC_033467.1, Wuhan insect virus 31; and NC_032948.1, Hubei toti-like virus 18).

An Expansion of the Giardia-Like Viruses

In addition to the Totiviridae pHMM, we also had a separate profile for the *Giardia lamblia* virus RdRp. In total, there were 65 sequences (nr/nt, ref, and TSA) which had the best match to this pHMM, with lengths ranging from 308 to 12,427 nt (the longest being LC333746.2, *Rosellinia necatrix* megatotivirus 1; Arjona-Lopez et al. 2018). Among our identified sequences, there were 24 with lengths >6,000 nt (cf. the *G. lamblia* virus reference sequence NC_003555.1 has 6,277 nt). As the *Giardiavirus* RdRp is translated via ribosomal frameshifting, for all these 65 sequences, we defined ORFs between stop codons to ensure coverage of the entire RdRp. For each contig, the ORF with longest BLASTP match to the *G. lamblia* virus

RdRp a.a. sequence was selected. The translated ORFs were aligned, and a phylogenetic tree was generated with PhyML (supplementary fig. S12, Supplementary Material online).

The original *G. lamblia* virus sequence formed a clade together with seven unclassified nr/nt and ref sequences and one TSA sequence (GBYF01047348.1), which was identical to a known nr/nt sequence (MG256177, *Gigaspora margarita* giardia-like virus 1). A neighboring clade to this one encompassed 5 nr/nt and ref sequences together with 13 TSA sequences—1 divergent sequence from a microfungus TSA sample (*Rhizopus oryzae*, GDUK01008098.1) and 12 sequences from 5 crustacean TSA samples (swimming crab, great spider crab, signal crayfish, and two species of *Proassellus* isopods). These 12 TSA sequences together with Wenzhou crab virus 5 form a strong clade (bootstrap support 1.0) of crustacean-associated giardia-like viruses. The remaining sequences comprised 30 TSA sequences from many different hosts which loosely clustered together with 7 GenBank unclassified viruses.

Amphibian-Associated Orthomyxo-Like Viruses

Within the Orthomyxoviridae family, the four influenzavirus genera form a highly supported monophyletic clade of vertebrate-infecting viruses (fig. 7). In our study, we identified additional influenza-like sequences in three different amphibian species. In the TSA data set GFMT01 (cane toad, *Rhinella marina*), we found a full PB1-encoding sequence (GFMT01051794.1, 2,350 nt), and in the TSA data set GECV01 (ornate chorus frog, *Microhyla fissipes*), we found three PB1-matching contigs (GECV01084760.1, 713 nt; GECV01039268.1, 338 nt; and GECV01050644.1, 473 nt). Using amino acid sequences derived from these PB1 contigs, we searched online amphibian TSA data sets (TBLASTN) and identified another PB1-encoding contig, namely, JV207023.1 (1,419 nt) in an *Ambystoma mexicanum* (axolotl) TSA data set (NCBI BioProject PRJNA157225). Comparison of the 7,332 contigs in BioProject PRJNA157225 with Orthomyxoviridae NCBI reference proteins using BLASTX revealed additional influenzavirus-like contigs, including JV207023.1, JV205532.1, and JV206720.1, that matched PB1 and could be merged into a 2018 nt sequence. These amphibian-associated sequences—which were also found concurrently by Parry et al. (2020)—all cluster within the influenzavirus clade (fig. 7).

Surprisingly, we also identified influenza-like PB1-encoding sequences in two fish TSA data sets (*Salaria pavo*, NCBI BioProject PRJNA329073, and *Nibea albiflora*, NCBI BioProject PRJNA359138). We identified sequences for the other segments by searching the respective TSA data sets using TBLASTN with Orthomyxoviridae NCBI reference proteins as queries. This revealed 9 and 15 influenza A virus–like contigs for *S. pavo* and *N. albiflora*, respectively. Using BLASTX, we identified all eight virus segments and found the encoded proteins to have 93–100% a.a. identity to influenza A virus proteins (supplementary tables S2 and S3, Supplementary Material online). These identity levels are typical for different strains of influenza A virus (supplementary table S4, Supplementary Material online) and much higher than identity levels between, for example, the homologous proteins of influenza A and B viruses (supplementary table S5, Supplementary Material online). Thus, the fish TSA data set sequences may be considered to represent the influenza A virus species and, since fish are not known hosts of influenza A virus, are likely contaminants (a conclusion also reached by Parry et al. 2020)—perhaps from bird fecal material or laboratory contamination. Similarly to Mifsud et al. (2022), who found influenza A virus in 16 plant transcriptomic data sets, we also found influenza A virus sequences in a plant data set (*Thlaspi arvense*—NCBI BioProject PRJNA183631), with the PB1 fragment GAKE01008984.1 having 99% nt identity to the avian influenza A virus sequence CY149610.

Among the four unclassified sequences which had the best match to the Orthomyxoviridae PB1 pHMM (supplementary fig. S8, Supplementary Material online), there was a 2,551 nt sequence, GFBM010604515.1, from another *A. mexicanum* data set. In contrast to the influenzavirus-like amphibian-associated sequences mentioned above and in Shi et al. (2018) and Parry et al. (2020), this sequence was far more divergent (e.g., 22.7% a.a. identity, 60% coverage in a TBLASTN comparison with influenza B virus PB1) and did not cluster within established genera (fig. 7). Note that the tree of orthomyxo-like viruses (fig. 7) does not extend to other Articulavirales clades such as the Amnoonviridae (Turnbull et al. 2020) and the unclassified gecko-derived Lauta virus (Ortiz-Baez et al. 2020), and the PB1 encoded by GFBM010604515.1 is more closely related to Orthomyxoviridae PB1 proteins than the PB1 proteins of these other virus groups.

Orthomyxoviruses have segmented genomes, with typically 6–8 segments. In an attempt to find other segments of this novel virus, we used TBLASTN to query 164 NCBI Orthomyxoviridae protein reference sequences (covering all segments) against the *A. mexicanum* TSA data set (NCBI BioProject PRJNA300706). However, the only match was to the original contig, GFBM010604515.1. To increase sensitivity, we downloaded all TSA contigs ($\sim 1.5 \times 10^6$) from BioProject PRJNA300706 and, using HMMsearch, compared them with new pHMMs generated for Orthomyxoviridae reference proteins. In this manner, we identified contigs encoding the three replicase components: GFBM010604515.1 (2,551 nt, PB1), GFBM010554880.1 (2,714 nt, PB2), and GFBM010538345.1 (2,195 nt, PA). All three contigs contain the ORF stop codon. Furthermore (after reverse complementing where appropriate), all three contigs have an identical 5′-end AAAAAGCAGU sequence (plus 0–2 extra 5′-terminal nucleotides) consistent with the conserved segment ends expected for orthomyxovirus sequences. Thus, the encoded proteins PB1, PB2, and PA appear to be full length.

We applied BLASTP to the three retrieved protein sequences, querying against the NCBI nr protein database.

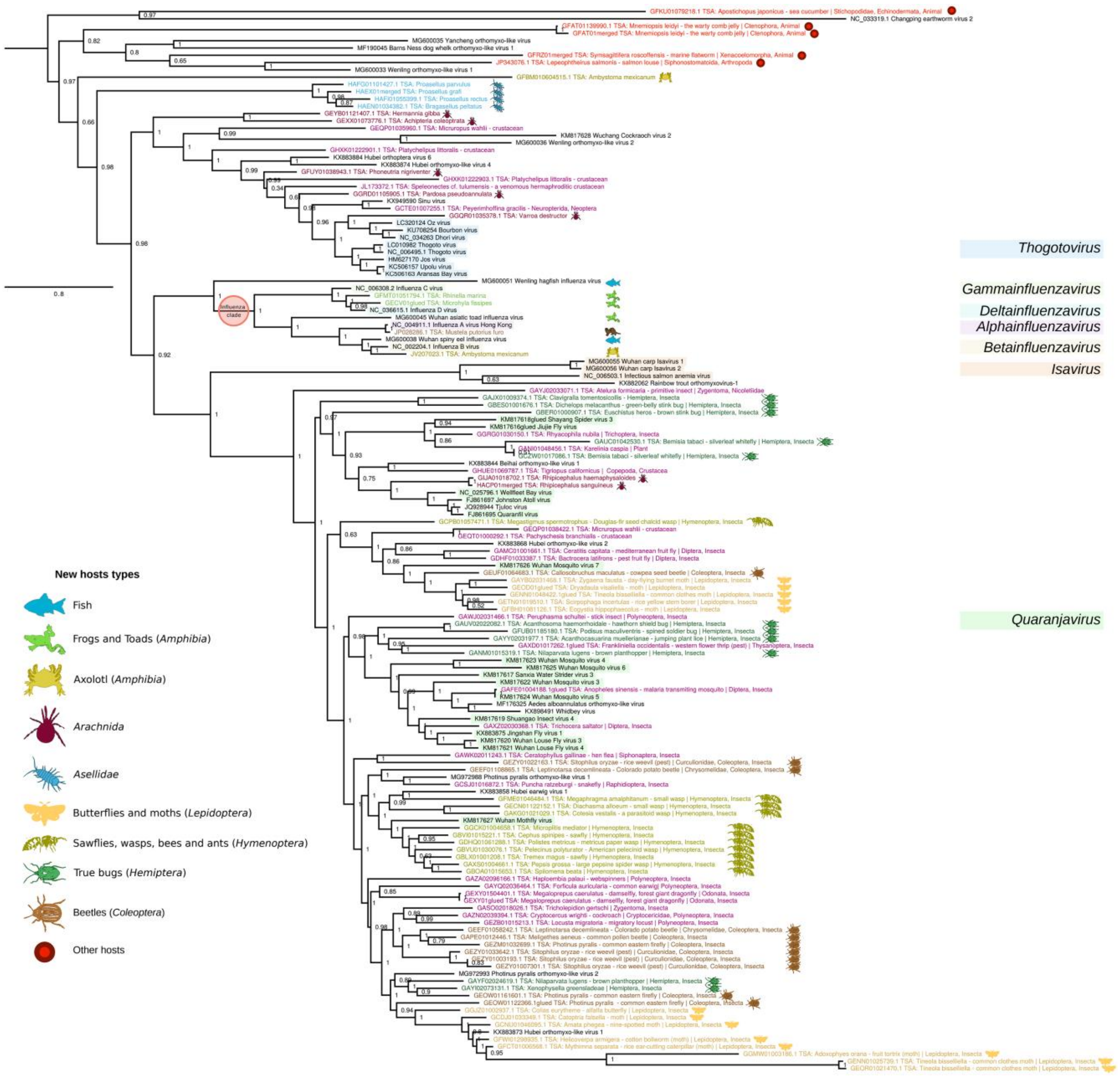

Fig. 7. Phylogenetic tree of sequences (classified or unclassified) with best match to the orthomyxovirus-like pHMM. Sequences shorter than 100 a.a. were removed, and then, sequences with >95% identity were clustered, and only the longest sequence in each cluster was retained as a representative. NCBI accession numbers, virus names, TSA target organism names, and group-representative icons/colors are shown (key at left). Currently defined genera are identified with colored highlighting (key at right). Sequences labeled as "merged" derive from merged overlapping contigs. Sequences labeled as "glued" comprise multiple concatenated ORFs from a contig that was inferred to likely have sequence quality issues which introduce stop codons (e.g., via frameshift errors—common with 454 sequencing) or potentially derive from mutated endogenized viral elements, EVEs). Ferret icon has been added but it is a well-known host type.

The best match for PB1 was Sinu virus (bitscore 72.4, 91% cover, 20% identity, $E = 6 \times 10^{-9}$), followed by Neke Harbour virus (bitscore 71.2, 33% cover, 25% identity, $E = 10^{-8}$), Wuhan mosquito virus 4 (bitscore 69.3, 41% cover, 24% identity, $E = 5 \times 10^{-8}$), and many influenza B virus sequences. The best matches for PB2 were all influenza A virus sequences (top match bitscore 60.8, 43% cover, 21% identity, $E = 2 \times 10^{-5}$). Finally, the PA protein matched only Barns Ness dog whelk orthomyxo-like virus 1 (bitscore 58.2, 42% cover, 24% identity, $E = 10^{-4}$). Thus, this novel amphibian-associated virus appears more closely related to orthomyxoviruses than to other virus families, though it falls outside of currently defined Orthomyxoviridae genera. It is possible that the host species of this virus is not in fact the axolotl but might instead be a contaminant (e.g., an invertebrate).

A New Clade of Orthomyxo-Like Viruses Associated With Isopods

In addition to the amphibian-associated sequences noted above, within the tree of all sequences with the best match to the Orthomyxoviridae pHMM, we also saw some new clades that fall outside of currently recognized genera (e.g., the top 14 sequences in fig. 7). One of these clades was specific to TSA data sets sampling the family Asellidae—a group of isopod crustaceans. A second clade contained a mixture of previously published virus sequences together with sequences from marine flatworm (*Symsagittifera roscoffensis*), salmon louse (*Lepeophtheirus salmonis*), and warty comb jelly (*Mnemiopsis leidyi*) TSA data sets. The *L. salmonis* sequence was previously identified by Waldron et al. (2018) and clusters with their Barns Ness dog whelk orthomyxo-like virus, whereas the other two sequences appear to be novel. A third clade contained the previously published Changping earthworm virus 2 and a sequence from a sea cucumber (*Apostichopus japonicus*) TSA data set.

For the Asellidae-associated clade, we initially identified nine contigs, but, after merging two overlapping contigs, there were eight PB1-encoding sequences from six different TSA data sets representing six different Asellidae species within NCBI BioProject PRJEB14193 (supplementary table S6, Supplementary Material online). The eight contigs have 65–98% nucleotide identity to each other (supplementary fig. S13A, Supplementary Material online). Using BLASTX, we compared these sequences with Orthomyxoviridae NCBI reference PB1 proteins and found them to be highly divergent, with all a.a. identity levels <32% (supplementary table S7 and fig. S13B, Supplementary Material online). When we compared the Asellidae-associated PB1 sequences with the entire NCBI nonredundant (nr) protein database using BLASTX, aside from orthomyxovirus-like sequences, there were no other significant matches, thus confirming a closer relationship to the Orthomyxoviridae family than to any other viruses (including other viruses in the Articulavirales order). For the phylogenetic tree, we discarded shorter sequences that had >95% amino acid identity to longer sequences, leaving the four PB1 Asellidae-associated sequences shown in figure 7. The four sequences are all >1,000 nt, and two are >2,000 nt, making them possibly full-length coding sequences (supplementary table S6, Supplementary Material online).

Using TBLASTN, we queried all Orthomyxoviridae NCBI reference proteins against the Asellidae BioProject PRJEB14193 to search for contigs matching to orthomyxovirus proteins other than PB1. Four contigs showed significant matches to the PA protein of thogotoviruses. When these contigs were queried against the Asellidae BioProject using TBLASTX, a total of 11 contigs from 5 host species were found that separated into 3 groups (97–100% a.a. identity within a group) plus a fourth small fragment. The longest segments by group were HAEN01028927.1 (2,008 nt), HAFG01097557.1 (1,986 nt), and HAEX01036584.1 (1,799 nt). In comparisons with thogotovirus PA (YP_145795), these sequences had coverage values of 32–51% and amino acid identities of 21.7–24.0%. Compared with each other, coverage and amino acid identities were in the range of 86–97% and 51.2–59.6%, respectively.

The other divergent orthomyxovirus-like clades and sequences mentioned above comprise the *A. mexicanum* TSA sequence GFBM010604515.1 discussed above, five other TSA sequences from a variety of organisms, and four previously identified but not currently classified NCBI nr/nt viruses (Shi et al. 2018; Waldron et al. 2018; upper part of tree in fig. 7). When these sequences were compared with the entire nr protein database using BLASTX, most of the best hits were encoded by the other nr/nt sequences identified in this group, with identities (variable coverage) mostly in the range 21–34% (supplementary table S8, Supplementary Material online). There were no hits to viruses other than orthomyxo-like viruses. Among the Orthomyxoviridae, there were unclassified Orthomyxoviridae, *Quaranjavirus*, and *Isavirus* PB1 matches (identities 20–30%, coverage 15–50%, bit scores 41.2–70.9, $E = 0.038\text{–}3 \times 10^{-11}$). Since these sequences do not cluster according to host taxa, and pairwise identity scores are relatively low, it is perhaps premature to propose new taxa at this stage. Nonetheless, these sequences enrich the apparent host range of orthomyxo-like viruses to include (with the caveat of potential contamination) such hosts as the marine flatworm (phylum Xenacoelomorpha), warty comb jelly (phylum Ctenophora), and sea cucumber (phylum Echinodermata), in addition to the published orthomyxo-like viruses from whelk (phylum Mollusca), earthworm (phylum Annelida), and the well-established Orthomyxoviridae host taxa Arthropoda and Vertebrata.

A Divergent Mononegavirus with Splicing-Dependent RdRp Expression

While looking at genome graphs of unclassified sequences, we noticed an unusual and divergent TSA sequence, GEZL01043288.1 (from a common ragweed data set, *Ambrosia artemisiifolia*). By analyzing the original TSA data set, GEZL01, we were able to extend GEZL010 43288.1 at the 3′ end with contigs GEZL01043287.1, GEZL01043289.1, and GEZL01043290.1. The resulting 12,569 nt sequence had best TBLASTX hit to Wuchan romanomeris nematode virus 2 (NCBI nr/nt; KX884441.1; Nematovirus, Lispiviridae, and Mononegavirales). This match mapped around RdRp motif C (QGDNQ) where there was 38% identity over 187 a.a. of the RdRp. The original TSA sequence best matched the Artoviridae pHMM and, when extended, showed features typical of viruses in the order Mononegavirales (fig. 8A). Phylogenetically, the sequence falls in a sister clade to the Rhabdoviridae-like family Lispiviridae (fig. 8B and supplementary fig. S13, Supplementary Material online). ORF1 apparently encodes the nucleoprotein (HHpred $E = 3.6 \times 10^{-9}$, PDB_mmCIF70_14Oct:1N93_X), whereas we were not able to identify the putative ORF2 and ORF3 products by homology search with HHpred. Downstream of ORF3, the 3′ region contains several disjoint ORFs, three of which have highly significant a.a. matches to the Mononegavirales L protein (TBLASTN against KX884441.1, $E = 5 \times 10^{-13}$ for the shortest fragment) and together cover its RdRp, capping, connector, and methyltransferase domains (HHpred of concatenated a.a. sequences has an $E = 5.1 \times 10^{-189}$ hit to PDB_mmCIF70_14Oct:6V85_A; fig. 8A). Further inspection revealed the presence of three introns (see below; fig. 8A and supplementary fig. S15, Supplementary Material online). When these introns are not spliced, the RdRp core is split between disjoint ORFs, whereas removal of all three introns fuses the RdRp/L protein-coding region into a single long ORF.

Evidence for the introns initially came from a comparison of the TSA contigs GEZL01043287.1 and GEZL01043290.1 which are identical except that GEZL01043290.1 has a single 98 nt deletion (intron 2) with canonical GU/AG exon/intron boundaries and flanking sequences that closely match the splice site consensus sequence of *Arabidopsis thaliana* (supplementary fig. S15, Supplementary Material online; Brown et al. 1996); the human splice site consensus sequence is very similar. Comparison of the TSA contigs GEZL01043288.1 and GEZL01043289.1 revealed a 132 nt intron (intron 1) that is deleted in GEZL01043289.1, whereas we found the 131 nt intron 3 manually. Like intron 2, introns 1 and 3 have canonical GU/AG boundaries and favorable (though less extensive) flanking intron–exon junction sequences (supplementary fig. S15, Supplementary Material online). All three introns are AU-rich (68.2%, 71.4%, and 68.7%, respectively) compared with a mean AU fraction for the entire 12,569 nt sequence of 56.5%.

The antigenomes of viruses in the −ssRNA virus order Mononegavirales typically contain a number of consecutive coding ORFs separated by intergenic regions. In the negative-sense template, these intergenic regions contain transcription stop–start signals (reviewed in Ogino and Green 2019) which direct the production of a series of positive-sense polyadenylated mRNAs, one mRNA species for each main ORF. To better define the RdRp ORF initiation site, we next identified the transcription stop–start motif. For this, we aligned intragenic regions and used GLAM2 (Frith et al. 2008) to identify enriched motifs. The motif comprising the stop–start signal ACU(U/A)UAAAAAGUGAAA(G/A)(G/A)C (represented in the positive sense; fig. 8C) was found ending at nucleotide positions 1,500, 2,595, and 4,758. The first two signals could produce transcripts for ORF2 (1,503–2,177) and ORF3 (2,684–4,540), respectively. The third signal could produce a transcript for the L protein ORF if all three introns are removed (AUG initiation codon at nt 4,771–4,773; see supplementary fig. S16, Supplementary Material online for protein sequence). In the absence of other transcription stop–start signals, initiation at this AUG codon and utilization of all three splice sites appears to be the only way for the virus to express a functional RdRp/L protein. We used TBLASTN to compare the entire 2,338 a.a. predicted L protein sequence against the related TSA sequences (identified by TBLASTN)—GDQF01122294.1 (*Erigeron breviscapus* TSA) and GGQG01009943.1 (*Cichorium intybus* TSA) (fig. 8B). Both sequences display 99% coverage of the predicted L protein and, importantly, a lack of alignment gaps at the identified exon–exon junctions in the GEZL01-derived sequence supports introns 1–3 being functionally utilized (supplementary fig. S17, Supplementary Material online).

Next, we wanted to check the abundance of viral reads in the data sets and also identify reads which would cover or span spliced regions of the sequence. For this purpose, we used HISAT2 (Kim et al. 2019) to map the raw reads from the corresponding BioProject (NCBI accession PRJNA335689, three different data sets) to spliced and unspliced versions of the viral sequence. The three data sets represent three different *A. artemisiifolia* (common ragweed) plant tissues (female flowers, male flowers, leaves) (Virág et al. 2016), and there was a clear difference in coverage between the three data sets (supplementary fig. S18*A*, Supplementary Material online). Interestingly, much higher levels of virus were found in the male flower sample, whereas the lowest levels were present in the leaf sample. As expected, we saw dips in mapped read counts corresponding to the transcription stop–start sites (supplementary fig. S18*A*, Supplementary Material online). We also saw dips corresponding to the exon–exon junctions, indicating that a substantial proportion of the RNA density in the L ORF region comes from RNA containing the intron sequences. This could be from prespliced L mRNAs, genomic vRNA, antigenomic cRNA, or transcripts from which the introns are never spliced. The latter could provide a mechanism to reduce or even temporally regulate L protein expression. Next, we plotted only spliced reads as determined by HISAT2 over the L protein–encoding region (supplementary fig. S18*B*, Supplementary Material online). All three introns mentioned above were supported by multiple spliced reads. In addition, HISAT2 uncovered an additional intron (intron 4) besides alternative 5′ donor sites for introns 1 and 4 (supplementary fig. S15, Supplementary Material online). However, utilization of alternative introns 1 or 4 would disrupt the L ORF leading to a greatly truncated L protein, whereas excision of intron 4 would lead to a 70 a.a. deletion compared with the *E. breviscapus* and *C. intybus* sequences (see above; supplementary fig. S17, Supplementary Material online).

The presence of these three related sequences in *A. artemisiifolia*, *E. breviscapus*, and *C. intybus* (three members of the Asteraceae family of flowering plants) besides related virus-derived sequences in more recent plant TSA data sets such as *Cenchrus americanus* (GEUY01006481.1), *Gymnadenia rhellicani* (GHXH01342866.1, GHXH01342865.1, GHXH01230927.1), *Ophrys sphegodes* (GHXJ01055800.1), and *Ophrys fusca* (GHXI01221739.1) suggests plants are the bona fide hosts of these viruses.

Splicing is a very rare phenomenon in RNA viruses (excluding retroviruses) and is known only in a few cases—notably bornaviruses (Cubitt et al. 1994; Schneider et al. 1994; Tomonaga et al. 2000; Pfaff and Rubbenstroth 2021) and various orthomyxoviruses (Inglis and Brown 1981; Shih et al. 1998; Kochs et al. 2000; Wise et al. 2012) and Culex tritaeniorhynchus rhabdovirus (family

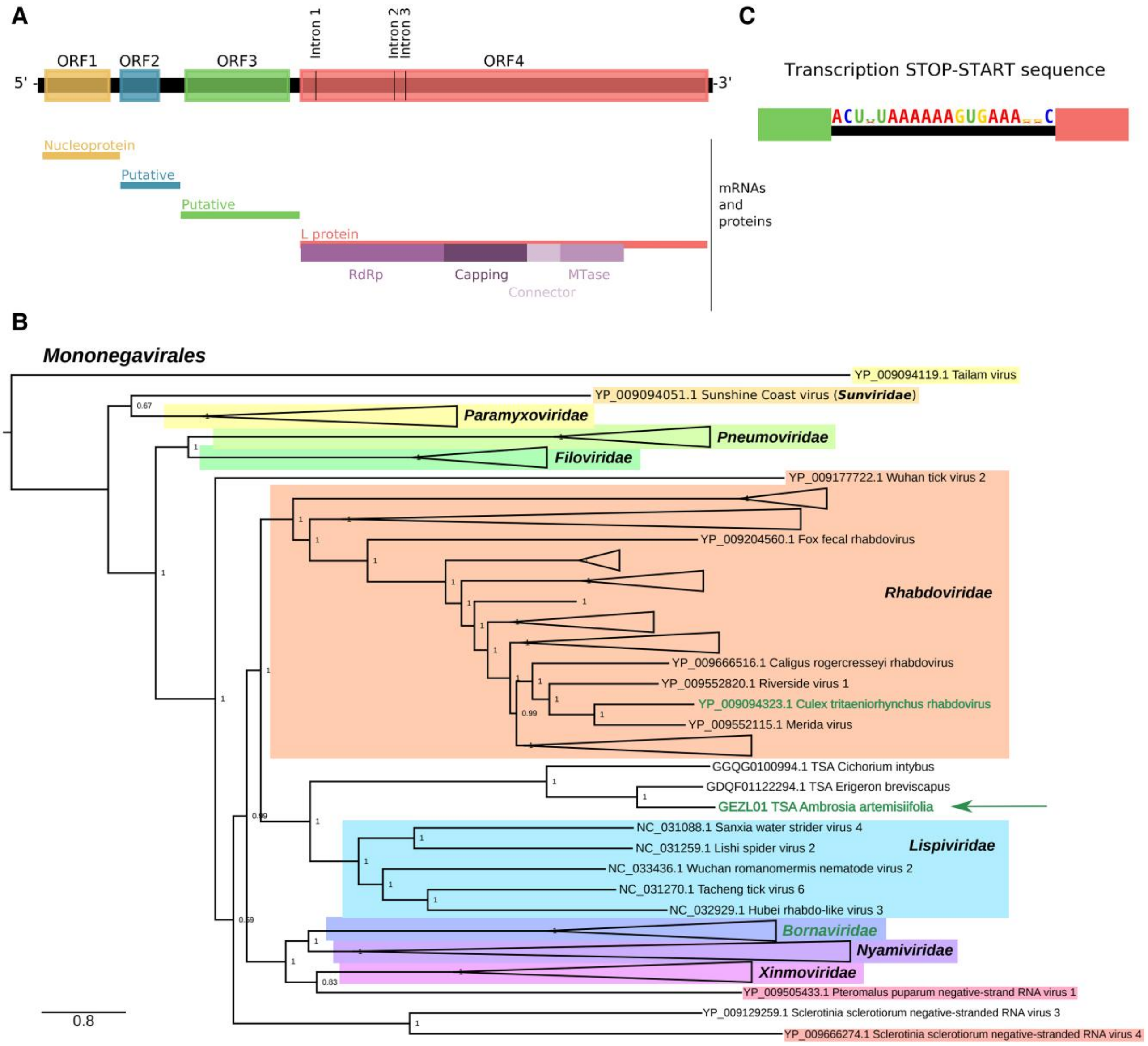
A
Intron 1
Intron 2
Intron 3
ORF1
ORF2
ORF3
ORF4
5'
3'
Nucleoprotein
Putative
Putative
L protein
RdRp
Capping
MTase
Connector
mRNAs and proteins
C
Transcription STOP-START sequence
ACU UAAAAAAGUGAAA C
B
Mononegavirales
YP_009094119.1 Tailam virus
YP_009094051.1 Sunshine Coast virus (Sunviridae)
Paramyxoviridae
Pneumoviridae
Filoviridae
YP_009177722.1 Wuhan tick virus 2
YP_009204560.1 Fox fecal rhabdovirus
Rhabdoviridae
YP_009666516.1 Caligus rogercresseyi rhabdovirus
YP_009552820.1 Riverside virus 1
YP_009094323.1 Culex tritaeniorhynchus rhabdovirus
YP_009552115.1 Merida virus
GGQG0100994.1 TSA Cichorium intybus
GDQF01122294.1 TSA Erigeron breviscapus
GEZL01 TSA Ambrosia artemisiifolia
NC_031088.1 Sanxia water strider virus 4
NC_031259.1 Lishi spider virus 2
Lispiviridae
NC_033436.1 Wuchan romanomermis nematode virus 2
NC_031270.1 Tacheng tick virus 6
NC_032929.1 Hubei rhabdo-like virus 3
Bornaviridae
Nyamiviridae
Xinmoviridae
YP_009505433.1 Pteromalus puparum negative-strand RNA virus 1
YP_009129259.1 Sclerotinia sclerotiorum negative-stranded RNA virus 3
YP_009666274.1 Sclerotinia sclerotiorum negative-stranded RNA virus 4
0.67
0.99
0.83
0.59
0.8

FIG. 8. Splicing in a new rhabdo-like virus sequence. (*A*) Genome map of the rhabdo-like virus derived from the GEZL01 TSA data set. The diagram illustrates ORFs in the antigenome after removal of the identified introns. The positions of the removed introns are indicated. Putative transcription stop–start (TSS) sequences were identified between the ORFs, and the corresponding inferred mRNAs and their products (where identified) are indicated below as well as domains of the L protein. (*B*) Phylogenetic tree of Mononegavirales L protein sequences showing the placement of the GEZL01-derived rhabdo-like virus. For visual convenience, some clades are collapsed into isosceles triangles. Names of sequences/clades with known splicing are written in a different color (green; the GEZL01 sequence is marked with an arrow). See supplementary figure S14, Supplementary Material online for the complete tree. (*C*) Sequence logo generated from the three identified copies of the putative TSS sequence (shown in the antigenome sense), using CIAlign v 1.1.0 (Tumescheit et al. 2022).

Rhabdoviridae) where a single 76 nt intron was identified in the RdRp ORF (Kuwata et al. 2011).

Comparison with BLASTP

As discussed above, all our results were verified with BLASTP against the NCBI nr protein database. In all but 25 cases, the most significant BLAST match was to a protein likely to be derived from an RNA virus.

Using BLAST at this low level of stringency also results in many false positives, reducing its usefulness for large-scale screening. To provide a direct comparison, we performed a BLASTP search using the input RdRp sequences from our pHMMs as a database against two query data sets, our identified viral ORFs and a curated set of known human proteins, the 20,398 reviewed SwissProt proteins in UniprotKB (UniProt Consortium 2021) which should not be RNA viral in origin. Using the default BLASTP settings plus a relaxed *E*-value cutoff of 0.05, all of our viral ORFs were detected. However, using this cutoff, significant hits were also detected for 824 of the human proteins. Using HMMSearch, with cutoffs as discussed in the Materials and Methods, with the Uniprot reviewed proteins, there were no false positives.

Comparison with Structure-Based pHMMs

In order to test if structure-based pHMMs would classify our sequences differently to our taxonomy-based pHMMs, prebuilt pHMMs from the structure-based SUPERFAMILY Hidden Markov Model set, version 1.75 (Gough et al. 2001; see supplementary methods, Supplementary Material online for model IDs), were used to reclassify our sequences. For viral RdRps, this database is not comprehensive; however, 11 RdRp pHMMs are available. These are not classified by viral family, but each was built based on an initial seed sequence. We therefore used the family (or order where family was unavailable) of this seed sequence to approximately classify the pHMMs, giving us five members of the Picornaviridae, two Caliciviridae, two Flaviviridae, one Birnaviridae, and one Reovirales. The Reovirales pHMM was excluded as its seed sequence is poorly taxonomically classified.

From our ORF database, we extracted sequences which, based on our own classification, best matched anything in the "picorna-like" cluster in figure 6 (Astroviridae, Partitiviridae, Solinviviridae, Iflaviridae, Marnaviridae, Polycipiviridae, Chipolycivirdae, Secoviridae, Caliciviridae, Dicistroviridae, Potyviridae, Picornaviridae), the "flavi-like" cluster in figure 6 (Flaviviridae, Tombusviridae), or the Birnaviridae family and used HMMsearch to match these to the SUPERFAMILY pHMMs. Ninety-six percent of the picorna-like sequences best matched a SUPERFAMILY pHMM with a Picornaviridae or Caliciviridae seed sequence, 90% of the flavi-like sequences best matched a SUPERFAMILY pHMM with a Flaviviridae seed sequence, and 100% of the Birnaviridae sequences best matched the SUPERFAMILY pHMM with a Birnaviridae seed sequence. Although this test was limited by the data available in the SUPERFAMILY database, it nevertheless suggests that our classifications are broadly similar to the classifications we would observe using structure-based pHMMs.

Discussion

We generated pHMMs for 77 RNA virus groups and used them to sensitively search the NCBI Transcriptome Shotgun Assembly database. We identified 5,867 RNA virus–derived RdRp-encoding TSA sequences. We supplemented these via a similar search of the NCBI nr/nt virus database. Through this work, we have expanded known virus clades and identified new virus clades. We have also illustrated how we can assess virus gene expression strategies by referring back to raw read data to analyze splicing in a new mononegavirus. Our pHMM search was fast enough to make searching >10 billion ORFs feasible, and postanalysis of the identified sequences confirmed high specificity (zero false positives among over 12,000 hits). We were able to detect many viral sequences in the "twilight zone" of sequence similarity (<35% similarity) (Rost 1999; Cobbin et al. 2021). A list of all the sequences found (as well as representative sequences after clustering by high similarity) and PhyML trees of the different virus groups are available as supplementary data sets 1 and 3, Supplementary Material online. Although a small proportion of these sequences may represent transcribed EVEs, or incorporate in silico misassemblies, we expect that most are likely to represent bona fide RNA viruses. The pHMM models and associated sequence alignments generated are available as supplementary data set 4, Supplementary Material online. We hope that these will be a useful resource for other virus discovery projects. Profile HMMs are not only a more sensitive and specific method than BLAST for finding distant homologues, but also faster (as their use can reduce the number of query-to-subject comparisons).

Although we have shown that our pHMM approach can identify virus groups not included in the original set of pHMMs, they may not be able to identify even more divergent RdRps. Employing an iterative pHMM search method such as JackHMMER (Johnson et al. 2010), whereby newly identified divergent sequences are used to update pHMMs for subsequent searches, might enable identification of even more divergent RdRps (cf. Callanan et al. 2020). Approaches based on predicted secondary or tertiary protein structure such as HHpred (Zimmermann et al. 2018), Phyre2 (Kelley et al. 2015), or AlphaFold (Jumper et al. 2021) could also be useful to find more divergent RdRp sequences (Wolf et al. 2020; Charon et al. 2022; Forgia, Chiapello, et al. 2022; Lee et al. 2022). For example, homology of the quenyavirus RdRp to previously known RNA virus RdRps was detectable with HHpred but not with BLASTP (Obbard et al. 2020) or our approach. On the other hand, pHMM searches are much faster than structural approaches, and this can be a critical issue for high-throughput searches.

Profile HMMs are sensitive to the input sequences used. In our study, we found that in some cases, a family-level pHMM was able to identify many more sequences from one genus than another genus within the same family. Often, this could be traced to a bias in sequence representation during pHMM construction. HMMbuild does not phylogenetically weight input sequences. Therefore, if one genus is highly "overrepresented" in the profile, the profile will be better at finding similar such sequences. On the other hand, one family-level profile may accommodate the possibility of very high divergence, whereas another may be very specific, depending on the diversity provided during pHMM construction. When identifying the best match profile for a given sequence, a sequence from the latter family might have a higher score with the former profile as it tolerates more variation. Thus, when building pHMMs, one should focus on representation of diversity rather than just number of sequences.

Reuse of public transcriptomic data is a cost-effective means of exploring the diversity of the RNA virosphere. In most cases, the data sets were obtained for purposes completely different from virus identification—for example, the divergent axolotl-associated orthomyxo-like sequence came from a transcriptomic data set for a study on limb regeneration (Bryant et al. 2017). A variety of other studies have also searched for RNA viruses in the NCBI TSA database or other transcriptomic studies generated without the express purpose of virus discovery, including Cook et al. (2013), Longdon et al. (2015), Mushegian et al. (2016), Olendraite et al. (2017), Gilbert et al. (2019), Käfer et al. (2019), Lauber et al. (2019, 2021), Rosani et al. (2019), Starr et al. (2019), Callanan et al. (2020), Obbard et al. (2020), Ott Rutar and Kordis (2020), Parry et al. (2020), Wu et al. (2020), Chang et al. (2021), Charon et al. (2021), Paraskevopoulou et al. (2021), Bejerman and Debat (2022), Dheilly et al. (2022), Lee et al. (2022), Mifsud et al. (2022), Neri et al. (2022), Sidharthan et al. (2022), and Zayed et al. (2022). Most of these studies have been limited to certain virus groups and/or certain host groups. Furthermore, although pHMM-based search strategies are being used more frequently (e.g., Gilbert et al. 2019; Käfer et al. 2019; Callanan et al. 2020; Charon et al. 2021, 2022; Lauber et al. 2021; Paraskevopoulou et al. 2021; Zayed et al. 2022), most studies to date have relied on BLAST-type search tools.

A recent study by Edgar et al. (2022) queried the entire SRA database of over 3 million RNA-seq data sets using a novel approach, combining read mapping to RdRp sequences and a novel tool, PalmScan (Babaian and Edgar 2022). PalmScan identifies RdRp-like sequences using the order and composition of the RdRp A, B, and C motifs. This methodology, which allows screening of unassembled sequence read data sets, alongside development of Serratus, a highly optimized computational architecture, allowed screening on an unprecedented scale and identification of over 10^5 putative RdRp sequences. This work contributes a monumental step forward in the field of virus discovery and provides an extremely valuable resource while also demonstrating the enormous potential of exploring the diversity of the RNA virosphere using public transcriptomic data. However, the approach used is likely to have reduced sensitivity for more divergent RdRp sequences.

When analyzing publicly available data, it is difficult to assess the likelihood of contamination, and one must therefore be particularly cautious with host species assignment. Potential sources of contamination include gut contents and microbiota, mold and insects on plant leaves, other internal and external parasites and commensal organisms, contamination during sampling and sample preparation, and contaminated reagents (Cobbin et al. 2021; Harvey and Holmes 2022; Mifsud et al. 2022). One sign of potential contamination is when an identified virus has a very different TSA target host compared with the host species of similar previously known viruses. For example, in our analysis, we found sequences with high similarity to known influenza A virus sequences in two fish data sets. Influenza A virus is known to infect birds and mammals, whereas known fish orthomyxoviruses are much more divergent. The sequences are therefore very likely to derive from contamination, for example, bird fecal material or laboratory contamination. Although beyond the scope of our study, identification of RNA virus fragments occasionally integrated into host genomes (i.e., EVEs) has been used by others as a means to support linkage of uncharacterized virus taxa to broad host taxonomic groups (Shi et al. 2016).

Despite the aforementioned caveats, the identification of multiple related virus sequences from multiple related host species in different studies lends credence to the assignment of virus–host associations—for example, the Asellidae-associated clade of orthomyxo-like viruses and the crustacean-associated clade of giardia-like viruses. Thus, our study allowed us to assess large-scale taxonomic associations and trends in sampled virus diversity. For vertebrate-specific groups such as paramyxoviruses and picornaviruses, we found relatively few new sequences in TSA data sets compared with viruses from nonvertebrate hosts. We also found a substantially larger number of RdRp sequences per data set in nonvertebrate hosts (~16-fold higher in plants and arthropods than in vertebrates). One possible explanation may be that vertebrate samples tend not to comprise whole organism samples, with exclusion of contaminating nontarget organisms from gut contents and surface material, besides sampling a reduced number of tissues and cell types compared with nonvertebrate studies that often comprise the whole organism or even multiple pooled whole organisms. Another possible explanation may lie in the different purposes for which TSA data sets are generated (e.g., many samples from the same species under different experimental laboratory conditions vs. samples from many different species obtained from the wild). Alternatively, it may be linked to differences in the immune systems of vertebrates (e.g., adaptive immunity) and nonvertebrates (e.g., RNA interference) or stem from other major events in the evolution of eukaryotes (cf. Harvey and Holmes 2022). In any case, it is clear that there is an enormous amount of unsampled RNA virus diversity—especially in nonvertebrates—and repurposing of existing data sets provides a valuable route to increasing our understanding of virus diversity, taxonomy, evolution, and ecology.

Materials and Methods

Construction of RNA-Dependent RNA Polymerase pHMMs

Initially, we used RNA virus groups and sequences from the "GRAViTy" analysis of Aiewsakun and Simmonds (2018). In cases where there was only a small number of sequences in a family-like group, or where some more recently published groups of viruses were not mentioned in Aiewsakun and Simmonds, we searched the NCBI taxonomy and nucleotide databases (April 2018) for additional reference sequences in order to make more representative pHMMs. In total, we used 1,793 RdRp protein and RdRp-containing polyprotein sequences to create 77 pHMMs.

Because many RdRps are contained within longer polyproteins, we wished to trim sequences to a core RdRp region. Therefore, we aligned the sequences within each group with MUSCLE v3.8.31 (Edgar 2004) and compared the alignments using HHpred (Zimmermann et al. 2018) with the Pfam and PDB databases (Berman et al. 2000; Finn et al. 2014). Based on the second best matching RdRp (to avoid overfitting if the best match Pfam pHMM contained sequences from the same group), we cropped each alignment from both ends. Where appropriate, the alignments and coordinates for trimming were manually curated based on current knowledge of the families. The cropped alignments were formatted to Stockholm format using AlignIO (Biopython; Cock et al. 2009), and then, pHMM profiles were created using HMMbuild (HMMER 3.1b2; Eddy 2011) with the option −singlemx, to enable profile building if only one sequence was given, and default parameters, except MAP (yes) and STATS (LOCAL: MSV/VITERBI/FORWARD).

The pHMMs were further curated by running HMMsearch (HMMER 3.1b2; Eddy 2011) on all the proteins which had been used to create the pHMMs. The results of this search were used to guide the selection of the threshold values (supplementary fig. S19, Supplementary Material online) for grouping sequences into the “classified,” “ambiguously classified,” and “unclassified” categories.

A list of the pHMMs, number of input sequences, cropping coordinates, and the HMMbuild output information is provided in supplementary table S9, Supplementary Material online, and the pHMMs and alignments with accession numbers are available in supplementary data set 4, Supplementary Material online and at github.com/ingridole/ViralRdRp_pHMMs. To quantify the diversity present among the input sequences for each pHMM, we provide measures of alignment diversity in supplementary data set 5, Supplementary Material online, available at github.com/ingridole/ViralRdRp_pHMMs_2.

Mean alignment identity is the mean pairwise amino acid identity between sequences, calculated as the mean of the output of the “make_similiarity_matrix_input” function of CIAlign version 1.0.18 (Tumescheit et al. 2022), excluding the diagonal and sites which are gaps in both sequences in a pair. Shannon entropy was calculated with the entropy function from the scipy stats library (Virtanen et al. 2020; version 1.7.0).

Sequence Databases for the RdRp Search

To obtain viral reference sequences, we used the assembly_summary.txt file at ftp.ncbi.nih.gov/genomes/refseq/viral/ (May 10, 2018). Viral sequences from this file with a complete genome and the latest version number were downloaded with wget from the NCBI path above. In total, there were 9,566 viral nucleotide reference sequences.

Nonredundant nucleotide (nr/nt) sequences were downloaded (May 14–15, 2018) from the NCBI nucleotide database in 12 sets covering all groups of RNA viruses as well as unclassified and unassigned viruses. The taxonomy and a minimum sequence length of 1,000 nt were specified, and two overrepresented species (hepatitis C virus and influenza A virus) were excluded. The exact search queries are provided in supplementary table S10, Supplementary Material online. In total, there were 274,579 nr/nt sequences. All sequences which were not defined as dsRNA, −ssRNA, or +ssRNA viral sequences were combined into one file. Within that file, we removed sequences named as phages or where the molecule type was specified as genomic DNA in the GenBank format files and sequences that were longer than 30,360 nucleotides (so as to remove many phage and DNA virus sequences). Following this, 7,059 sequences remained in the “others” file. Next, clustering within each of the four groups (dsRNA, −ssRNA, +ssRNA, others) was performed using CDHIT (versions 4.6 and 4.7; Li and Godzik 2006; Fu et al. 2012) to remove similar sequences (>80% pairwise nucleotide identity), in each case retaining the longest sequence from a group of similar sequences. After this step, 14,832 sequences were left for further processing (supplementary table S11, Supplementary Material online).

Next, we used BLASTN (v2.2.31+, built January 7, 2016; Altschul et al. 1990; Camacho et al. 2009) to compare the remaining 14,832 nr/nt sequences with the 9,566 reference sequences, and we removed nr/nt sequences that had ≥80% nucleotide identity to and ≥80% coverage by at least one reference sequence. After this step, 9,855 nr/nt sequences were left and used in the further analysis.

Sequences from the NCBI TSA database (National Center for Biotechnology Information (NCBI), 2017) were downloaded (October 29, 2017) based on the wgs_selector file from the TSA browser at https://www.ncbi.nlm.nih.gov/Traces/wgs/? view=TSA. We downloaded 2,648 different TSA data sets covering ~1,800 unique taxonomic groups.

The exact commands for downloading and clustering of sequences are provided in supplementary methods, Supplementary Material online.

Analysis of Protein Sequences

For all TSA, reference, and nr/nt nucleotide sequences, open reading frames (ORFs) were retrieved using GETORF (EMBOSS v5.5 and v6.6; Rice et al. 2000) using three different genetic code tables (the standard genetic code, table = 1; stop codon UGA redefined as Trp, table = 4; and stop codons UAA and UAG redefined as Gln, table = 6) and identifying regions between consecutive stop codons (parameter -find 0) with a minimum length of 60 nucleotides (parameter -minsize 60) to allow for detection even of RdRp fragments. In total, this resulted in $>13 \times 10^9$ ORFs (supplementary table S12, Supplementary Material online).

HMMsearch

To search for viral RdRps, we searched the retrieved ORFs using HMMsearch (HMMER 3.1b2; Eddy 2011) with the 77 family-level pHMM profiles. HMMsearch was performed for each genetic code table data set separately, adjusting the *E*-value based on database size to maintain a constant *P*-value threshold of 1×10^{-6}. The next step was to find the best hit for each ORF among the different matched pHMMs and sort the match to one of three groups: classified, ambiguously classified, or unclassified based on our IDscore metric, which is the bit score divided by the length in amino acids of the alignment between an ORF and a matched pHMM (supplementary fig. S19, Supplementary Material online). If the highest IDscore was lower than 0.25, a sequence was sorted into the unclassified group. Otherwise, if a sequence had statistically significant hits to more than one of the pHMMs and the second best IDscore was less than 20% lower than the best IDscore, the sequence was sorted into the ambiguously classified group. Sequences with an IDscore of 0.25 or higher, and at least 20% difference in IDscore between the first and second best hits, were sorted into the classified group and classified according to the best match pHMM.

Processing of RdRp-Encoding ORFs

Since there could be multiple partially identical ORFs due to the use of three genetic code tables, we next used pairwise global alignments (Biopython pairwise2.align.globalxx; Cock et al. 2009) to compare all ORF sequences with the same original nucleotide accession. If, in any pairwise alignment, the number of identities divided by the shorter ORF length was 1, then the shorter ORF was removed from further analysis.

The remaining ORFs were trimmed according to the start and end positions of the amino acid region which mapped to the best hit pHMM to approximate the core part of the RdRp. Then, alignments were generated for each pHMM group, combining reference, nr/nt, and TSA-derived trimmed ORF sequences, using MUSCLE v3.8.31 (Edgar 2004).

Grouping into 60 Clusters

After applying the grouping scheme and discarding similar sequences using CDHIT-EST, for some of the 77 viral pHMM-based groups, there were fewer than 10 classified sequences remaining. For convenience, these sequences were joined with other groups which were somewhat taxonomically similar, with preference given to those groups which themselves had fewer sequences, aiming for each cluster to contain close to or more than 20 sequences. Thus, 60 clusters of classified sequences were created. Clustering was not performed for the ambiguously classified or unclassified groups of sequences.

Verification

In order to verify that sequences were true viral RdRps rather than false positives, and to identify chimeric sequences, all ORFs identified as encoding (part of) an RdRp were checked using a BLAST-based approach. For verification with a third approach, HHSearch (part of HHSuite v3.3.0; Steinegger et al. 2019) was used to compare our set of putative viral ORFs to the Pfam database (Finn et al. 2014). We compared the false positive rate for HMMER and BLASTP using a control set of Uniprot proteins. All verification steps are described in full in the supplementary methods, Supplementary Material online. Full results are provided in supplementary data set 2, Supplementary Material online.

Heatmap for Relationships between Viral Groups

For all ORFs in the classified group, we identified the first and second best hit pHMMs based on IDscore. Then, for each represented pHMM group, we counted the number of co-occurrences with each of the pHMMs. This resulted in a large matrix, where rows represent the first best hit pHMM and columns the second best hit pHMM. Then, we normalized the values by row. Results were visualized in Python (Matplotlib, pyplot and colors, and Numpy packages). The virus taxonomic groups of the pHMMs were manually sorted in an order consistent with ICTV taxonomy and/or the phylogeny of Wolf et al. (2018).

Phylogenetic Trees

Phylogenetic trees were constructed using amino acid alignments and PhyML v3.1 (amino acid model LG, tree topology search operation SPR, and gamma distribution shape parameter 20; Guindon and Gascuel 2003; Guindon et al. 2010). In some cases, similar sequences were discarded using pairwise comparisons in Python or using CDHIT. Alignments were prepared with MUSCLE v3.8.31 (Edgar 2004) and their format changed to Phylip using SEQRET (EMBOSS v5.5 and v6.6; Rice et al. 2000). FigTree v1.4.3 (Rambaut 2006) was used for visualization of trees using the midpoint rooting option and showing the bootstrap support values as node labels.

Host Identification for TSA Sequences

To identify the putative host species and taxonomic group for RdRp sequences found in the TSA database, we used EFETCH (Entrez Programming Utilities; Sayers et al. 2022) and each TSA contig accession number to retrieve a corresponding taxonomy line. Next, we assigned the type of putative host as follows. If "Eukaryota" was not in the taxonomy line, the type was set to metagenomics (or environmental). In other cases, we searched for keywords in the taxonomy line: fungi (keyword "Fungi"), plants (keyword "Viridiplantae"), vertebrates (keyword "Vertebrata"), and arthropods (keyword "Arthropoda"). For invertebrates, we required the keyword "Metazoa" but not "Vertebrata" or "Arthropoda," whereas for Protista, the keywords "Metazoa," "Fungi," and "Viridiplantae" had to be absent.

To generate the updated taxonomy column in supplementary data set 1, Supplementary Material online, corresponding to the most recent (June 28, 2022) NCBI taxonomy, derived from the ICTV 2021 taxonomy (Schoch et al. 2020; Walker et al. 2021), the taxonomic classification assigned to the GenBank record for each accession number was used, again retrieved using EFETCH. Where this ID corresponded to the host rather than the virus, it was not included. To assign a taxonomic lineage to the pHMM family/genus level classifications, the name of the pHMM was used except in the cases of the Sobemoviridae (which is now the genus *Sobemovirus*), Ophioviridae (which is now renamed as Aspaviridae), and Rubellavirus (which is now Rubivirus). Zhaovirus, Yanvirus, and Weivirus are not classified in the current ICTV taxonomy.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

This work was funded by Wellcome Trust Senior Research Fellowships (106207/Z/14/Z, 220814/Z/20/Z) and a European Research Council grant (646891). For the purpose of open access, the author has applied a CC BY public copyright license to any Author Accepted Manuscript version arising from this submission.

Data Availability

The table of identified virus RdRp-derived sequences (supplementary data set 1, Supplementary Material online), results of our BLAST and HHSearch verification (supplementary data set 2, Supplementary Material online), PhyML trees of the different groups of sequences (supplementary data set 3, Supplementary Material online), and alignment statistics for the input alignments for the pHMMs (supplementary data set 5, Supplementary Material online) are available at github.com/ingridole/ViralRdRp_pHMMs_2; release v1.0.1 is associated with this manuscript. The pHMMs and the alignments used to create them (supplementary data set 4, Supplementary Material online) are available at github.com/ingridole/ViralRdRp_pHMMs; release v1.0.1 is associated with this manuscript.

References

Aiewsakun P, Katzourakis A. 2015. Endogenous viruses: connecting recent and ancient viral evolution. *Virology* **479–480**:26–37.

Aiewsakun P, Simmonds P. 2018. The genomic underpinnings of eukaryotic virus taxonomy: creating a sequence-based framework for family-level virus classification. *Microbiome* **6**(1):38.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol.* **215**(3):403–410.

Arjona-Lopez JM, Telengech P, Jamal A, Hisano S, Kondo H, Yelin MD, Arjona-Girona I, Kanematsu S, Lopez-Herrera CJ, Suzuki N. 2018. Novel, diverse RNA viruses from Mediterranean isolates of the phytopathogenic fungus, *Rosellinia necatrix*: insights into evolutionary biology of fungal viruses. *Environ Microbiol.* **20**(4): 1464–1483.

Babaian A, Edgar RC. 2022. Ribovirus classification by a polymerase barcode sequence. *PeerJ.* **10**:e14055. https://doi.org/10.7717/peerj.14055

Baltimore D. 1971. Expression of animal virus genomes. *Bacteriol Rev.* **35**(3):235–241.

Batson J, Dudas G, Haas-Stapleton E, Kistler AL, Li LM, Logan P, Ratnasiri K, Retallack H. 2021. Single mosquito metatranscriptomics identifies vectors, emerging pathogens and reservoirs in one assay. *eLife* **10**:e68353.

Bejerman N, Debat H. 2022. Exploring the tymovirales landscape through metatranscriptomics data. *Arch Virol.* **167**(9): 1785–1803.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. 2000. The protein data bank. *Nucleic Acids Res.* **28**(1):235–242.

Brown JW, Smith P, Simpson CG. 1996. Arabidopsis consensus intron sequences. *Plant Mol Biol.* **32**(3):531–535.

Bruenn JA. 2003. A structural and primary sequence comparison of the viral RNA-dependent RNA polymerases. *Nucleic Acids Res.* **31**(7):1821–1829.

Bryant DM, Johnson K, DiTommaso T, Tickle T, Couger MB, Payzin-Dogru D, Lee TJ, Leigh ND, Kuo T-H, Davis FG, *et al.* 2017. A tissue-mapped axolotl de novo transcriptome enables identification of limb regeneration factors. *Cell Rep.* **18**(3):762–776.

Bushmanova E, Antipov D, Lapidus A, Prjibelski AD. 2019. rnaSPAdes: a de novo transcriptome assembler and its application to RNA-seq data. *GigaScience* **8**(9):giz100.

Callanan J, Stockdale SR, Shkoporov A, Draper LA, Ross RP, Hill C. 2020. Expansion of known ssRNA phage genomes: from tens to over a thousand. *Sci Adv.* **6**(6):eaay5981.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics.* **10**:421.

Chang T, Hirai J, Hunt BPV, Suttle CA. 2021. Arthropods and the evolution of RNA viruses. *bioRxiv.* 2021.05.30.446314.

Charon J, Buchmann JP, Sadiq S, Holmes EC. 2022. RdRp-scan: a bioinformatic resource to identify and annotate divergent RNA viruses in metagenomic sequence data. *Virus Evol.* **8**(2):veac082.

Charon J, Marcelino VR, Wetherbee R, Verbruggen H, Holmes EC. 2020. Metatranscriptomic identification of diverse and divergent RNA viruses in green and chlorarachniophyte algae cultures. *Viruses* **12**(10):1180.

Charon J, Murray S, Holmes EC. 2021. Revealing RNA virus diversity and evolution in unicellular algae transcriptomes. *Virus Evol.* **7**(2):veab070.

Chen Y-M, Sadiq S, Tian J-H, Chen X, Lin X-D, Shen J-J, Chen H, Hao Z-Y, Wille M, Zhou Z-C, *et al.* 2022. RNA viromes from terrestrial sites across China expand environmental viral diversity. *Nat Microbiol.* **7**(8):1312–1323.

Chiapello M, Rodríguez-Romero J, Ayllón MA, Turina M. 2020. Analysis of the virome associated to grapevine downy mildew lesions reveals new mycovirus lineages. *Virus Evol.* **6**(2):veaa058.

Chiba S, Kondo H, Tani A, Saisho D, Sakamoto W, Kanematsu S, Suzuki N. 2011. Widespread endogenization of genome sequences of non-retroviral RNA viruses into plant genomes. *PLoS Pathog.* **7**(7):e1002146.

Cobbin JC, Charon J, Harvey E, Holmes EC, Mahar JE. 2021. Current challenges to virus discovery by meta-transcriptomics. *Curr Opin Virol.* **51**:48–55.

Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, *et al.* 2009. Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics.* **25**(11): 1422–1423.

Cook S, Chung BY-W, Bass D, Moureau G, Tang S, McAlister E, Culverwell CL, Glücksman E, Wang H, Brown TDK, *et al.* 2013. Novel virus discovery and genome reconstruction from field RNA samples reveals highly divergent viruses in dipteran hosts. *PLoS One.* **8**(11):e80720.

Crooks GE, Hon G, Chandonia J-M, Brenner SE. 2004. Weblogo: a sequence logo generator. *Genome Res.* **14**(6):1188–1190.

Cubitt B, Oldstone C, Valcarcel J, Carlos de la Torre J. 1994. RNA splicing contributes to the generation of mature mRNAs of Borna disease virus, a non-segmented negative strand RNA virus. *Virus Res.* **34**(1):69–79.

Dance A. 2021. Beyond coronavirus: the virus discoveries transforming biology. *Nature* **595**(7865):22–25.

Debat HJ. 2018. Expanding the size limit of RNA viruses: evidence of a novel divergent nidovirus in California sea hare, with a ∼35.9 kb virus genome. *bioRxiv* 307678.

Dheilly NM, Lucas P, Blanchard Y, Rosario K. 2022. A world of viruses nested within parasites: unraveling viral diversity within parasitic flatworms (Platyhelminthes). *Microbiol Spectr.* **10**(3):e0013822.

Di Paola N, Dheilly NM, Junglen S, Paraskevopoulou S, Postler TS, Shi M, Kuhn JH. 2022. *Jingchuvirales*: a new taxonomical framework for a rapidly expanding order of unusual monjiviricete viruses broadly distributed among arthropod subphyla. *Appl Environ Microbiol.* **88**(6):e0195421.

Dinan AM, Lukhovitskaya NI, Olendraite I, Firth AE. 2020. A case for a negative-strand coding sequence in a group of positive-sense RNA viruses. *Virus Evol.* **6**(1):veaa007.

Dolja VV, Koonin EV. 2018. Metagenomics reshapes the concepts of RNA virus evolution by revealing extensive horizontal virus transfer. *Virus Res.* **244**:36–52.

Eddy SR. 2011. Accelerated profile HMM searches. *PLoS Comput Biol.* **7**(10):e1002195.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**(5):1792–1797.

Edgar RC, 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**(7895):142–147.

Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, *et al.* 2014. Pfam: the protein families database. *Nucleic Acids Res.* **42**(Database issue):D222–D230.

Forgia M, Chiapello M, Daghino S, Pacifico D, Crucitti D, Oliva D, Ayllon M, Turina M, Turina M. 2022. Three new clades of putative viral RNA-dependent RNA polymerases with rare or unique catalytic triads discovered in libraries of ORFans from powdery mildews and the yeast of oenological interest *Starmerella bacillaris*. *Virus Evol.* **8**(1):veac038.

Forgia M, Navarro B, Daghino S, Cervera A, Gisel A, Perotto S, Aghayeva DN, Akinyuwa MF, Gobbi E, Zheludev IN, *et al.* 2022. Extant hybrids of RNA viruses and viroid-like elements. *bioRxiv* 2022.08.21.504695.

Frith MC, Saunders NFW, Kobe B, Bailey TL. 2008. Discovering sequence motifs with arbitrary insertions and deletions. *PLoS Comput Biol.* **4**(4):e1000071.

Fu L, Niu B, Zhu Z, Wu S, Li W. 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*. **28**(23): 3150–3152.

Geoghegan JL, Holmes EC. 2017. Predicting virus emergence amid evolutionary noise. *Open Biol.* **7**(10):170189.

Gilbert C, Belliardo C. 2022. The diversity of endogenous viral elements in insects. *Curr Opin Insect Sci.* **49**:48–55.

Gilbert KB, Holcomb EE, Allscheid RL, Carrington JC. 2019. Hiding in plain sight: new virus genomes discovered via a systematic analysis of fungal public transcriptomes. *PLoS One*. **14**(7): e0219207.

Gildow FE, DArcy CJ. 1988. Barley and oats as reservoirs for an aphid virus and the influence on barley yellow dwarf virus transmission. *Phytopathology* **78**:811–881.

Gorbalenya AE, Pringle FM, Zeddam J-L, Luke BT, Cameron CE, Kalmakoff J, Hanzlik TN, Gordon KHJ, Ward VK. 2002. The palm subdomain-based active site is internally permuted in viral RNA-dependent RNA polymerases of an ancient lineage. *J Mol Biol.* **324**(1):47–62.

Gough J, Karplus K, Hughey R, Chothia C. 2001. Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *J Mol Virol.* **313**(4):903–919.

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

Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol.* **59**(3):307–321.

Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol.* **52**(5):696–704.

Harvey E, Holmes EC. 2022. Diversity and evolution of the animal virome. *Nat Rev Microbiol.* **20**(6):321–334.

Inglis SC, Brown CM. 1981. Spliced and unspliced RNAs encoded by virion RNA segment 7 of influenza virus. *Nucleic Acids Res.* **9**(12): 2727–2740.

Jácome R, Campillo-Balderas JA, Becerra A, Lazcano A. 2022. Structural analysis of monomeric RNA-dependent polymerases revisited. *J Mol Evol.* **90**(3–4):283–295.

Johnson LS, Eddy SR, Portugaly E. 2010. Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics*. **11**:431.

Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, *et al.* 2021. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**(7873):583–589.

Käfer S, Paraskevopoulou S, Zirkel F, Wieseke N, Donath A, Petersen M, Jones TC, Liu S, Zhou X, Middendorf M, *et al.* 2019. Re-assessing the diversity of negative strand RNA viruses in insects. *PLoS Pathog.* **15**(12):e1008224.

Kartali T, Nyilasi I, Szabó B, Kocsubé S, Patai R, Polgár TF, Nagy G, Vágvölgyi C, Papp T. 2019. Detection and molecular characterization of novel dsRNA viruses related to the *Totiviridae* family in *Umbelopsis ramanniana*. *Front Cell Infect Microbiol.* **9**:249.

Katzourakis A, Gifford RJ. 2010. Endogenous viral elements in animal genomes. *PLoS Genet.* **6**(11):e1001191.

Keese PK, Gibbs A. 1992. Origins of genes: "big bang" or continuous creation? *Proc Natl Acad Sci U S A.* **89**(20):9489–9493.

Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJE. 2015. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat Protoc.* **10**(6):845–858.

Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. 2019. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol.* **37**(8):907–915.

King AMQ, Lefkowitz EJ, Mushegian AR, Adams MJ, Dutilh BE, Gorbalenya AE, Harrach B, Harrison RL, Junglen S, Knowles NJ, *et al.* 2018. Changes to taxonomy and the International Code of Virus Classification and Nomenclature ratified by the International Committee on Taxonomy of Viruses (2018). *Arch Virol.* **163**(9):2601–2631.

Kinsella CM, Deijs M, Gittelbauer HM, van der Hoek L, van Dijk K. 2022. Human clinical isolates of pathogenic fungi are host to diverse mycoviruses. *Microbiol Spectr.* **10**(5):e0161022.

Kochs G, Weber F, Gruber S, Delvendahl A, Leitz C, Haller O. 2000. Thogoto virus matrix protein is encoded by a spliced mRNA. *J Virol.* **74**(22):10785–9.

Koonin EV. 1991. The phylogeny of RNA-dependent RNA polymerases of positive-strand RNA viruses. *J Gen Virol.* **72**(Pt 9): 2197–2206.

Koonin EV, Dolja VV. 1993. Evolution and taxonomy of positive-strand RNA viruses: implications of comparative analysis of amino acid sequences. *Crit Rev Biochem Mol Biol.* **28**(5):375–430.

Koonin EV, Dolja VV, Krupovic M. 2015. Origins and evolution of viruses of eukaryotes: the ultimate modularity. *Virology* **479–480**:2–25.

Koonin EV, Dolja VV, Krupovic M, Varsani A, Wolf YI, Yutin N, Zerbini FM, Kuhn JH. 2020. Global organization and proposed megataxonomy of the virus world. *Microbiol Mol Biol Rev.* **84**(2): e00061-19.

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

Kuhn JH, Wolf YI, Krupovic M, Zhang Y-Z, Maes P, Dolja VV, Koonin EV. 2019. Classify viruses—the gain is worth the pain. *Nature* **566**(7744):318–320.

Kuwata R, Isawa H, Hoshino K, Tsuda Y, Yanase T, Sasaki T, Kobayashi M, Sawabe K. 2011. RNA splicing in a new rhabdovirus from *Culex* mosquitoes. *J Virol.* **85**(13):6185–6196.

Lauber C, Seifert M, Bartenschlager R, 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.

Lauber C, Vaas J, Klingler F, Mutz P, Gorbalenya AE, Bartenschlager R, Seitz S. 2021. Deep mining of the sequence read archive reveals bipartite coronavirus genomes and inter-family spike glycoprotein recombination. *bioRxiv* 2021.10.20.465146.

Lee BD, Neri U, Roux S, Wolf YI, Camargo AP, Krupovic M, Simmonds P, Kyrpides N, Gophna U, Dolja VV, *et al.* 2022. Mining metatranscriptomes reveals a vast world of viroid-like circular RNAs. *Cell* **186**(3):646–661.

Lee SH, Yun S-H, Chun J, Kim D-H. 2017. Characterization of a novel dsRNA mycovirus of Trichoderma atroviride NFCF028. *Arch Virol.* **162**(4):1073–1077.

Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*. **22**(13):1658–1659.

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

Longdon B, Murray GGR, Palmer WJ, Day JP, Parker DJ, Welch JJ, Obbard DJ, Jiggins FM. 2015. The evolution, diversity, and host associations of rhabdoviruses. *Virus Evol.* **1**(1):vev014.

Mifsud JCO, Gallagher RV, Holmes EC, Geoghegan JL. 2022. Transcriptome mining expands knowledge of RNA viruses across the plant kingdom. *J Virol.* **96**(24):e0026022.

Mönttinen HAM, Ravantti JJ, Poranen MM. 2021. Structure unveils relationships between RNA virus polymerases. *Viruses* **13**(2):313.

Mushegian A, Shipunov A, Elena SF. 2016. Changes in the composition of the RNA virome mark evolutionary transitions in green plants. *BMC Biol.* **14**:68.

Nasir A, Forterre P, Kim KM, Caetano-Anolles G. 2014. The distribution and impact of viral lineages in domains of life. *Front Microbiol.* **5**: 194.

National Center for Biotechnology Information (NCBI). 2017. Transcriptome shotgun assembly sequence database. Retrieved from https://www.ncbi.nlm.nih.gov/genbank/tsa/

Neri U, Wolf YI, Roux S, Camargo AP, Lee B, Kazlauskas D, Chen IM, Ivanova N, Zeigler Allen L, Paez-Espino D, *et al.* 2022. Expansion of the global RNA virome reveals diverse clades of bacteriophages. *Cell* **185**:4023–4037.e18.

Nguyen M, Haenni A-L. 2003. Expression strategies of ambisense viruses. *Virus Res.* **93**(2):141–150.

Nibert ML. 2017. Mitovirus UGA(Trp) codon usage parallels that of host mitochondria. *Virology* **507**:96–100.

Obbard DJ. 2018. Expansion of the metazoan virosphere: progress, pitfalls, and prospects. *Curr Opin Virol.* **31**:17–23.

Obbard DJ, Shi M, Roberts KE, Longdon B, Dennis AB. 2020. A new lineage of segmented RNA viruses infecting animals. *Virus Evol.* **6**(1):vez061.

Ogino T, Green TJ. 2019. RNA synthesis and capping by non-segmented negative strand RNA viral polymerases: lessons from a prototypic virus. *Front Microbiol.* **10**:1490.

Olendraite I, Lukhovitskaya NI, Porter SD, Valles SM, Firth AE. 2017. Polycipiviridae: a proposed new family of polycistronic picorna-like RNA viruses. *J Gen Virol.* **98**(9):2368–2378.

Ortiz-Baez AS, Eden J-S, Moritz C, Holmes EC. 2020. A divergent articulavirus in an Australian gecko identified using metatranscriptomics and protein structure comparisons. *Viruses* **12**(6):613.

Ott Rutar S, Kordis D. 2020. Analysis of the RNA virome of basal hexapods. *PeerJ* **8**:e8336.

Paraskevopoulou S, Käfer S, Zirkel F, Donath A, Petersen M, Liu S, Zhou X, Drosten C, Misof B, Junglen S. 2021. Viromics of extant insect orders unveil the evolution of the flavi-like superfamily. *Virus Evol.* **7**(1):veab030.

Parry R, Wille M, Turnbull OMH, Geoghegan JL, Holmes EC. 2020. Divergent influenza-like viruses of amphibians and fish support an ancient evolutionary association. *Viruses* **12**(9):1042.

Pfaff F, Rubbenstroth D. 2021. Two novel bornaviruses identified in colubrid and viperid snakes. *Arch Virol.* **166**(9):2611–2614.

Rambaut A. 2006. FigTree. Retrieved from http://tree.bio.ed.ac.uk/software/figtree/

Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.* **16**(6):276–277.

Roossinck MJ, Sabanadzovic S, Okada R, Valverde RA. 2011. The remarkable evolutionary history of endornaviruses. *J Gen Virol.* **92**(Pt 11):2674–2678.

Rosani U, Shapiro M, Venier P, Allam B. 2019. A needle in a haystack: tracing bivalve-associated viruses in high-throughput transcriptomic data. *Viruses* **11**(3):205.

Rosario K, Van Bogaert N, López-Figueroa NB, Paliogiannis H, Kerr M, Breitbart M. 2022. Freshwater macrophytes harbor viruses representing all five major phyla of the RNA viral kingdom *Orthornavirae*. *PeerJ* **10**:e13875.

Rost B. 1999. Twilight zone of protein sequence alignments. *Protein Eng Des Sel.* **12**(2):85–94.

Saberi A, Gulyaeva AA, Brubacher JL, Newmark PA, Gorbalenya AE. 2018. A planarian nidovirus expands the limits of RNA genome size. *PLoS Pathog.* **14**(11):e1007314.

Sayers EW, Bolton EE, Brister JR, Canese K, Chan J, Comeau D, Connor R, Funk K, Kelly C, Kim S, *et al.* 2022. Database resources of the national center for biotechnology information. *Nucleic Acids Res.* **50**(D1):D20–D26.

Schneider PA, Schneemann A, Lipkin WI. 1994. RNA splicing in borna disease virus, a nonsegmented, negative-strand RNA virus. *J Virol.* **68**(8):5007–5012.

Schoch CL, Ciufo S, Domrachev M, Hotton CL, Kannan S, Khovanskaya R, Leipe D, Mcveigh R, O'Neill K, Robbertse B, *et al.* 2020. NCBI taxonomy: a comprehensive update on curation, resources and tools. *Database* **2020**:baaa062.

Shi M, Lin X-D, Chen X, Tian J-H, Chen L-J, Li K, Wang W, Eden J-S, Shen J-J, Liu L, *et al.* 2018. The evolutionary history of vertebrate RNA viruses. *Nature* **556**(7700):197–202.

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**(7634):539–543.

Shih SR, Suen PC, Chen YS, Chang SC. 1998. A novel spliced transcript of influenza A/WSN/33 virus. *Virus Genes.* **17**(2):179–183.

Sidharthan VK, Rajeswari V, Baranwal VK. 2022. Analysis of public domain plant transcriptomes expands the phylogenetic diversity of the family Secoviridae. *Virus Genes.* **58**(6):598–604.

Starr EP, Nuccio EE, Pett-Ridge J, Banfield JF, Firestone MK. 2019. Metatranscriptomic reconstruction reveals RNA viruses with the potential to shape carbon cycling in soil. *Proc Natl Acad Sci U S A.* **116**(51):25900–25908.

Steinegger M, Meier M, Mirdita M, Vöhringer H, Haunsberger SJ, Söding J. 2019. HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics.* **20**(1):473.

Sutela S, Forgia M, Vainio EJ, Chiapello M, Daghino S, Vallino M, Martino E, Girlanda M, Perotto S, Turina M. 2020. The virome from a collection of endomycorrhizal fungi reveals new viral taxa with unprecedented genome organization. *Virus Evol.* **6**(2):veaa076.

te Velthuis AJW. 2014. Common and unique features of viral RNA-dependent polymerases. *Cell Mol Life Sci.* **71**(22):4403–4420.

Tomonaga K, Kobayashi T, Lee BJ, Watanabe M, Kamitani W, Ikuta K. 2000. Identification of alternative splicing and negative splicing activity of a nonsegmented negative-strand RNA virus, Borna disease virus. *Proc Natl Acad Sci U S A.* **97**(23):12788–12793.

Tumescheit C, Firth AE, Brown K. 2022. CIAlign: a highly customisable command line tool to clean, interpret and visualise multiple sequence alignments. *PeerJ* **10**:e12983.

Turnbull OMH, Ortiz-Baez AS, Eden J-S, Shi M, Williamson JE, Gaston TF, Zhang Y-Z, Holmes EC, Geoghegan JL. 2020. Meta-transcriptomic identification of divergent *Amnoonviridae* in fish. *Viruses* **12**(11):1254.

UniProt Consortium. 2021. Uniprot: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* **49**(D1):D480–D489.

Virág E, Hegedűs G, Barta E, Nagy E, Mátyás K, Kolics B, Taller J. 2016. Illumina sequencing of common (short) ragweed (*Ambrosia artemisiifolia* L.) reproductive organs and leaves. *Front Plant Sci.* **7**: 1506.

Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, *et al.* 2020. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nat Methods.* **17**(3):261–272.

Waldron FM, Stone GN, Obbard DJ. 2018. Metagenomic sequencing suggests a diversity of RNA interference-like responses to viruses across multicellular eukaryotes. *PLoS Genet.* **14**(7):e1007533.

Walker PJ, Siddell SG, Lefkowitz EJ, Mushegian AR, Adriaenssens EM, Alfenas-Zerbini P, Davison AJ, Dempsey DM, Dutilh BE, García ML, *et al.* 2021. Changes to virus taxonomy and to the International Code of Virus Classification and Nomenclature ratified by the International Committee on Taxonomy of Viruses (2021). *Arch Virol.* **166**(9):2633–2648.

Wamonje FO, Michuki GN, Braidwood LA, Njuguna JN, Musembi Mutuku J, Djikeng A, Harvey JJW, Carr JP. 2017. Viral metagenomics of aphids present in bean and maize plots on mixed-use farms in Kenya reveals the presence of three dicistroviruses including a novel Big Sioux River virus-like dicistrovirus. *Virol J.* **14**(1):188.

Wang D. 2022. The enigma of picobirnaviruses: viruses of animals, fungi, or bacteria? *Curr Opin Virol.* **54**:101232.

Wang L, Zhang J, Zhang H, Qiu D, Guo L. 2016. Two novel relative double-stranded RNA mycoviruses infecting *Fusarium poae* strain SX63. *Int J Mol Sci.* **17**(5):641.

Wise HM, Hutchinson EC, Jagger BW, Stuart AD, Kang ZH, Robb N, Schwartzman LM, Kash JC, Fodor E, Firth AE, *et al.* 2012. Identification of a novel splice variant form of the influenza A virus M2 ion channel with an antigenically distinct ectodomain. *PLoS Pathog.* **8**(11):e1002998.

Wolf YI, Kazlauskas D, Iranzo J, Lucía-Sanz A, Kuhn JH, Krupovic M, Dolja VV, Koonin EV. 2018. Origins and evolution of the global RNA virome. *mBio* **9**(6):e02329-18.

Wolf YI, Silas S, Wang Y, Wu S, Bocek M, Kazlauskas D, Krupovic M, Fire A, Dolja VV, Koonin EV. 2020. Doubling of the known set of RNA viruses by metagenomic analysis of an aquatic virome. *Nat Microbiol.* **5**(10):1262–1270.

Wu H, Pang R, Cheng T, Xue L, Zeng H, Lei T, Chen M, Wu S, Ding Y, Zhang J, *et al.* 2020. Abundant and diverse RNA viruses in insects revealed by RNA-seq analysis: ecological and evolutionary implications. *mSystems* **5**(4):e00039-20.

Zayed AA, Wainaina JM, Dominguez-Huerta G, Pelletier E, Guo J, Mohssen M, Tian F, Pratama AA, Bolduc B, Zablocki O, *et al.* 2022. Cryptic and abundant marine viruses at the evolutionary origins of earths RNA virome. *Science* **376**(6589):156–162.

Zhang Y-Z, Chen Y-M, Wang W, Qin X-C, Holmes EC. 2019. Expanding the RNA virosphere by unbiased metagenomics. *Annu Rev Virol.* **6**(1):119–139.

Zimmermann L, Stephens A, Nam S-Z, Rau D, Kübler J, Lozajic M, Gabler F, Söding J, Lupas AN, Alva V. 2018. A completely reimplemented MPI bioinformatics toolkit with a new HHpred server at its core. *J Mol Biol.* **430**(15):2237–2243.