Using artificial intelligence to document the hidden RNA virosphere

Xin Hou[1,16], Yong He[2,16], Pan Fang[2], Shi-Qiang Mei[1], Zan Xu[2], Wei-Chen Wu[1], Jun-Hua Tian[3], Shun Zhang[2], Zhen-Yu Zeng[2], Qin-Yu Gou[1], Gen-Yang Xin[1], Shi-Jia Le[1], Yin-Yue Xia[4], Yu-Lan Zhou[5], Feng-Ming Hui[6,7], Yuan-Fei Pan[8], John-Sebastian Eden[9], Zhao-Hui Yang[10], Chong Han[11], Yue-Long Shu[12], Deyin Guo[13], Jun Li[14], Edward C. Holmes[9,15], Zhao-Rong Li[2] & Mang Shi[1]

[1]State key laboratory for biocontrol, the Centre for Infection and Immunity Studies, School of Medicine, Shenzhen Campus of Sun Yat-sen University, Sun Yat-sen University, Shenzhen, China.

[2]Industry Research and Development Department, Alibaba Cloud Intelligence, Alibaba Group, Hangzhou, China.

[3]Wuhan Centers for Disease Control and Prevention, Wuhan, China.

[4]Polar Research Institute of China, Shanghai, China.

[5]Department of Nursing, The Fifth Affiliated Hospital, Sun Yat-sen University, Zhuhai, China.

[6]School of Geospatial Engineering and Science, Southern Marine Science and Engineering Guangdong Laboratory (Zhuhai), Sun Yat-sen University, Zhuhai, China.

[7]Key Laboratory of Comprehensive Observation of Polar Environment, Ministry of Education, Sun Yat-sen University, Zhuhai, China.

[8]Ministry of Education Key Laboratory of Biodiversity Science and Ecological Engineering, National Observations and Research Station for Wetland Ecosystems of the Yangtze Estuary, Institute of Biodiversity Science and Institute of Eco-Chongming, School of Life Sciences, Fudan University, Shanghai, China.

[9]Sydney Institute for Infectious Diseases, School of Medical Sciences, The University of Sydney, Sydney, NSW 2006, Australia.

[10]College of Life Sciences, Zhejiang University, Hangzhou, China.

[11]School of life science, Guangzhou University, Guangzhou, China.

[12]School of Public Health (Shenzhen), Shenzhen campus of Sun Yat-sen University, Sun Yat-sen University, Shenzhen, China.

[13]Guangzhou National Laboratory, Guangzhou International Bio-Island, Guangzhou, China.

[14]Department of Infectious Diseases and Public Health, Jockey Club College of Veterinary Medicine and Life Sciences, City University of Hong Kong, Hong Kong, China.

[15]Laboratory of Data Discovery for Health Limited, Hong Kong SAR, China.

[16]These authors contributed equally

[a]Corresponding authors:

Zhao-Rong Li (zhaorong.lzr@alibaba-inc.com)

Mang Shi (shim23@mail.sysu.edu.cn)

Abstract

RNA viruses are diverse and abundant components of global ecosystems. The metagenomic identification of RNA viruses is currently limited to those that exhibit sequence similarity to known viruses. Consequently, the detection of highly divergent viruses with poor sequence similarity to known viruses remains a challenging task. We developed a deep learning algorithm, termed LucaProt, to identify highly divergent RNA-dependent RNA polymerase (RdRP) sequences in 10,487 metatranscriptomes from diverse global ecosystems. LucaProt integrates both sequence and structural information to accurately and efficiently detect RdRP sequences. With this approach we identified 161,979 putative RNA virus species and 180 RNA virus supergroups, among which only 21 contained members of phyla or classes currently defined by the International Committee on Taxonomy of Viruses, and includes many groups that were either undescribed or poorly characterized in previous studies. The newly identified RNA viruses were present in diverse ecological settings, including the air, hot springs and hydrothermal vents, and both virus diversity and abundance varied substantially among ecosystems. We also identified the longest RNA virus genome (nido-like virus) documented to date, at 47,250 nucleotides. This study marks the beginning of a new era of virus discovery, providing computational tools that will help expand our understanding of the global RNA virosphere and of virus evolution.

Key Words:

RNA virus; virome; virus discovery; artificial intelligence; metatranscriptomes; phylogeny

Introduction

RNA viruses infect a diverse array of host species. Despite their omnipresence, the pivotal role of RNA viruses as major constituents of global ecosystems has only recently garnered recognition due to systematic and large-scale virus discovery initiatives performed in animals[1,2], plants[3], fungi[4], aquatic environments[5], marine[6], soil environments[7], and planetary metatranscirptomes[8]. A common characteristic of these studies is their reliance on the analysis of RNA-dependent RNA polymerase (RdRP) sequences, a canonical component of RNA virus genomes. Collectively, these studies have led to the identification of tens of thousands of novel virus species, resulting in at least a 10-fold expansion of the virosphere and the inclusion of new phylum-level virus groups such as the "*Taraviricota*" (i.e., "quenyaviruses")[6,9]. Similarly, the data mining of over 10^9 meta-transcriptomic contigs from diverse ecosystems has revealed several divergent clades of RNA bacteriophage[10], while recent meta-transcriptomic studies have led to a remarkable 5-fold expansion in the diversity of viroid-like circular RNAs[11]. Despite such progress in uncovering RNA virus diversity through ecological sampling and sequencing, it is probable that more divergent groups of RNA viruses remain to be discovered[9,12]. This is in part because the BLAST-based sequence similarity searching approaches used to discover new RNA virus sequences have limitations in detecting highly divergent RdRPs[13], while profile alignment-based approaches (i.e., HMM) necessarily entail a high false positive rate when detecting more divergent RNA virus sequences[14,15]. The development of innovative strategies is therefore imperative to efficiently uncover the full spectrum of RNA virus diversity.

Over the past decade, artificial intelligence (AI) related approaches, especially deep learning algorithms, have had a huge impact on various research fields in the life sciences, such as molecular docking, compound screening and interaction, protein structure prediction and functional annotation, and the modelling of infectious diseases[16-21]. Such progress can be attributed to the advantages of deep learning algorithms over classic bioinformatic approaches, including enhanced accuracy, superior performance, reduced reliance on feature engineering, flexible model architectures, and self-learning capabilities[22,23]. Recently, deep learning approaches, such as CHEER, VirHunter, Virtifier and RNN-VirSeeker have been developed and applied for identification of viruses from genomic and metagenomic data[24-27].

However, many of these approaches rely solely on nucleotide sequence information without incorporating protein sequence or structural information, making them less effective in identifying highly divergent RNA viruses. Transformer architecture was recently developed and applied to sequence-based protein function predictions, outperforming the convolutional neural networks (CNN) and recurrent neural network (RNN) algorithms implemented in previous virus discovery algorithms[28-30]. Consequently, transformer architecture can be leveraged to design a better tool for uncovering highly divergent RNA viruses. Herein, we show how AI can accurately can efficiently detect RNA viruses that exhibit sequence divergence beyond the detection capability of traditional similarity-based methods, in doing so unveiling a concealed realm of viral diversity.

Results

Deep learning reveals the dark matter of the RNA virosphere

We performed systematic searches to expand the diversity of RNA viruses in a variety of ecological systems sampled on a global scale (Extended Data Fig. 1, Supplementary Table 1 and 2). Accordingly, a total of 10,487 metatranscriptomes (51 Tb of total sequencing data) were assembled, which resulted in more than 1,368 million contigs and 872 million predicted proteins. Based on this data set, potential viral RdRPs were revealed and cross-validated using two different strategies (Fig. 1, Extended Data Fig. 2-4). The major AI algorithm used here, denoted “LucaProt”, is a deep learning, transformer-based model established based on sequence and structural features of 5,979 well-characterized RdRPs and 229,434 non-RdRPs. LucaProt demonstrated exceptional accuracy (0.014% false positives) and specificity (1.72% false negatives) when evaluated on the test data set (Fig. 1b). Independently to LucaProt, we applied a more conventional approach (i.e., “ClstrSearch”) that clustered all proteins based on their sequence homology and then used BLAST or HMM models to identify resemblance to viral RdRPs or non-RdRP proteins. ClstrSearch is distinguished from previous BLAST or HMM-based approaches because it queries on protein clusters (i.e., alignments) instead of individual sequences, resulting in a great reduction in both false positive and negative rates of virus identification.

By merging the results of the two search strategies we discovered 513,134 RNA viral

contigs, representing 161,979 putative viral species (i.e., >90% RdRP identity), and 180 RNA viral supergroups that are comparable to existing viral classes and phyla as defined by the International Committee on Taxonomy of Viruses (ICTV) (Fig. 1, Supplementary Table 3 and see Methods). Among these, 512,690 viral contigs (99.9% of total contigs) and 157 supergroups (87.2%) were identified by both "LucaProt" and "ClstrSearch", whereas an additional 444 contigs and 23 supergroups were only identified by "LucaProt". Both strategies out-performed previous RNA virus discovery studies of aquatic environments[5], ocean[6], soil[7], and more diverse ecosystems[8,10,31] (Fig. 1c). Indeed, "LucaProt" was able to identify 93.48% to ~100.0% of RdRPs discovered in these previous studies, even though none were used in either training or testing of the models (Extended Data Fig. 5g).

To demonstrate the sensitivity and specificity of LucaProt, we benchmarked it to four other virus discovery tools, utilizing the same data set and RdRP database (see Methods). Notably, LucaProt exhibited the highest recall rate (i.e., proportion of correctly predicted true positives, Extended Data Fig. 5a), while maintaining a relatively low false positive rate (i.e., proportion of incorrectly predicted true negatives as positives, Extended Data Fig. 5c) and false discovery rate (i.e., proportion of all predictions incorrectly identified as positives, Extended Data Fig. 5d), along with reasonable computational efficiency (Extended Data Fig. 5h). Specifically, in comparisons, HMMscan and Palmscan showed the highest precision rates (99.80%, 99.46%) but the lowest recall rates (65.00%, 65.03%), while Diamond blast and HH-suite exhibited both the highest false positive rate (0.195%, 0.573%) and false discovery rate (7.07%, 21.72%, Extended Data Fig. 5b-5d). When considering all RdRPs identified in this study, LucaProt exhibited the most comprehensive discovery (98.22%), whereas the other four virus discovery tools could only identify a portion of the entire collection (76.82%-87.81%, Extended Data Fig. 5e). Importantly, our benchmarking results demonstrate that the other virus discovery tools only identified a minority (<42%) of the new viruses identified by LucaProt. Hence, most of the new viruses described here (see below) cannot be identified through homology or HMM-based approaches (Extended Data Fig. 5f). Additionally, LucaProt successfully recalled over 98% of the virus RdRPs identified by six other published studies[5-8,10,31]. The exception was the study of Olendraite *et al.*,[31] for which LucaProt had a slightly lower recall rate (93.48%, Extended Data Fig. 5g). Manual inspection

and sequence alignment revealed that the sequences from Olendraite *et al.*,[31] not detected by LucaProt lacked the core RdRP domain region. We also performed an automatic comparison with a uniform definition (identity ≥0.9, coverage ≥0.3) among this study and the six previous studies. This revealed a total of 70,458 putative novel viral species and 60 RNA virus supergroups that were newly identified by LucaProt (Fig. 1c, Extended Data Fig. 6). A novel supergroup was defined as one that did not contain previously described RdRPs, including classified reference sequences and those from previous metagenomic studies[5-8,10,31].

All the RNA viral sequences identified in this study were classified into clusters and supergroups, with the latter were subsequently compared to viral classes and phyla as defined by ICTV[32,33]. Among the supergroups identified in this study, only 21 contained those from viral phyla/classes currently classified by ICTV, resulting in an 8.6-fold expansion of RNA virus diversity at supergroup level compared to the latest ICTV report[32], and a 1.5-fold expansion of all RNA viruses described so far[5-8,10,31]. This expansion encompasses both existing viral supergroups as well as the discovery of 60 novel and highly divergent supergroups that have largely been overlooked by previous efforts of RNA virus discovery (Fig. 1d). The virus supergroups identified here were mostly comparable to the existing classification system at the phylum (e.g., phylum *Lenarviricota* in the case of the Narna-Levi supergroup) or class level (e.g., the *Stelpaviricetes, Alsuviricetes, Flasuviricetes* classes for the Astro-Poty, Hepe-Virga, Flavi supergroups) (Supplementary Table 4). However, the virus supergroups defined here exhibited much greater phylogenetic diversity than those defined previously.

Verification and confirmation of newly identified viral supergroups

That the 180 viral supergroups identified here represented RNA rather than DNA sequences was substantiated through multiple lines of evidence. At the sequence level, two criteria were employed to establish a viral supergroup: the absence of homology to cellular proteins and the presence of key RdRP motifs (Fig. 2a, Extended Data Fig. 7). Moreover, the majority (60/180) of the newly identified supergroups, including those from LucaProt, shared varying degrees of sequence similarity with existing reference RdRPs (i.e., BLAST *e*-value ≤1*E*-3

and/or had HMM model score ≥10) (Fig. 2a, Extended Data Fig. 8, Supplementary Table 5). To validate computational predictions, simultaneous DNA and RNA extraction and sequencing were performed on the 50 environmental samples collected in this study to examine the presence of the 115 viral supergroups identified in these samples (Supplementary Table 6). This revealed that only RNA sequencing reads mapped to contigs associated with viral RdRPs, whereas both RNA and DNA sequencing reads mapped to contigs linked with DNA viruses, reverse-transcriptase (RT), and cellular organisms (Fig. 2b, Extended Data Fig. 9 and 10). Seventeen of 115 viral supergroups were further confirmed by a more sensitive RT-PCR approach; this revealed an absence of sequences encoding viral RdRP in the DNA extractions, suggesting that these viral supergroups are *bona fide* RNA organisms (Fig. 2c, Extended Data Fig. 9b). Finally, a 3D alignment was used to compare the newly identified viral RdRPs with known viral RdRPs, eukaryotic RdRPs, eukaryotic DdRPs and RTs to determine their degree of structural similarity (Fig. 2d). The new RdRP supergroups from LucaProt contained at least three signature components of viral RdRP structures that resulted in significantly higher structural similarity to known RNA virus proteins (average structure similarity = 3.0) than their cellular counterparts (average structure similarity = 5.8).

Genomic structures reveal modularity and flexibility within the RNA virosphere

We next analyzed the composition and structure of the putative RNA virus genomes identified in this study. The length of the RdRP-encoding genomes or genome segments differed markedly within and between viral supergroups, although most were centered around 2,131 nt (Fig. 3). Notably, our data set contained exceptionally long RNA virus genomes identified from soil that belonged to the Nido-like supergroup. The length of one of these genomes, at 47.3 kb, surpassed *Planarian secretory cell nidovirus* (41.2 kb)[34] as the longest RNA virus genome identified to date (Fig. 3c, Extended Data Fig. 11, Supplementary Table 7). Interestingly, one more ORF was identified between the 5' end of the genome and the ORF encoding RdRP, although the function of the encoded protein is unclear due to a lack of sequence homology. In addition to the RdRP, we characterized the remaining proteins encoded by the newly identified virus genomes. While most of these predicted proteins had

no homologs in the existing databases, we identified some that were related to structural (i.e., coat, capsid, glycoprotein and envelope proteins, amongst others) and non-structural (i.e., helicase, protease, methyltransferase, movement protein, immune or host-related regulatory proteins, amongst others) proteins from known viruses (Fig. 3d, Extended Data Fig. 12). Importantly, the presence of these additional virus proteins in newly identified supergroups provided further evidence that these were from *bona fide* RNA viruses. Furthermore, the presence of phage-related proteins (i.e., phage coat, phage mat-A and phage integrase) indicates that some of the viruses are likely infecting prokaryotic hosts, although further validation is required. In addition, the occurrence of these proteins was incongruent with virus phylogenetic history as inferred from RdRP sequences (Fig. 3e), indicative of a modular-like configuration of RNA virus genomes.

Expanded phylogenetic diversity of RNA viruses

The expansion of the RNA virosphere at the level of virus species – a 55.9-fold increase (251,846/4,502) from those defined by ICTV and a 1.4-fold increase (251,846/181,388) from all previously described RdRP sequences – is also evident in both the enlargement of known virus groups (e.g., phyla, orders, families) and the identification of entirely novel groups (Fig. 4). Many of viruses discovered here formed clades that were distinct from the lineages observed in previously described virus supergroups (Fig. 4). Interestingly, several groups previously represented by only a limited number of genomes – namely the Astro-Poty, Hypo, Yan and Cysto clades – experienced major expansion to encompass larger viral clusters with greater phylogenetic diversity (Fig. 4). Several newly identified supergroups also had high levels of phylogenetic diversity, including SG023 (1,232 virus species), SG025 (466 virus species), and SG27 (475 virus species), suggesting that more highly divergent RNA viruses will be discovered in environmental samples. Following our analysis, the supergroups with the greatest number of virus species here were the Narna-Levi (58,063 virus species), Picorna-Calici (19,970 virus species), and Tombus-Nada (15,520 virus species).

Ecological structure of the global RNA virome

Our analysis revealed the ubiquitous presence of RNA viruses across diverse ecosystem

subtypes (32 categories) and in 1,612 locations globally. Despite repeated efforts to uncover the diversity of RNA viruses from such ecological samples[5-8,10,31], 33.3% of viral groups detected by LucaProt were not described previously (Fig. 5a). Indeed, that the rate of RNA virus discovery did not plateau (Fig. 5b) underscores the vast untapped potential for exploring the global diversity of RNA viruses, with a particularly notable surge in soil environments as previously documented[10].

We compared alpha diversity (measured by the Shannon index) and abundance levels (measured by the number of reads per million total non-rRNA reads, i.e., RPM) of the RNA virome among diverse ecosystem subtypes to help identify any ecological patterns among them (Fig. 5c and 5d, Supplementary Table 8). In general, average alpha diversity was highest in leaf litter, wetland, freshwater and wastewater environments, while virus abundance reached its peak in Antarctic sediment, marine sediment, and freshwater ecosystem subtypes, with average RPMs ranging from 18424.6 to 46685.5 (Fig. 5c). In contrast, the lowest average diversity and abundance were observed in halite and subsurface environments (Fig. 5c), which as expected due to their low biomass (i.e., host cells). For extreme ecological subtypes such as hot springs and hydrothermal vents, the associated RNA viruses were characterized by low diversity but moderate abundance (1528.9 ~ 3726.9 average RPM) (Fig. 5c). It is also noteworthy that the new viral supergroups established in this study were predominantly found in aquatic and sediment samples, with only a few occurrences in vertebrate and invertebrate animal samples (Fig. 5c).

Our results further revealed the prevalence and abundance levels of particular viral groups across different ecosystem subtypes (Fig. 5d). The majority (85.9%) of the viruses discovered here only occurred in a single ecosystem subtype. Others, however, could be considered ecological generalists. For example, members of the Narna-Levi, Partiti-Picobirna and Picorna supergroups as well as Tombus-Noda supergroup were among the prominent generalist RNA viruses, found in more than 32 ecosystem subtypes (Extended Data Fig. 13). Finally, we identified “marker” virus species, which exhibited high prevalence and abundance exclusively within specific ecosystem subtypes (Fig. 5d), consistent with previous reports[6,10]. Among these, *Partiti-Picobirna sp.*1991 and *Partiti-Picobirna sp.* 5447 were associated with hot springs while *Tombus-Noda sp.* 2765 and *Supergroup 026 sp.* 2205 were

associated with hydrothermal vents, compatible with the idea that these viruses help modulate host metabolism related to high temperatures[35,36]. However, since the data sets analysed here were generated by different laboratories employing distinct sample processing, library preparation and sequencing procedures, the comparisons of viral diversity and abundance among different ecosystem subtypes were necessarily subject to systemic biases.

Discussion

The accurate identification of highly divergent RNA viruses remains a major challenge, impeding a comprehensive understanding of the genetic diversity in the RNA virosphere and hindering advancements in RNA virus evolution and ecology[37,38]. Indeed, the conventional approach for RNA virus discovery has heavily relied on sequence similarity comparisons and hence the completeness of sequence databases[15,37]. To address these issues, we developed a data-driven deep learning model (i.e., LucaProt) that also outperforms conventional methods in accuracy, efficiency, and most importantly, the breadth of virus diversity detected. Importantly, LucaProt not only incorporated sequence data but also integrated structural information, which is crucial for accurate prediction of protein function, especially in the case of the RdRP[39]. Without implementing the structural model, our model had only 41.8% and 94.9% specificity and accuracy, respectively, on the test data set, and could only detect 44.5% of the predicted RdRP proteins. In addition, the advanced transformer architecture incorporated into our model allowed the parallel processing of longer amino acid sequences[40,41], which can better capture the relationships between residues from distant parts of sequence space, and outperforming the CNN and/or RNN encoders implemented in the CHEER, VirHunter, Virtifier and RNN-VirSeeker RNA virus discovery tools (Extended Data Fig. 14)[24-27]. Collectively, we have established an AI framework for large-scale RNA virus discovery which can be easily extended to the accurate description of any biological “dark matter”.

Despite the large expansion in RNA virus diversity documented here, major gaps remain in our understanding of the evolution and ecology of the newly discovered viruses. In particular, nothing is known about the hosts of the viruses identified, including that (a nido-like virus) with the longest virus genome identified to date. It is possible that the viral clades

and supergroups identified here were largely associated with diverse microbial eukaryotic hosts, given that the majority of current known RNA viruses infect eukaryotes[42,43] and microbial eukaryotes exist in great abundance and diversity in natural environments[44,45]. Nevertheless, it is also likely that a substantial proportion of the novel viruses discovered are associated with bacterial (and perhaps archaeal) hosts[46-48]. Indeed, mounting evidence[10,49] strongly supports the notion that more groups of RNA viruses than currently documented are associated with bacteria. Indeed, the presence of RNA bacteriophage in multiple RNA viral supergroups underlines the evolutionary connection between RNA viruses from bacterial and eukaryotic hosts. If viewed through the lens of virus-host co-divergence[1,2,50], such a link suggests that the evolutionary history of RNA viruses is at least as long, if not longer, than that of the cellular organisms.

Methods

Samples and data sets

This study comprised RNA virus discovery through the meta-transcriptomic analysis of 10,487 samples. The majority of these samples (n = 10,437) were mined from the NCBI Sequence Read Archive (SRA) database (https://www.ncbi.nlm.nih.gov/sra) between January 16 - August 14, 2020. We targeted samples collected from a wide range of environmental types globally (Extended Data Fig. 1), including: aquatic (such as marine, riverine and lake water), soil (such as sediment, sludge and wetland), host-related (such as biofilm, wood decay, and rhizosphere), and extreme environmental samples (such as hydrothermal vent, hypersaline lake and salt marsh), that were subject to high quality meta-transcriptomics sequencing to ensure the generation of ≥50 MB total RNA Q20 sequencing data. In addition, 50 data sets were generated in this study (see below), all of which were subject to high-quality short-read sequencing utilizing Illumina sequencing platforms. The raw sequencing data output ranged from 35.1 to 204.1 Gbp, and no enrichment for microbial organisms was performed during sample processing or library preparations. For highly abundant environmental types, such as “soil” and “marine”, representative samples were selected to include as many projects (i.e., independent studies), geographic locations and ecological

niches as possible.

In addition to data mined from the SRA database, we generated 50 meta-transcriptomic data sets from Antarctica and China. The sample types included marine (n = 5), freshwater (n = 12), soil (n = 19), and sediment (n = 14), of which nine sediment samples were collected at the Ross Sea station in Antarctica between January and February 2022, with the others from Zhejiang, Guangdong, Hubei and Heilongjiang provinces of China collected between August and October 2022. For each of these samples, DNA and RNA were simultaneously extracted: the soil and sediment samples were extracted using the RNeasy® PowerSoil® Total RNA Kit and RNeasy® PowerSoil® DNA Elution Kit (QIAGEN, Germany), while the marine and freshwater samples were extracted using the DNeasy® PowerWater® Kit and RNeasy® PowerWater® Kit (QIAGEN, Germany). The extracted nucleic acid was then subject to library construction using NEBNext Ultra RNA Library Prep Kit and NEB Next Ultra DNA Library Prep Kit (LTD.NEB, China) for RNA and DNA samples, respectively. Paired-end (150 bp) sequencing of these libraries was performed using the Illumina NovaSeq 6000 platform (Illumina, the United States).

For all 10,487 data sets used this study, reads were assembled *de novo* into contigs using MEGAHIT v1.2.8[51] with default parameters. Potential encoded proteins were predicted from contigs using ORFfinder v0.4.3 (https://ftp.ncbi.nlm.nih.gov/genomes/TOOLS/ORFfinder/linux-i64/; parameters, -g 1, -s 2).

Identification of RNA viruses based on deep learning

We developed a new deep learning, transformer-based model, termed "Deep Sequential and Structural Information Fusion Network for Protein Function Prediction" (i.e., LucaProt), that effectively integrated protein sequence composition and structure information to facilitate the precise identification of viral RdRPs. The model includes five modules: Input, Tokenizer, Encoder, Pooling, and Output (Extended Data Fig. 2e).

Input Layer: The amino acid sequence serves as the input in our model.

Tokenizer Layer: This module consists of two components. A corpus consisting of 5,979 viral RdRPs and 150,000 non-virus RDRP sequences was constructed. The BPE (Byte Pair Encoder) algorithm[52] of the Subword-NMT tool (https://github.com/rsennrich/subword-

nmt) was applied to automatically generate about 20,000 frequently occurring short subsequences of variable length as the "vocabulary", which was in turn used to generate words from the input sequence, treated co-occurring amino acids as a whole (namely, "words"). The other component broke down each protein sequence into a combination of single amino acid characters, subsequently utilized when modeling protein structure.

Encoder Layer: This module processes two types of input into sequence and structural representation matrices. In the case of subword processing, an advanced Transformer-Encoder was applied to obtain the sequence representation matrix, while for structural processing, two strategies were considered to calculate the protein structure representation matrix. The first strategy used a structural model (such as RoseTTAFold[53], AlpahFold[19], and ESMFold[54]) to predict 3D protein structure, calculated from the distance between the C-atoms (Alpha-C or Beta-C) of all amino acid residues into a Contact Map matrix, and then applied Graph Convolutional Network (GCN)[55] to encode the Contact Map into a representation matrix. The second approach was to directly use the intermediate matrix from the structural model and employ it as the structural representation matrix. This method not only addressed the issue of the insufficient number of 3D structures observed in experiments, but also circumvented the need to perform the encoder, resulting in a cost-effective approach suitable for large-scale implementation such as this study. We therefore adopted the second strategy here and used the faster ESMFold[54] for structural representation.

Pooling Layer: The previous module obtained the sequence and structure representation matrices. A value-level attention pooling (VLAP) approach[56] was then used to transform these two matrices into two vectors.

Output Layer: A concatenation operator was used to join the two vectors generated by the pooling layer. A fully connected layer and the sigmoid function (Extended Data Fig. 4a) were then used to generate the probability values between 0.0 and 1.0 as a measure of confidence, and a threshold of 0.5 was used to determine whether it represents viral RNA (Extended Data Fig. 4).

Model Building: We constructed a data set with 235,413 samples for model building. This included 5,979 positive samples of known viral RdRPs (i.e., the well-curated RdRP database described below) and 229,434 randomly selected negative samples of confirmed

non-virus RdRPs (as the positive samples account for a very small portion of the total data, we constructed the training data set using the conventional 1:40 ratio of positive to negative data). The non-virus RdRP representative sequences contained proteins from eukaryotic RNA-dependent RNA polymerase (Eu RdRP, n = 2,233), eukaryotic DNA-dependent RNA polymerase (Eu DdRP, n = 1,184), reverse transcriptase (RT, n = 48,490), proteins obtained from DNA viruses (n = 1,533), non-RdRP proteins obtained from RNA viruses (n = 1,574), as well as an array of cellular proteins from different functional categories (n = 174,420). We randomly divided the data set into training, validation, and testing sets with a ratio of 8.5:1:1, which were used for model fitting, model finalization (based on the best F1-score training iteration), and performance reporting (including accuracy, precision, recall, F1-score, and Area under the ROC Curve (AUC)), respectively (Extended Data Fig. 4).

Finally, LucaProt successfully identified 792,436 putative RdRP signatures from a data set of 144.6 million proteins derived from the 10,487 metatranscriptomes. These results were first compared with the RdRPs identified based on ClstrSearch (see below). RdRPs that were identified by deep learning algorithms (LucaProt) were either incorporated into RNA virus supergroups using the Diamond blastp program v0.9.25.126[57] with an *e*-value threshold of 1*E*-3, or, if they remained unclassified, were subjected to clustering, merging, and manual alignment inspection. The presence of key viral RdRP motifs, including motif A [DxxxxD], motif B [(S/T)Gxxx(T/G)xxxN], and motif C [(S/G/N/A)D(D/N)], was rigorously examined for all supergroups through multiple sequence alignments.

Identification of RNA viruses based on homologous clustered proteins

Another strategy for the identification of RNA viruses at cluster level involved utilization of homologous clustered proteins (Extended Data Fig. 2a). A total of 871.8 million amino acid sequences predicted by ORFfinder (see Samples and data sets) were compared against a well-curated RdRP database (N = 5,979) that contained only those derived from reference RNA virus genomes downloaded from the NCBI GenBank database and their close relatives in vertebrate and invertebrate hosts[1,2]. The comparisons were performed using the Diamond blastp program v0.9.25.126[57], with the *e*-value threshold set at 1*E*+5 to identify more divergent RdRP sequences (Extended Data Fig. 2a, Extended Data Fig. 3a). This process

resulted in 75.3 million hits which were further subjected to homology-based and multi-step clustering (three iterations with 90%, 60%, and 20% amino acid identity, respectively) using CD-HIT v4.8.1 (https://github.com/weizhongli/cdhit); this process resulted in 3,805,584 clusters. False positives and hits to known RdRP sequences were removed by comparison against the NCBI non-redundant (nr) protein database (version 2023.01.09), the NCBI RefSeq protein database (version 2023.03.26) and the virus RdRP database using Diamond blastp program v0.9.25.126[57], with the *e*-value threshold set at 1*E*-3 (Extended Data Fig. 2b). The remaining unknown protein clusters were subject to viral RdRP domain search using hidden Markov models (HMMs) built from a manually reviewed profile of known RdRP clusters using hmmscan v3.3.2 (e =10, hits ≥1)[58]. The RdRP sequences of each known viral group were aligned using Mafft v7.475[59], employing the L-INS-I algorithm. HMM profiles were then built from each MSA using hmmbuild v3.3.2 with standard parameters[58]. Clusters that contained more than one hmmscan hit were subsequently aligned and inspected for the presence of well-conserved RdRP core sequences referred to as the "palmprint", which were delineated by three essential motifs (A, B and C) that collectively constitute the catalytic core within the RdRP structure[60]. As a result of our rigorous screening and checking steps, a total of 713 potential RdRP clusters were retained.

To further expand the RdRP collection based on the viruses newly discovered here, we updated the RdRP protein database with the 713 RdRP clusters newly identified here and used it to detect additional RdRP sequences from the original 144.6 million amino acid sequences using Diamond blastp v0.9.25.126[57] with an *e*-value threshold of 1*E*-3. The newly detected RdRPs were again incorporated into the RdRP database for another round of virus detection. This procedure was repeated ten times. The resulting hits (21,747,015 in total) were compared to the homology-based clusters for the removal of false positives and virus classification. The false positives and non-RdRP viral proteins which did not contain RdRP domain were removed by a HMMs-based search using an updated HMMs profile derived from each RdRP cluster and built as described above. The cut-off was set to score ≥40 and the aligned fraction to ≥0.4 to remove false positives and partial virus RdRPs (Extended Data Fig. 2c, Extended Data Fig. 3b).

Finally, the expanded clusters were merged into virus supergroups using a hierarchical method employing the Girvan-Newman algorithm[61], with the edge betweenness determined based on median *e*-value threshold of 1*E*-3 for each pair of clusters (Extended Data Fig. 2d, Extended Data Fig. 3c and 3d). Briefly, the merging of clusters employed the following steps: (i) the betweenness of all edges (median *e*-value between clusters) in the network was calculated; (ii) the edge(s) with the highest betweenness were removed; (iii) the betweenness of all edges affected by the removal was recalculated; (iv) steps ii and iii were repeated until no edges remained. All processes related to merging were performed using igraph package v1.3.5[62] implemented in *R*.

Benchmarking LucaProt and comparisons to Diamond, HMMscan, HH-suite and Palmscan

We assessed the performance of LucaProt relative to other four virus discovery tools: Diamond v0.9.25.126[57] (with *e*-value threshold of 1*E*-3), HMMscan v3.3.2[58] (with score ≥30 and aligned fraction ≥0.7 as referenced from Zayed *et al.*,[6]), HH-suite v3.3.0[63] (with default parameters), and Palmscan[60] (with default parameters). This comparison included recall, precision, false positive and false discovery based on the 2.05 million amino acid sequences (length ≥300aa) from contigs assembled from the 50 samples collected in this study, employing a unified criterion and utilizing the same well-curated RdRP database (N = 5,979).

True positive (TP) virus hits were determined by simultaneously identifying the true viral RdRPs using both aforementioned tools, and subsequently confirming through comparison against the nr protein database (version 2023.01.09) using the Diamond blastp program v0.9.25.126[57] with an *e*-value threshold of 1*E*-3. True negative (TN) hits were defined as the subset of sequences recognized as cellular proteins and/or with reads from DNA sequencing results mapping against the corresponding contigs, among a data set comprising 2.05 million sequences. A data set of 1,029,342 true negative sequences was obtained through both approaches. Recall was calculated as the ratio of correctly predicted true positives to the total number of true positives (recall rate = TP * 100% / total TP). Precision was calculated as the proportion of all predictions that were correctly predicted as true positives (precision rate =

TP*100% / total predictions). The false positive (FP) rate was calculated as the proportion of all true negatives that were incorrectly predicted as positives (false positive rate = FP*100%/total TN). False discovery (FD) was calculated as the proportion of all predictions that were incorrectly predicted as positives (false discovery rate = FP*100%/total predictions).

To compare the computational time required by LucaProt and other conventional methods (i.e., HH-suit, Diamond, HMMscan, and Palmscan) for RdRP inference, we categorized the test data sets into six groups based on protein sequence length, including [300, 500), [500, 800), [800, 1000), [1000, 3000), [3000, 5000), and [5000,). Each group comprised 50 randomly sampled viral RdRPs and 50 non-viral sequences. A single sequence was used as input to calculate the average time on the same machine of the Alibaba Cloud Elastic Compute Service (ECS). The CPU execution environment comprised 96 cores (vCPU) and 192 GiB memory (instance type: ecs.c8i.24xlarge). For LucaProt, we also recorded the time spent on a single NVIDIA A100 80G GPU (instance type: ecs.gn7e-c16g1.16xlarge).

All scripts of the tools used, containing exact parameters and the database utilized in the benchmarking, are available at http://47.93.21.181/Benchmark/LucaProtApp/.

Virus verification

To determine whether the newly discovered viral RdRPs belonged to RNA viruses rather than organisms with DNA genomes, we performed three verification experiments. First, the RdRPs were compared against the NCBI nr protein database (version 2023.01.09) using Diamond blastp v0.9.25.126[57], with the *e*-value threshold set to 1*E*-3. Proteins with homology to cellular proteins were removed. In addition, we confirmed the presence of the key RdRP motifs (i.e., the A, B and C motifs) through alignment to all RNA supergroups that possessed these motifs (i.e. motif A [DxxxxD], motif B [(S/T)Gxxx(T/G)xxxN], motifC [(S/G/N)DD]). Sequences that failed to cover these motifs were removed.

Second, to ensure the accuracy of DNA read mapping, an initial quality control step was performed on viral contigs using bbduk.sh (https://sourceforge.net/projects/bbmap/). As well as the simultaneous RNA and DNA extraction and sequencing on 50 environmental samples collected in this study, we searched for those published RNA studies that performed both RNA and DNA sequencing on the samples. To confirm the detected contigs represented *bona*

fide RNA viral genomes, reads from the DNA sequencing data were mapped against the viral contigs using Bowtie2 v2.4.2[64] with the "end-to-end" setting.

Finally, RT-PCR assays were conducted to validate the presence of RNA organisms within the viral supergroups identified. Two pairs of validation primers were designed for each of the representative RdRP sequences from 17 of the 115 RNA viral supergroups involved in 50 samples collected in this study. These comprised: (i) two known supergroups defined by ICTV: Astro-Poty, Bunya-Arena; (ii) ten unclassified supergroups derived from this study and other studies: Supergroup022, Supergroup034, Supergroup038, Supergroup053, Supergroup055, Supergroup063, Supergroup086, Supergroup104, Supergroup124, Supergroup180; and (iii) five new supergroups identified here: Supergroup102, Supergroup167, Supergroup175, Supergroup184, Supergroup187. We also included gene sequences from two DNA virus families (i.e., the *Podoviridae* and *Siphoviridae*), and RT sequences identified in this study, with a product length of 300-550 bp. For each of these samples, both the reverse-transcribed RNA and the matching DNA underwent simultaneous PCR amplification, and the amplification products were subject to electrophoresis using a 1% agarose gel with GelRed dye, which was subsequently visualized under UV.

Structural prediction and comparisons between viral RdRPs and homologous proteins

Three-dimensional structures of newly identified viral RdRPs from diverse RNA viral supergroups were predicted from primary sequences using AlphaFold 2 v2.3[19] and visualized using the PyMol software v2.5.4 (http://www.pymol.org/pymol). The pLDDT (predicted local distance difference test) score was measured for each structure prediction as its per-residue estimate of the prediction confidence on a scale from 0-100. Considering the computational resource limitations with such a large data set, ten representative RdRPs from each supergroup were predicted; more than 52.03% of the structures predicted had accuracy >70%, showing that Alphafold 2 prediction was a relatively reliable source of structural information. The previously resolved or predicted structures of viral RdRPs, eukaryotic RdRPs, eukaryotic DdRPs and RTs were compared using the Super algorithm[65]. Considering that the protein structures have similar molecular weights but substantial conformational variation, the "number of aligned atoms after refinement" option was employed to evaluate the similarity

between each pair of proteins. Subsequently, networkX (https://networkx.org/) was employed to construct a three-dimensional structure diagram using the "edge-weighted spring embedded" approach, with results then mapped as a scatter plot (depicted in the Fig. 2d). Simultaneously, we visualized four viral RdRP domain proteins using PyMol.

Annotation and characterization of virus genomes

Potential open reading frames (ORFs) in virus genomes were predicted in the newly identified virus genomes based on two criteria: (i) the predicted amino acid sequences were longer than 200 amino acids in length, and (ii) they were not completely nested within larger ORFs. The annotation of non-RdRP ORFs was mainly based on comparisons of predicted proteins to hidden Markov models (HMMs) collected from the Pfam database (version 2022.03.23, https://pfam-legacy.xfam.org/) using hmmscan v3.3.2 implemented in HMMER (e = 10, score ≥10)[58]. For the remaining ORFs, the annotation was carried out by blastp comparisons against the nr protein database with an e-value threshold of 1E-3.

Analyses of virome diversity, evolution and ecology

To unveil the diversity of the RNA viruses identified and establish putative novel viral species, we employed a protein sequence identity threshold of 90% for the RdRP domain and a minimum aligned fraction of 30%, as implemented by CD-HIT v4.8.1 (https://github.com/weizhongli/cdhit). Abundance levels were subsequently estimated for every virus species-level taxonomy based on the number of non-rRNA reads per million (RPM) within each sample (i.e., sequencing runs) mapped to the RdRP sequences of all viral species using Bowtie2 v2.4.2[64] with the “end-to-end” setting. The abundance of each putative virus species was estimated as the number of mapped reads per million total non-rRNA reads (RPM) in each library (cat XX.sorted.bam.txt | grep -v "#" | cut -f1,4 | awk -v readcount=`cat XX_non-RNA_readcount` -F "\t" 'BEGIN{OFS="\t"} {print $1,$2,$2/readcount*1000000}' > XX.rpm.txt).

Virus alpha diversity (measured using the Shannon index) and overall abundance were subsequently estimated and compared across different geographic locations and ecosystem

subtypes, namely; soil, marine, freshwater, wetland, hot spring, salt marsh, and other subtypes. “Marker virus species”, that were greatly enriched in certain ecosystem subtypes, were also identified based on the virus mapping results. The marker virus species were defined as those present only in one ecosystem subtype with RPM ≥1 and coverage (percentage of horizontal covered bases with depth ≥1) ≥20%.

To reveal the diversity and evolutionary relationships of RNA viruses within individual virus supergroups, all previously documented RNA viruses were incorporated into phylogenetic analyses, including RefSeq RdRPs sequences as well as those from previous studies. For supergroups containing more than 500 RdRPs, homology-based clustering using CD-HIT v4.8.1 (https://github.com/weizhongli/cdhit) with a threshold of identity = 0.6 was employed to select representative RdRPs to ensure phylogenetic diversity, while for those containing fewer than 500 RdRPs, all sequences were used in the phylogenetic analysis. The RdRPs of each supergroup were aligned using L-INS-I algorithm implemented in Mafft v7.475[59]. Phylogenetic analyses were performed based on the sequence alignment using a maximum likelihood algorithm, employing the LG model of amino acid substitution, a Subtree Pruning and Regrafting (SPR) branch swapping algorithm, and a Shimodaira–Hasegawa-like procedure as implemented in the Phyml program v3.1[66].

Data availability

The raw sequence reads newly generated in this study are available at the NCBI Sequence Read Archive (SRA) database under the BioProject accession PRJNA956286 and PRJNA956287 (Extended Data Table. 2). All virus sequence data produced in this study are publicly available at http://47.93.21.181/; this includes all RNA virus contigs, RdRP sequences, RdRP proteins, RdRP HMM profiles and phylogenetic tree files. Additionally, this website includes related data sets for model building and validation, and the trained LucaProt model

Code availability

The original source code for ClstrSearch and LucaProt are stored at GitHub repository

(ClstrSearch: https://github.com/alibaba/LucaProt/tree/master/ClstrSearch, LucaProt: https://github.com/alibaba/LucaProt). The link will be available upon acceptance of the paper. Currently, the codes are provided for the review process only. Any additional information required to reanalyze the data reported in this paper is available from the lead contact author upon request.

Acknowledgements

This work was supported by National Natural Science Foundation of China (82341118, 32270160), Shenzhen Science and Technology Program (KQTD20200820145822023 and JCYJ20210324124414040), Natural Science Foundation of Guangdong Province (2022A1515011854), Guangdong Province "Pearl River Talent Plan" Innovation and Entrepreneurship Team Project (2019ZT08Y464), Hong Kong Innovation and Technology Fund (ITF) (MRP/071/20X), and Health and Medical Research Fund (COVID190206). E.C.H. is funded by a National Health and Medical Research Council (Australia) Investigator grant (GNT2017197) and by AIR@InnoHK administered by the Innovation and Technology Commission, Hong Kong Special Administrative Region, China.

We sincerely thank the Computing and Storage teams of Alibaba Cloud Computing Co., Ltd. and Zhejiang laboratory for their contribution of 15 machines with 128 CPUs and 1T RAM of Elastic High-Performance Computing (EHPC), 64 Nvidia A100 Graphics Processing Units (GPUs), and 500TB of Network Attached Storage (NAS) resources.

Author contributions

Conceptualization, X.H., Y.H., E.C.H., Z.-R.L. and M.S.; Methodology, X.H., Y.H., J.-S.E., J.L., Z.-R.L. and M.S.; Investigation, X.H., Y.H., P.F., S.-Q.M., Z.X. and Q.-Y.G.; Writing – Original Draft, X.H., Y.H., E.-C.H. and M.S.; Writing – Review and Editing, All authors. Funding Acquisition, F.-M.H., Y.-L.S., D.-Y.G., Z.-R.L. and M.S.; Resources (sampling), X.H., S.-Q.M., W.-W.C., J.-H.T., G.-Y.X., S.-J.L., Y.-Y.X., Y.-L.Z., F.-M.H., Y.-F.P., Z.-H.Y. and C.H.; Resources (computational), S.Z., Z.-Y.Z. and Z.-R.L.; Supervision, Z.R.L. and M.S.

Competing interests

The authors declare no competing interests.

1 Shi, M. *et al.* Redefining the invertebrate RNA virosphere. *Nature* **540**, 539-543, doi:10.1038/nature20167 (2016).

2 Shi, M. *et al.* The evolutionary history of vertebrate RNA viruses. *Nature* **556**, 197-202, doi:10.1038/s41586-018-0012-7 (2018).

3 Rivarez, M. P. S. *et al.* In-depth study of tomato and weed viromes reveals undiscovered plant virus diversity in an agroecosystem. *Microbiome* **11**, 60, doi:10.1186/s40168-023-01500-6 (2023).

4 Sutela, S. *et al.* The virome from a collection of endomycorrhizal fungi reveals new viral taxa with unprecedented genome organization. *Virus Evolution* **6**, veaa076, doi:10.1093/ve/veaa076 (2020).

5 Wolf, Y. I. *et al.* Doubling of the known set of RNA viruses by metagenomic analysis of an aquatic virome. *Nat Microbiol* **5**, 1262-1270, doi:10.1038/s41564-020-0755-4 (2020).

6 Zayed, A. A. *et al.* Cryptic and abundant marine viruses at the evolutionary origins of Earth's RNA virome. *Science* **376**, 156-162, doi:10.1126/science.abm5847 (2022).

7 Chen, Y. M. *et al.* RNA viromes from terrestrial sites across China expand environmental viral diversity. *Nat Microbiol* **7**, 1312-1323, doi:10.1038/s41564-022-01180-2 (2022).

8 Edgar, R. C. *et al.* Petabase-scale sequence alignment catalyses viral discovery. *Nature* **602**, 142-147, doi:10.1038/s41586-021-04332-2 (2022).

9 Obbard, D. J., Shi, M., Roberts, K. E., Longdon, B. & Dennis, A. B. A new lineage of segmented RNA viruses infecting animals. *Virus Evol* **6**, vez061, doi:10.1093/ve/vez061 (2020).

10 Neri, U. *et al.* Expansion of the global RNA virome reveals diverse clades of bacteriophages. *Cell* **185**, 4023-4037 e4018, doi:10.1016/j.cell.2022.08.023 (2022).

11 Lee, B. D. *et al.* Mining metatranscriptomes reveals a vast world of viroid-like circular RNAs. *Cell* **186**, 646-661 e644, doi:10.1016/j.cell.2022.12.039 (2023).

12 Dominguez-Huerta, G. *et al.* The RNA virosphere: How big and diverse is it? *Environ Microbiol* **25**, 209-215, doi:10.1111/1462-2920.16312 (2023).

13 Rost, B. Twilight zone of protein sequence alignments. *Protein Eng* **12**, 85-94 (1999).

14 Skewes-Cox, P., Sharpton, T. J., Pollard, K. S. & DeRisi, J. L. Profile hidden Markov models for the detection of viruses within metagenomic sequence data. *PLoS One* **9**, e105067, doi:10.1371/journal.pone.0105067 (2014).

15 Chen, J., Guo, M., Wang, X. & Liu, B. A comprehensive review and comparison of different computational methods for protein remote homology detection. *Brief Bioinform* **19**, 231-244, doi:10.1093/bib/bbw108 (2016).

16 McNutt, A. T. *et al.* GNINA 1.0: molecular docking with deep learning. *J Cheminform* **13**, 43, doi:10.1186/s13321-021-00522-2 (2021).

17 Pham, T. H., Qiu, Y., Zeng, J., Xie, L. & Zhang, P. A deep learning framework for high-throughput mechanism-driven phenotype compound screening and its application to COVID-19 drug repurposing. *Nat Mach Intell* **3**, 247-257, doi:10.1038/s42256-020-00285-9 (2021).

18 Du, B.-X. *et al.* Compound-protein interaction prediction by deep learning: Databases, descriptors and models. *Drug Discov Today* **27**, 1350-1366, doi:10.1016/j.drudis.2022.02.023 (2022).

19 Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583-589, doi:10.1038/s41586-021-03819-2 (2021).

20 Gligorijevic, V. *et al.* Structure-based protein function prediction using graph convolutional networks. *Nat Commun* **12**, 3168, doi:10.1038/s41467-021-23303-9 (2021).

21 Xu, L., Magar, R. & Barati Farimani, A. Forecasting COVID-19 new cases using deep learning methods. *Comput Biol Med* **144**, 105342, doi:10.1016/j.compbiomed.2022.105342 (2022).

22 Deng, L. Deep Learning: Methods and Applications. *Foundations and Trends® in Signal Processing* **7**, 197-387, doi:10.1561/2000000039 (2014).

23 Sarker, I. H. Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions. *SN Comput Sci* **2**, 420, doi:10.1007/s42979-021-00815-1 (2021).

24 Shang, J. & Sun, Y. CHEER: HierarCHical taxonomic classification for viral mEtagEnomic data via deep leaRning. *Methods* **189**, 95-103, doi:10.1016/j.ymeth.2020.05.018 (2021).

25 Sukhorukov, G. *et al.* VirHunter: A Deep Learning-Based Method for Detection of Novel RNA Viruses in Plant Sequencing Data. *Front Bioinform* **2**, 867111, doi:10.3389/fbinf.2022.867111 (2022).

26 Miao, Y., Liu, F., Hou, T. & Liu, Y. Virtifier: a deep learning-based identifier for viral sequences from metagenomes. *Bioinformatics* **38**, 1216-1222, doi:10.1093/bioinformatics/btab845 (2022).

27 Liu, F., Miao, Y., Liu, Y. & Hou, T. RNN-VirSeeker: A Deep Learning Method for Identification of Short Viral Sequences From Metagenomes. *IEEE/ACM Trans Comput Biol Bioinform* **19**, 1840-1849, doi:10.1109/TCBB.2020.3044575 (2022).

28 Kabir, A. & Shehu, A. GOProFormer: A Multi-Modal Transformer Method for Gene Ontology Protein Function Prediction. *Biomolecules* **12**, doi:10.3390/biom12111709 (2022).

29 Cao, Y. & Shen, Y. TALE: Transformer-based protein function Annotation with joint sequence–Label Embedding. *Bioinformatics* **37**, 2825-2833, doi:10.1093/bioinformatics/btab198 (2021).

30 Nambiar, A. *et al.* Transforming the Language of Life: Transformer Neural Networks for Protein Prediction Tasks. *bioRvix* (2020).

31 Olendraite, I., Brown, K. & Firth, A. E. Identification of RNA Virus-Derived RdRp Sequences in Publicly Available Transcriptomic Data Sets. *Mol Biol Evol* **40**, doi:10.1093/molbev/msad060 (2023).

32 Wolf, Y. I. *et al.* Origins and Evolution of the Global RNA Virome. *mBio* **9**, doi:10.1128/mBio.02329-18 (2018).

33 Koonin, E. V. *et al.* Global Organization and Proposed Megataxonomy of the Virus World. *Microbiol Mol Biol Rev* **84**, doi:10.1128/MMBR.00061-19 (2020).

34 Saberi, A., Gulyaeva, A. A., Brubacher, J. L., Newmark, P. A. & Gorbalenya, A. E. A planarian nidovirus expands the limits of RNA genome size. *PLoS Pathog* **14**, e1007314, doi:10.1371/journal.ppat.1007314 (2018).

35 Jarett, J. K. *et al.* Insights into the dynamics between viruses and their hosts in a hot spring microbial mat. *ISME J* **14**, 2527-2541, doi:10.1038/s41396-020-0705-4 (2020).

36 Cheng, R. *et al.* Virus diversity and interactions with hosts in deep-sea hydrothermal vents. *Microbiome* **10**, 235, doi:10.1186/s40168-022-01441-6 (2022).

37 Krishnamurthy, S. R. & Wang, D. Origins and challenges of viral dark matter. *Virus Res* **239**, 136-142, doi:10.1016/j.virusres.2017.02.002 (2017).

38 Cobbin, J. C., Charon, J., Harvey, E., Holmes, E. C. & Mahar, J. E. Current challenges to virus discovery by meta-transcriptomics. *Curr Opin Virol* **51**, 48-55, doi:10.1016/j.coviro.2021.09.007 (2021).

39 Monttinen, H. A. M., Ravantti, J. J. & Poranen, M. M. Structure Unveils Relationships between RNA Virus Polymerases. *Viruses* **13**, doi:10.3390/v13020313 (2021).

40 Vaswani, A. *et al.* Attention Is All You Need. *arXiv* (2017).

41 Avsec, Ž. *et al.* Effective gene expression prediction from sequence by integrating long-range interactions. *Nature Methods* **18**, 1196-1203, doi:10.1038/s41592-021-01252-x (2021).

42 Wu, R. *et al.* RNA Viruses Linked to Eukaryotic Hosts in Thawed Permafrost. *mSystems* **7**, e0058222, doi:10.1128/msystems.00582-22 (2022).

43 Charon, J., Murray, S. & Holmes, E. C. Revealing RNA virus diversity and evolution in unicellular algae transcriptomes. *Virus Evolution* **7**, veab070, doi:10.1093/ve/veab070 (2021).

44 Ibarbalz, F. M. *et al.* Global Trends in Marine Plankton Diversity across Kingdoms of Life. *Cell* **179**, 1084-1097 e1021, doi:10.1016/j.cell.2019.10.008 (2019).

45 Kalu, E. I., Reyes-Prieto, A. & Barbeau, M. A. Community dynamics of microbial eukaryotes in intertidal mudflats in the hypertidal Bay of Fundy. *ISME Communications* **3**, 21, doi:10.1038/s43705-023-00226-8 (2023).

46 Bollback, J. P. & Huelsenbeck, J. P. Phylogeny, genome evolution, and host specificity of single-stranded RNA bacteriophage (family Leviviridae). *J Mol Evol* **52**, 117-128 (2001).

47 Poranen, M. M., Mäntynen, S. & Ictv Report, C. ICTV Virus Taxonomy Profile: Cystoviridae. *J Gen Virol* **98**, 2423-2424, doi:10.1099/jgv.0.000928 (2017).

48 Callanan, J. *et al.* RNA Phage Biology in a Metagenomic Era. *Viruses* **10**, doi:10.3390/v10070386 (2018).

49 Gan, T. & Wang, D. Picobirnaviruses encode proteins that are functional bacterial lysins. *Proc Natl Acad Sci U S A* **120**, e2309647120, doi:10.1073/pnas.2309647120 (2023).

50 Sharp, P. M. & Simmonds, P. Evaluating the evidence for virus/host co-evolution. *Curr Opin Virol* **1**, 436-441, doi:10.1016/j.coviro.2011.10.018 (2011).

51 Li, D., Liu, C.-M., Luo, R., Sadakane, K. & Lam, T.-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics (Oxford, England)* **31**, 1674-1676, doi:10.1093/bioinformatics/btv033 (2015).

52 Gage, P. A New Algorithm for Data Compression. *C Users J* **12**, 23–38, doi:10.5555/177910.177914 (1994).

53 Baek, M. *et al.* Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871-876, doi:10.1126/science.abj8754 (2021).

54 Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123-1130, doi:10.1126/science.ade2574 (2023).

55 Thomas N. Kipf, M. W. Semi-Supervised Classification with Graph Convolutional Networks. (2017).

56 Yong He, C. W., Shun Zhang, Nan Li, Zhaorong Li , Zhenyu Zeng. KG-MTT-BERT: Knowledge Graph Enhanced BERT for Multi-Type Medical Text Classification. *arXiv* (2022).

57 Buchfink, B. A.-O. X., Reuter, K. A.-O. & Drost, H. A.-O. X. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat Methods* (2021).

58 Potter, S. C. *et al.* HMMER web server: 2018 update. *Nucleic Acids Res* **46**, W200-W204, doi:10.1093/nar/gky448 (2018).

59 Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution* **30**, 772-780, doi:10.1093/molbev/mst010 (2013).

60 Babaian, A. & Edgar, R. Ribovirus classification by a polymerase barcode sequence. *PeerJ* **10**, e14055, doi:10.7717/peerj.14055 (2022).

61 Girvan, M. & Newman, M. E. J. Community structure in social and biological networks. *Proc Natl Acad Sci U S A* **99**, 7821-7826 (2002).

62 Newman, M. E. J. & Girvan, M. Finding and evaluating community structure in networks. *Phys Rev E Stat Nonlin Soft Matter Phys* **69**, 026113 (2004).

63 Steinegger, M. *et al.* HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics* **20**, 473, doi:10.1186/s12859-019-3019-7 (2019).

64 Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature methods* **9**, 357-359, doi:10.1038/nmeth.1923 (2012).

65 Hasegawa, H. & Holm, L. Advances and pitfalls of protein structural alignment. *Curr Opin Struct Biol* **19**, 341-348, doi:10.1016/j.sbi.2009.04.003 (2009).

66 Guindon, S. & Gascuel, O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic biology* **52**, 696-704 (2003).

67 Jianfeng Liu, H. Q., Ziyi Liu, Sheng Wang, Qiang Zhang, Zhengmin He. A Density-Based Spatial Clustering of Application with Noise Algorithm and its Empirical Research. *Highlights in Science, Engineering and Technology* **7**, 174-179 (2022).

Figure and Table Legends

Fig. 1. The global RNA virosphere. a, RNA virus discovery pipeline. The pathway for sequence homolog-based virus discovery is highlighted in blue on the left, including the clustering, expand and merge steps. The RdRP AI modeling pathway is highlighted in orange on the right, including the modeling, clustering and merge steps. **b**, Number of viral supergroups discovered using two methods (left), and the detection accuracy of LucaProt model (right). **c**, Venn diagram shows the shared putative virus species between available data from Wolf *et al*., Edgar *et al*., Zayed *et al*., Neri *et al*., Chen *et al*., Olendraite *et al*., and this study. The bar graph shows the shared viral supergroups between the seven studies and the unique viral supergroups identified in this study. **d**, Diverse RNA viruses clusters (dark colored small circle) and RNA virus supergroups (light colored large circle). The known viral clusters and supergroups defined by ICTV are denoted in dark grey and light grey, respectively. The viral clusters and supergroups discovered by two strategies are denoted with dark blue and light blue, respectively. The viral clusters and supergroups additionally discovered by LucaProt are denoted with dark orange and light orange, respectively.

Fig. 2. Evaluation of authenticity of RNA viral supergroups. a, Distribution of BLAST median *e*-value, HMM score and mean AI modeling probabilities of RNA virus supergroups, with the primary sequence-identified conserved RdRP motif C of each supergroup shown on the left. For each group selected, the top 20 - 30 supergroups are shown by size for clarity. The known viral supergroups defined by the ICTV show high sensitivity for all three methods and are shaded in grey. The unclassified supergroups and new viruses discovered in this study show declining homology but a relatively stable AI probability. **b**, The proportion of positive libraries and mean RPM (the number of mapped reads per million non-rRNA reads) of representative viral supergroups, DNA viruses, RT, and cell organisms in 50 samples collected in this study. DNA libraries are shown in purple and RNA libraries in yellow, the different groups of RNA viruses and DNA organisms are shown in different colors, and red asterisks refer to those subsequently validated by RT-PCR. **c**, RT-PCR results of the first pairs of validation primers for representative RdRP sequences from 17 RNA viral supergroups,

capsid sequences from two DNA viral families (*Podoviridae* and *Siphoviridae*), and RT sequences. **d**, Three-dimensional (3D) structure homology analysis of representative RdRPs from 180 viral supergroups with Eu DdRPs, Eu RdRPs, and RT. Each point stands for a representative structure. The distance between different points represents structural similarity: the greater the distance, the lower the structural similarity. Four RdRP domain structures of the AI-specific supergroups are displayed with the A, B and C motifs highlighted. Representative RdRP domain structures from all supergroups are available at http://47.93.21.181/SG_predicted_protein_structure_supplementation/.

Fig 3. Genomic features of viral supergroups. **a**, Size (i.e., the number of putative viral species) of all viral supergroups. The number of supergroups included in each group is displayed above the violin chart. Centre lines in the box plots represent the median bounds. **b**, Genome length of viral species of all viral supergroups. The number of viral species included in each group is displayed above the violin chart. Centre lines in the box plots represent the median bounds. **c**, Histogram of the genome size distribution of viral species from known, unclassified and new supergroups. **d**, The distribution of annotated functional protein in each viral supergroup. The supergroups with no other annotated genes than RdRP were not shown. **e**, Genome structure of representatives from six known supergroups, 17 unclassified supergroups and eight new supergroups. Grey stars represent reference virus genomes of known supergroups. Domains not commonly found in RNA viruses are shown in yellow and are labeled above their corresponding positions. At the bottom, scale length in nucleotides. Abbreviations: GOLGA2L5: golgin subfamily A member 2-like protein 5; Pentaxin: pentaxin family; Tme5 EGF: thrombomodulin like fifth domain, EGF-like; Mg trans NIPA: magnesium transporter NIPA; NUDIX, nucleoside diphosphate-X hydrolase; RecX: RecX family; TssO: type VI secretion system, TssO; Securin: securin sister-chromatid separation inhibitor; Rax2: cortical protein marker for cell polarity; Abhydrolase: alpha/beta-hydrolase family; OmdA: bacteriocin-protection, YdeI or OmpD-Associated; Blt1 C: Get5 carboxyl domain; DnaJ: DnaJ domain; Trypan PARP: procyclic acidic repetitive protein (PARP); SAM KSR: kinase suppressor RAS 1; CBD PlyG: PlyG cell wall binding domain; LydB: LydA-holin antagonist; RelB: RelB antitoxin; T2SSE: type II/IV secretion system protein; PARP

regulatory: poly A polymerase regulatory subunit; Pheromone: fungal mating-type pheromone; Y phosphatase2: tyrosine phosphatase family; PseudoU synth: RNA pseudouridylate synthase; Glyco hydro 35: glycosyl hydrolases family 35; TIP: tuftelin interacting protein.

Fig 4. Phylogenetic diversity of 31 RNA viral supergroups. Each phylogenetic tree was estimated using a maximum likelihood method based on an amino acid alignment of the conserved RdRP domain. Within each phylogenetic tree, the viruses newly identified here are shaded yellow, while those previously reported are shaded light blue and green. The name of each supergroup is shown on the top of each phylogeny and the names of the families within each supergroup are shown on right of the tree. All trees are midpoint-rooted for clarity only, and the scale bar indicates 0.5 amino acid substitutions per site. The tree file of all 180 supergroups are available at http://47.93.21.181/Results/tree_dataset/.

Fig 5. Ecological structure of the global RNA virome. **a**, Global distribution of RNA viral species identified in this study. The Density-Based Spatial Clustering of Applications with Noise (NBSCAN) clustering algorithm[67] was used to cluster the obtained 1,612 different latitude and longitude points from positive samples, where the points with a distance less than 800KM will be aggregated together, and resulting in the 70 aggregated points used to draw the map. Viral species from known RNA virus supergroups are shown in gray, from unclassified supergroups in blue, and from new supergroups in orange. Pie size reflects the number of viral species (log10). **b**, Rarefaction curve of all RNA viral species. Inset, Rarefaction curve of RNA virus species at the ecosystem level with colors indicating different ecosystems. **c**, Distribution of sample size (numbers in the bar chart), alpha diversity, RPM, virus species from known supergroups, virus species from unclassified supergroups and virus species from new supergroups at different ecosystem subtypes and colored by their ecosystem. “Extreme ecosystem” here refers to high salinity, high temperature, or low temperature environmental types. Only ecosystem subtypes with more than nine libraries (excluded poly-A sequencing) were retained for ecological comparison. The ecosystem subtypes on the y-axis are ordered from the highest to the lowest alpha diversity for each

ecosystem. **d**, Viral distribution patterns in environmental and animal samples. The relative abundance of viruses in each library was calculated and normalized by the number of mapped reads per million no-rRNA reads (RPM). RNA virus species from 11 ecosystem subtypes are shown and divided into three groups, indicated by the colors on the heatmap.

Extended Data Fig. 1 Geographic coverage of the meta-transcriptomic data analyzed in this study. a, Geographical distribution of samples at the ecosystem level. Pie size is positively correlated to the number of samples (log10). The DBSCAN clustering algorithm was applied to cluster 1837 latitude and longitude points from all metatranscriptomes into 70 clustered points. **b**, Total number of samples at different ecosystems.

Extended Data Fig. 2. Detailed RNA virus discovery pipeline. a, Schematic diagram of homology-based discovery (i.e., ClstrSearch) and RdRP AI modeling (i.e., LucaProt). **b**, Protein clustering process; only clusters with more than ten members are retained for viral cluster discovery. **c**, Ten iterations of RdRP expansion by recruiting newly detected RdRPs in this process. **d**, RdRP clusters merging into RdRP supergroups using BLAST median *e*-values. **e**, RdRP identification by a new deep learning model (i.e., LucaProt), includes five modules: Input, Tokenizer, Encoder, Pooling, and Output.

Extended Data Fig. 3. Benchmarking of the threshold used in three processes (clustering, expand and merge). a, Number of hits with Diamond blastp v0.9.25.126 using different *e*-values at the test stage. A total of 15,000 sequences were randomly sampled for the test, and host hits were recognized by comparison against the NCBI non-redundant (nr) protein database (version 2023.01.09) with an *e*-value threshold of 1*E*-3. **b**, Benchmarking of hmmscan bitscore and aligned fraction using the RdRP and non-RdRP data sets (including RT, Eu DdRP and Ed RdRP derived from the NCBI GenBank database). **c**, BLAST Median *e*-value within the same known RdRP cluster. **d**, BLAST Median *e*-value between pairwise comparisons of known RdRP clusters, with a 1*E*-3 cut-off used for cluster merging.

Extended Data Fig. 4. Benchmarking of the LucaProt for RdRP modeling. a, The

Sigmoid function of the LucaProt. The Sigmoid function, also called the Logistic function, is used for the output layer of a binary classification model, and mapping a real number to the probability of 0 ~ 1. **b,** Statistics of the data set for LucaProt building, including the entire data set, training set, validation set, and testing set. **c,** The confusion matrix on the validation dataset. **d,** The confusion matrix on the testing dataset. **e,** The metrics for the validation and testing datasets.

Extended Data Fig. 5. Benchmarking of LucaProt with other bioinformatic tools for virus discovery. a-d, The recall, precision, false positive rate, and false discovery rate of LucaProt, Diamond, HMMscan, HH-suite, and Palmscan compared on the 50 metatranscriptomes sequenced in this study. **e,** The recall of LucaProt, Diamond, HMMscan, HH-suite, and Palmscan in identifying the total RdRPs from this study was compared. **f,** The recall of LucaProt, Diamond, HMMscan, HH-suite, and Palmscan in this study was compared for the RdRPs additionally identified by LucaProt. **g,** Recall of LucaProt based on RdRPs (length ≥300aa) found in six previous studies. **h,** The average time used by each tool was calculated based on six datasets of varying lengths, each composed of 50 positive sequences and 50 negative sequences.

Extended Data Fig. 6. Comparison of RNA virus discovery between six previous studies and the current study. a, The distribution of putative viral species of seven studies at the supergroup level and the study-specific level. **b,** Upset plot shows the number of viral species found in each study and those shared/unique between and among seven studies. Rightmost histogram shows the number of metatranscriptomes analyzed in each study. It should be noted that Olendraite *et al*. extracted viral RdRPs from the NCBI Transcriptome Shotgun Assembly (TSA) database, making it impossible to determine the exact number of analyzed metatranscriptomes.

Extended Data Fig. 7. The three highly conserved RdRP sequence motifs (A, B and C) and the distribution of BLAST median e-value, HMM score and mean AI modeling

probabilities of the RNA virus supergroups. RNA viral supergroups were divided into three groups by homology to reference and other published data sets, with the conserved RdRP A, B and C motifs of each supergroup shown on the left.

Extended Data Fig. 8. The distribution of AI modeling probabilities of viral RdRPs. a, Distribution of AI modeling probabilities for all RdRPs from known viral supergroups (first left column) and representatively known RdRP supergroups (right four columns). **b,** Distribution of AI modeling probabilities for all RdRPs (first left column) from unclassified viral supergroups (first left column) and representatively unclassified RdRP supergroups (right four columns). **c,** Distribution of AI modeling probabilities for all RdRPs (first left column) from new supergroups and representatively new RdRP supergroups (right four columns).

Extended Data Fig. 9. Expression difference of RNA viruses and DNA organisms in our newly sequenced data. a, Abundance comparisons for 58 RNA viral supergroups, four DNA virus families, RT and cell organisms at DNA and RNA libraries. **b**, RT-PCR results of second pairs of validation primers for representative RdRP sequences from 17 RNA viral supergroups, capsid sequences from two DNA virus families (*Podoviridae* and *Siphoviridae*), and RT sequences.

Extended Data Fig. 10. Genome coverage for representative RNA viruses and DNA organisms in our newly sequenced data. Data comprises 42 RNA viral supergroups, four DNA virus families, RT, and cellular organisms. Genomes at high abundance in the RNA libraries were chosen to check read coverage in the DNA libraries.

Extended Data Fig. 11. Phylogenetic tree of the Nido-like supergroup and the genome structure of representatives viruses. The tree was estimated using a maximum likelihood method based on the conserved RdRP domain. The reference sequences reported previously are shaded grey, while the viruses newly identified here are shaded by different colors according to different ecotypes. The names of viral families are shown on right of the tree.

The tree was midpoint-rooted for clarity only, and the scale bar of tree indicates 0.2 amino acid substitutions per site. The genome structures of representative viruses are shown on right of the tree. At the bottom, scale indicates the length in nucleotides.

Extended Data Fig. 12. Association between RNA viral supergroups and other non-RdRP protein clusters. Grey, blue and green circles denote known, unclassified and new supergroups, respectively. Orange circles denote non-RdRP protein clusters.

Extended Data Fig. 13. Specificity and shareability of RNA viruses. a, Number of specific putative viral species (i.e., putative "marker" viral species) in each ecosystem subtype. b, Association between RNA viruses and different environmental ecosystems. The size of the colored circles indicates the number of putative viral species identified by each ecosystem type, while the thickness of the line indicates the number of viral species shared by each ecosystem.

Extended Data Fig. 14. Comparison of CHEER, VirHunter, Virtifier, RNN-VirSeeker and LucaProt. a, Positive rate of prediction results for CHEER, VirHunter, Virtifier, RNN-VirSeeker and LucaProt based on the test data set. **b,** False positive rate of prediction results for CHEER, VirHunter, Virtifier, RNN-VirSeeker and LucaProt based on the test data set. **c,** Recall rate of prediction results for CHEER, VirHunter, Virtifier, RNN-VirSeeker and LucaProt based on all the RdRPs identified this study. **d,** Number of viral sequences of different groups by contig length identified by CHEER, VirHunter, Virtifier, RNN-VirSeeker and LucaProt. The training machines, training data sets, training strategies, and final model selection of all comparison models are consistent with LucaProt. All comparison models were built using multiple sets of hyperparameters with the best results selected for the comparison.

Supplementary Table 1. Detailed information of the 10,437 meta-transcriptomics retrieved from the SRA database.

Supplementary Table 2. Detailed information on the 50 environmental samples collected in this study.

Supplementary Table 3. Information on the RdRP sequences identified in this study.

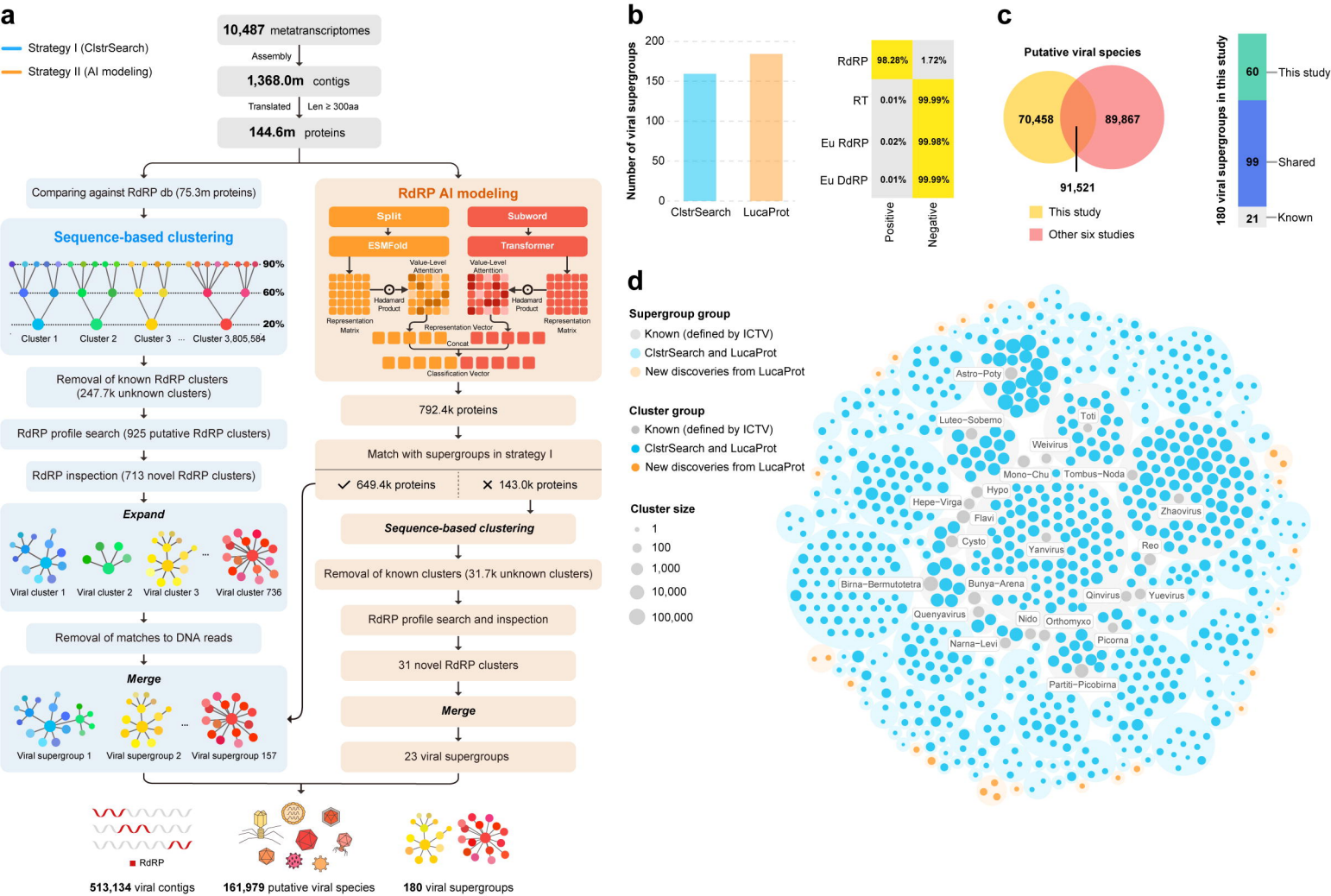
Supplementary Table 4. Comparison of known viral supergroups between this study and the current ICTV RNA virus classification system. Viral supergroups that are incompatible with current viral phyla or classes are shown in orange.

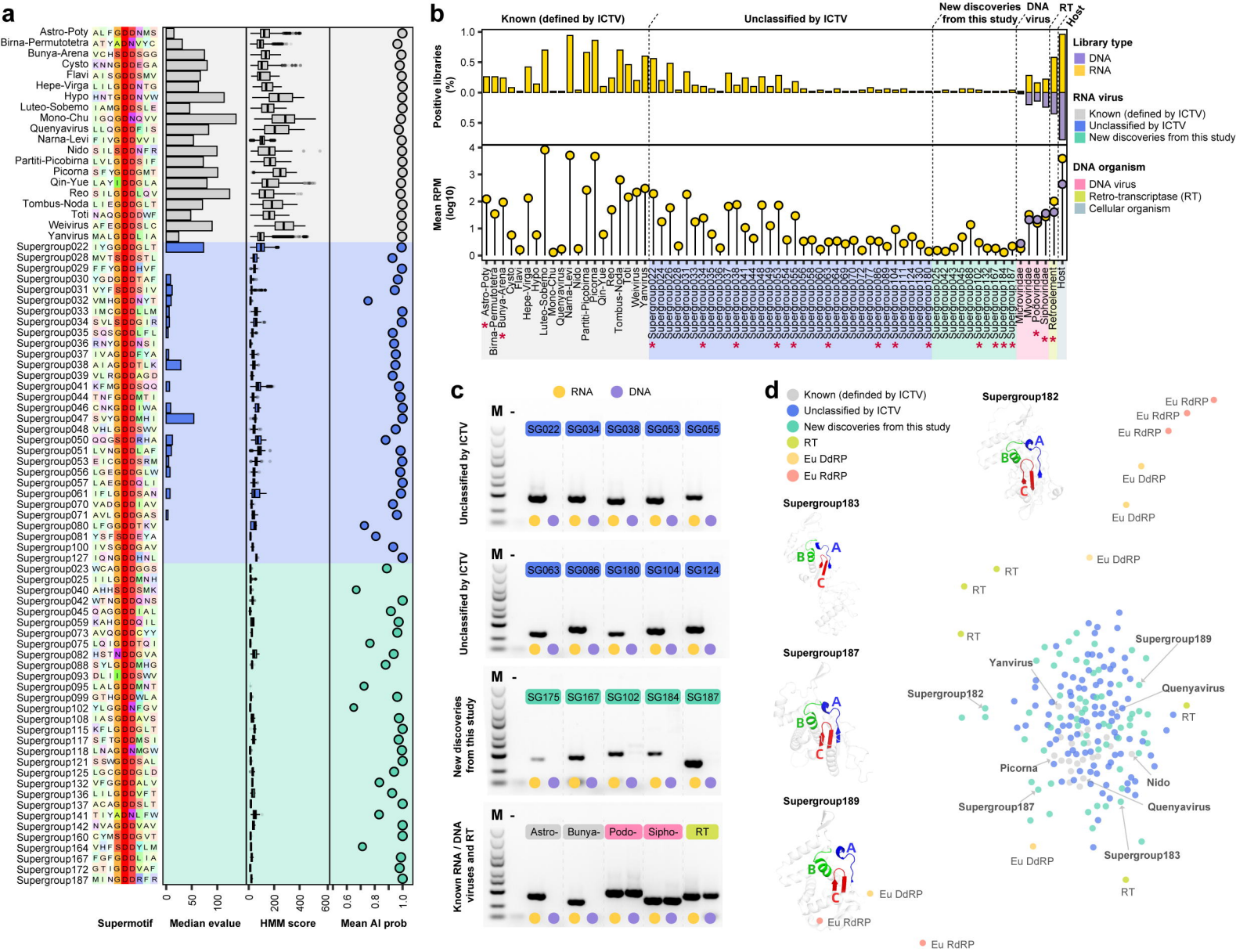
Supplementary Table 5. Predicted results of RdRPs identified in this study using three methods (Threshold: BLAST: $e \leq$1E-3; HMM: score $\geq$10; AI: prob $\geq$0.5).

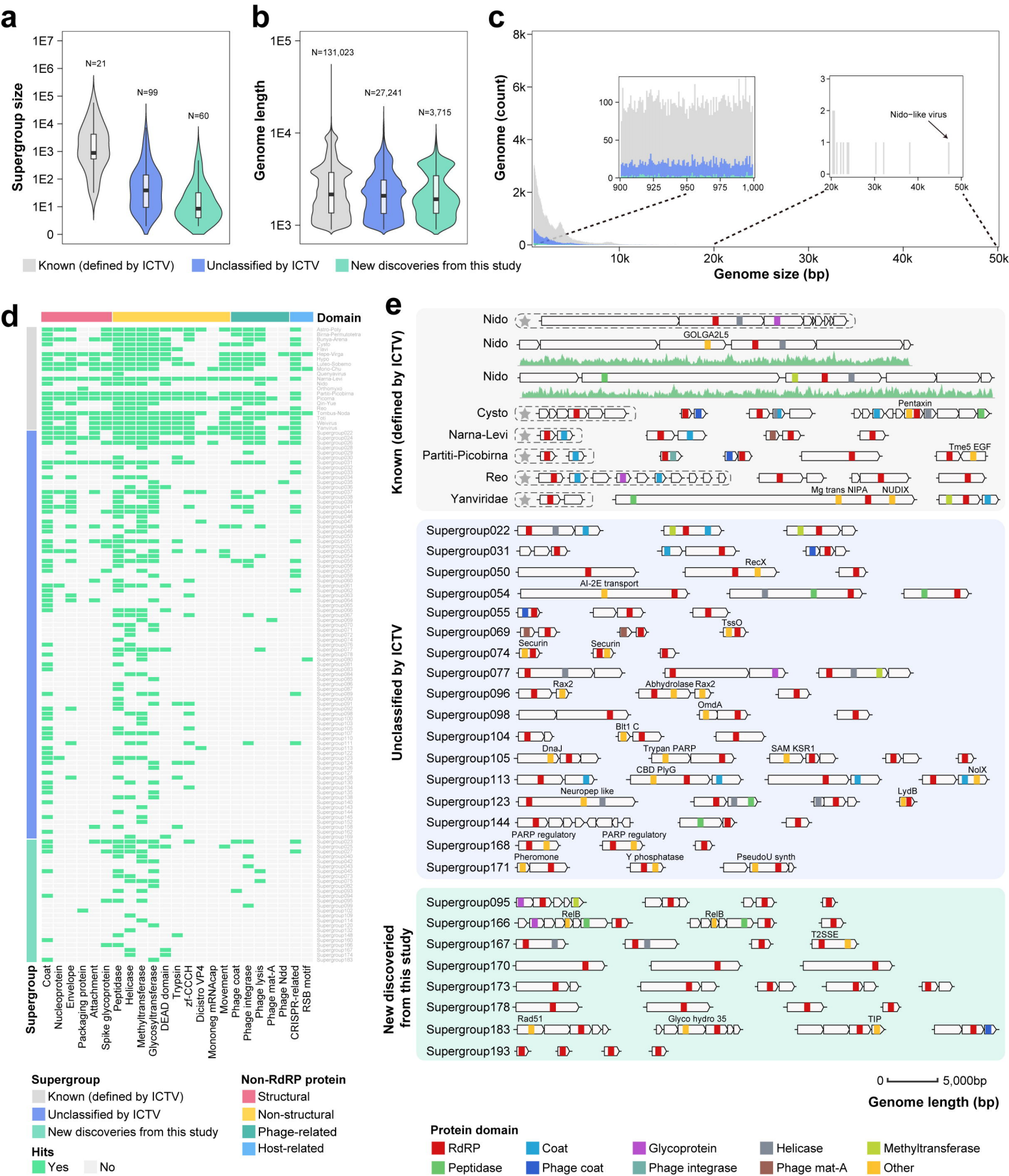
Supplementary Table 6. Viral supergroups identified in 50 collected samples were confirmed through DNA read mapping and RT-PCR.

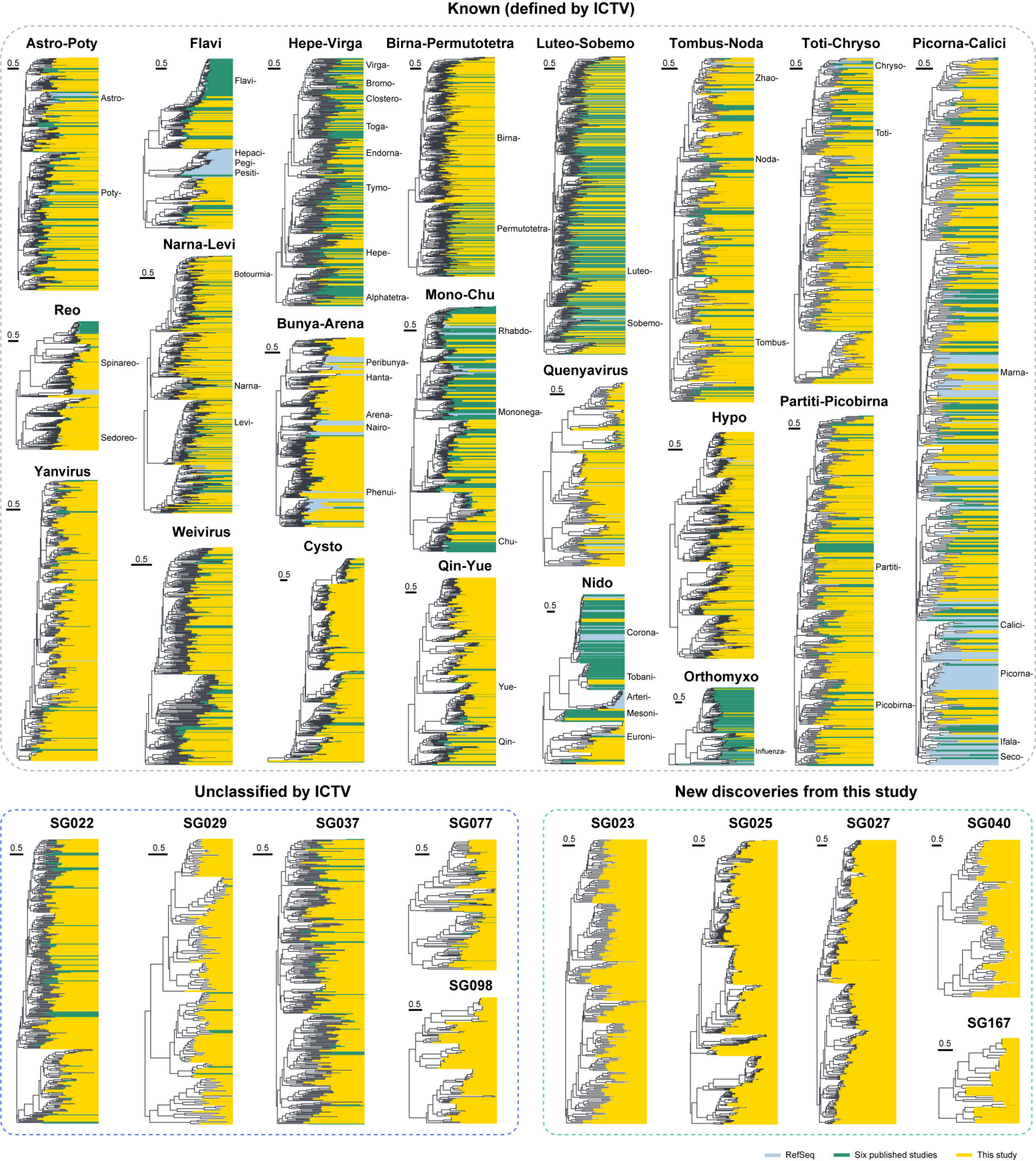
Supplementary Table 7. Size information (the number of dedup-contigs) of all viral supergroups.

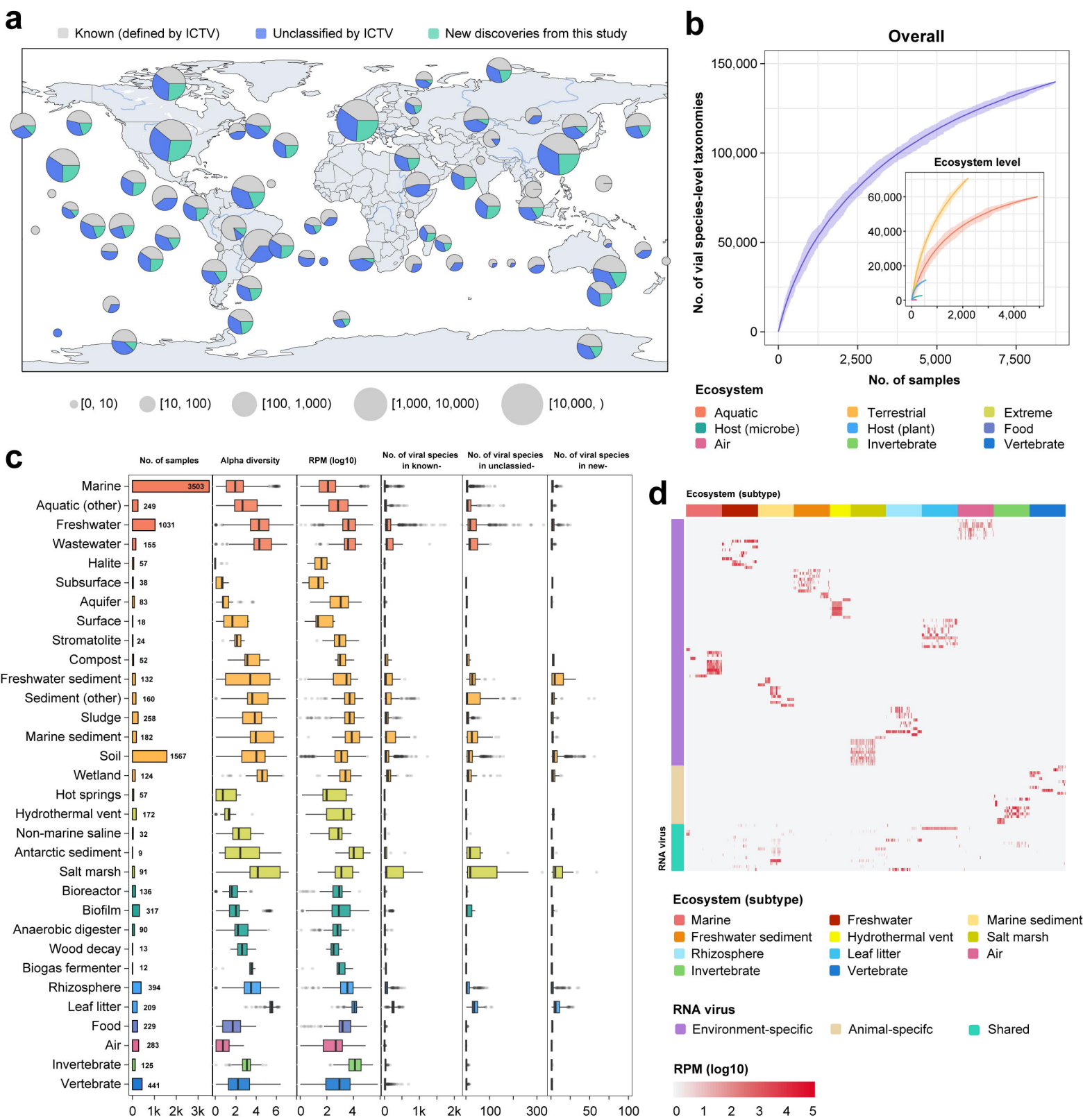
Supplementary Table 8: Normalized abundance levels (measured by RPM) of each viral species in environmental samples (RPM $\geq$1, coverage $\geq$20%).


a
Strategy I (ClstrSearch)
Strategy II (AI modeling)
10,487 metatranscriptomes
Assembly
1,368.0m contigs
Translated
Len ≥ 300aa
144.6m proteins
Comparing against RdRP db (75.3m proteins)
Sequence-based clustering
90%
60%
20%
Cluster 1
Cluster 2
Cluster 3
Cluster 3,805,584
Removal of known RdRP clusters
(247.7k unknown clusters)
RdRP profile search (925 putative RdRP clusters)
RdRP inspection (713 novel RdRP clusters)
Expand
Viral cluster 1
Viral cluster 2
Viral cluster 3
Viral cluster 736
Removal of matches to DNA reads
Merge
Viral supergroup 1
Viral supergroup 2
Viral supergroup 157
RdRP AI modeling
Split
Subword
ESMFold
Transformer
Value-Level Attention
Hadamard Product
Representation Matrix
Representation Vector
Concat
Classification Vector
792.4k proteins
Match with supergroups in strategy I
649.4k proteins
143.0k proteins
Sequence-based clustering
Removal of known clusters (31.7k unknown clusters)
RdRP profile search and inspection
31 novel RdRP clusters
Merge
23 viral supergroups
RdRP
513,134 viral contigs
161,979 putative viral species
180 viral supergroups
b
Number of viral supergroups
200
150
100
50
0
ClstrSearch
LucaProt
RdRP 98.28% 1.72%
RT 0.01% 99.99%
Eu RdRP 0.02% 99.98%
Eu DdRP 0.01% 99.99%
Positive
Negative
c
Putative viral species
70,458
89,867
91,521
This study
Other six studies
180 viral supergroups in this study
60 This study
99 Shared
21 Known
d
Supergroup group
Known (defined by ICTV)
ClstrSearch and LucaProt
New discoveries from LucaProt
Cluster group
Known (defined by ICTV)
ClstrSearch and LucaProt
New discoveries from LucaProt
Cluster size
1
100
1,000
10,000
100,000
Astro-Poty
Luteo-Sobemo
Toti
Weivirus
Mono-Chu
Tombus-Noda
Hypo
Hepe-Virga
Flavi
Zhaovirus
Cysto
Yanvirus
Reo
Birna-Bermutotetra
Bunya-Arena
Qinvirus
Yuevirus
Quenyavirus
Nido
Orthomyxo
Narna-Levi
Picorna
Partiti-Picobirna


a
Supermotif
Median evalue
HMM score
Mean AI prob
b
Known (defined by ICTV)
Unclassified by ICTV
New discoveries from this study
DNA virus
RT
Host
Positive libraries (%)
Mean RPM (log10)
Library type
DNA
RNA
RNA virus
Known (defined by ICTV)
Unclassified by ICTV
New discoveries from this study
DNA organism
DNA virus
Retro-transcriptase (RT)
Cellular organism
c
RNA
DNA
Unclassified by ICTV
SG022
SG034
SG038
SG053
SG055
Unclassified by ICTV
SG063
SG086
SG180
SG104
SG124
New discoveries from this study
SG175
SG167
SG102
SG184
SG187
Known RNA / DNA viruses and RT
Astro-
Bunya-
Podo-
Sipho-
RT
d
Known (defined by ICTV)
Unclassified by ICTV
New discoveries from this study
RT
Eu DdRP
Eu RdRP
Supergroup182
Supergroup183
Supergroup187
Supergroup189
Yanvirus
Quenyavirus
Picorna
Nido


a
Supergroup size
N=21
N=99
N=60
b
Genome length
N=131,023
N=27,241
N=3,715
c
Genome (count)
Nido−like virus
Genome size (bp)
Known (defined by ICTV)
Unclassified by ICTV
New discoveries from this study
d
Domain
Supergroup
Coat
Nucleoprotein
Envelope
Packaging protein
Attachment
Spike glycoprotein
Peptidase
Helicase
Methyltransferase
Glycosyltransferase
DEAD domain
Trypsin
zf-CCCH
Dicistro VP4
Mononeg mRNAcap
Movement
Phage coat
Phage integrase
Phage lysis
Phage mat-A
Phage Ndd
CRISPR-related
RSB motif
Supergroup
Known (defined by ICTV)
Unclassified by ICTV
New discoveries from this study
Hits
Yes
No
Non-RdRP protein
Structural
Non-structural
Phage-related
Host-related
e
Known (defined by ICTV)
Nido
GOLGA2L5
Cysto
Pentaxin
Narna-Levi
Partiti-Picobirna
Tme5 EGF
Reo
Yanviridae
Mg trans NIPA
NUDIX
Unclassified by ICTV
Supergroup022
Supergroup031
Supergroup050
RecX
Supergroup054
AI-2E transport
Supergroup055
Supergroup069
TssO
Supergroup074
Securin
Supergroup077
Supergroup096
Rax2
Abhydrolase Rax2
Supergroup098
OmdA
Supergroup104
Blt1 C
Supergroup105
DnaJ
Trypan PARP
SAM KSR1
Supergroup113
CBD PlyG
NolX
Supergroup123
Neuropep like
LydB
Supergroup144
Supergroup168
PARP regulatory
Supergroup171
Pheromone
Y phosphatase
PseudoU synth
New discoveried from this study
Supergroup095
Supergroup166
RelB
Supergroup167
T2SSE
Supergroup170
Supergroup173
Supergroup178
Supergroup183
Rad51
Glyco hydro 35
TIP
Supergroup193
0
5,000bp
Genome length (bp)
Protein domain
RdRP
Coat
Glycoprotein
Helicase
Methyltransferase
Peptidase
Phage coat
Phage integrase
Phage mat-A
Other


Known (defined by ICTV)
Astro-Poty
Astro-
Poty-
Flavi
Flavi-
Hepaci-
Pegi-
Pesiti-
Hepe-Virga
Virga-
Bromo-
Clostero-
Toga-
Endorna-
Tymo-
Hepe-
Alphatetra-
Birna-Permutotetra
Birna-
Permutotetra-
Luteo-Sobemo
Luteo-
Sobemo-
Tombus-Noda
Zhao-
Noda-
Tombus-
Toti-Chryso
Chryso-
Toti-
Picorna-Calici
Marna-
Calici-
Picorna-
Ifala-
Seco-
Narna-Levi
Botourmia-
Narna-
Levi-
Reo
Spinareo-
Sedoreo-
Bunya-Arena
Peribunya-
Hanta-
Arena-
Nairo-
Phenui-
Mono-Chu
Rhabdo-
Mononega-
Chu-
Quenyavirus
Hypo
Partiti-Picobirna
Partiti-
Picobirna-
Yanvirus
Weivirus
Cysto
Qin-Yue
Yue-
Qin-
Nido
Corona-
Tobani-
Arteri-
Mesoni-
Euroni-
Orthomyxo
Influenza-
0.5
Unclassified by ICTV
SG022
SG029
SG037
SG077
SG098
New discoveries from this study
SG023
SG025
SG027
SG040
SG167
RefSeq
Six published studies
This study


a
Known (defined by ICTV)
Unclassified by ICTV
New discoveries from this study
[0, 10)
[10, 100)
[100, 1,000)
[1,000, 10,000)
[10,000,)
b
Overall
No. of vial species-level taxonomies
No. of samples
Ecosystem level
Ecosystem
Aquatic
Terrestrial
Extreme
Host (microbe)
Host (plant)
Food
Air
Invertebrate
Vertebrate
c
No. of samples
Alpha diversity
RPM (log10)
No. of viral species in known-
No. of viral species in unclassied-
No. of viral species in new-
Marine
Aquatic (other)
Freshwater
Wastewater
Halite
Subsurface
Aquifer
Surface
Stromatolite
Compost
Freshwater sediment
Sediment (other)
Sludge
Marine sediment
Soil
Wetland
Hot springs
Hydrothermal vent
Non-marine saline
Antarctic sediment
Salt marsh
Bioreactor
Biofilm
Anaerobic digester
Wood decay
Biogas fermenter
Rhizosphere
Leaf litter
Food
Air
Invertebrate
Vertebrate
d
Ecosystem (subtype)
RNA virus
Ecosystem (subtype)
Marine
Freshwater
Marine sediment
Freshwater sediment
Hydrothermal vent
Salt marsh
Rhizosphere
Leaf litter
Air
Invertebrate
Vertebrate
RNA virus
Environment-specific
Animal-specifc
Shared
RPM (log10)