Artificial intelligence redefines RNA virus discovery

Xin Hou[1,15], Yong He[2,15], 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], 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]These authors contributed equally

⍰Corresponding authors:

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

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

Abstract

RNA viruses are diverse components of global ecosystems. The metagenomic identification of RNA viruses is currently limited to those with sequence similarity to known viruses, such that highly divergent viruses that comprise the “dark matter” of the virosphere remain challenging to detect. We developed a deep learning algorithm – LucaProt – to search for highly divergent RNA-dependent RNA polymerase (RdRP) sequences in 10,487 global meta-transcriptomes. LucaProt integrates both sequence and structural information to accurately and efficiently detect RdRP sequences. With this approach we identified 180,571 RNA viral species and 180 superclades (viral phyla/classes). This is the broadest diversity of RNA viruses described to date, including many viruses undetectable using BLAST or HMM approaches. The newly identified RNA viruses were present in diverse ecological niches, including the air, hot springs and hydrothermal vents, and both virus diversity and abundance varied substantially among ecological types. We also identified the longest RNA virus genome (nido-like) observed so far, at 47,250 nucleotides, and expanded the diversity of RNA bacteriophage to more than ten phyla/classes. This study marks the beginning of a new era of virus discovery, with the potential to redefine our understanding of the global virosphere and reshape our understanding of virus evolutionary history.

Key Words:

RNA virus; Virome; Virus discovery; Deep learning; Meta-transcriptomes; Phylogeny

Introduction

RNA viruses infect a huge array of host species. Despite this ubiquity, their role as an essential component of global ecosystems has only recently been recognized thanks to systematic and large-scale virus discovery projects performed in animals[1,2], plants[3], fungi[4], marine[5], and soil environments[6]. A common feature of these studies is that they were based on the analysis of RNA-dependent RNA polymerase (RdRP) sequences, a canonical component of RNA viruses. Combined, they have resulted in the discovery of tens of thousands of new virus species, leading to at least a ten-fold expansion of the virosphere and the addition of five new phyla of RNA viruses, including the "Taraviricota"[5]. Similarly, data mining exercises that reanalyzed over 10^9 meta-transcriptomic contigs associated with diverse ecosystems have identified several divergent clades of RNA bacteriophage[7]. Despite such significant progress in filling the gaps of RNA virus diversity through ecological sampling and sequencing, our understanding of the full spectrum of the RNA virosphere is likely limited[8,9]. 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[10], while the profile alignment (i.e., HMM) based approach misses a significant proportion of viruses due to a high false-negative rate[11]. To efficiently uncover the full range of RNA virus diversity, the development of novel strategies is therefore essential.

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, including molecular docking, compound screening and interaction, protein structure prediction and functional annotation, and the modelling of infectious diseases[12-17]. These advancements can be attributed to the advantages of deep learning algorithms over classic bioinformatic approaches, including greater accuracy, better performance, less feature engineering, flexible models, and self-learning capabilities[18,19]. Recently, deep learning approaches, such as CHEER, VirHunter, Virtifier and RNN-VirSeeker have also been developed and applied to identify viruses from genomic and metagenomic data[20-23]. However, many of these approaches rely on nucleotide sequence information without incorporating protein sequence or structural information, and are hence less likely to identify highly divergent RNA viruses. The 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[24-26]. As a consequence, transformer architecture can be used to design a better tool to uncover the hidden “dark matter” of highly divergent RNA viruses. Herein, we show how AI can be used to accurately and efficiently detect RNA viruses that are too divergent in sequence to be detected by traditional sequence similarity-based methods, in doing so revealing a hidden world of virus diversity.

Results

Deep learning to reveal the dark matter of the RNA virosphere

We performed all-inclusive searches to reveal the entirety of RNA virus diversity present in different ecological systems sampled at global scale (Extended Data Fig. 1, Supplementary Table 1 and 2). Accordingly, a total of 10,487 meta-transcriptomes (51 Tb of 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 (i.e., “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 had high accuracy (0.03% false positives) and specificity (0.20% false negatives) on the test data set (Fig. 1b, Extended Data Fig. 4). Independently to the deep-learning approach, 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 any resemblance to viral RdRPs or non-RdRP proteins. The latter approach is distinguished from previous BLAST or HMM based approaches because it queries on protein clusters (i.e., alignments) instead of individual sequences, which greatly reduced both the 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 180,571 RNA viral species (i.e., > 90% RdRP identity), and 180 RNA viral superclades at the phylum level taxonomic rank (Fig. 1, Supplementary Table 3 and see Methods). Among these, 512,691 viral contigs (0.04% of total contigs) and 157 superclades

(87.2%) were revealed by both “LucaProt” and “ClstrSearch”, whereas 443 contigs and 23 superclades were only predicted by “LucaProt”. Both strategies out-performed previous attempts at RNA virus discovery from ocean[5], soil[6], and more diverse ecosystems[7] (Fig. 1c). Indeed, “LucaProt” was able to identify 98.2% ~ 99.9% of RdRPs discovered in these previous studies, even though none were used in either training or testing of the models (Extended Data Fig. 5). To ensure the robustness and introduce innovative findings from the AI approach, we jointly applied the two strategies and merged the results; this enabled us to identify 93,580 viral species and 59 novel superclades, and resulted in a 9-fold expansion in RNA virus diversity (Fig. 1c). This was reflected in the expansion of both existing viral superclades and the identification of new superclades unlikely to be discovered by sequence homology and HMM based approaches alone (Fig. 1d).

All the RNA viral sequences discovered here were organized into clusters and superclades without the influence of the current virus classification system[27,28]. These superclades were then placed back onto the classification system at the phylum (such as phylum Lenarviricota in the case of the Narna-Levi superclade) or class (such as the Stelpaviricetes, Alsuviricetes, Flasuviricetes classes for the Astro-Poty, Hepe-Virga, Flavi superclades) levels (Supplementary Table 4)[28]. Notably, however, the virus superclades comprised much greater phylogenetic diversity than their corresponding phyla/classes. Also of note was that our data did not conform to several of the higher taxonomic ranks, such as the phyla Duplornaviricota and Negarnaviricota, which were now too broad to be regarded as single phyla. Indeed, even the Markov cluster algorithm (MCL) approach, on which the existing virus classification scheme is derived[29,30], fails to re-group these expanded classes into the existing phyla[5].

Verification and confirmation of newly identified viral superclades

That the 180 RNA viral superclades identified represented RNA-based organisms was verified by multiple lines of evidence. At the sequence level, two criteria were used to establish a viral superclade: a lack of homology to cellular proteins and the presence of key RdRP motifs (Fig. 2a). Furthermore, the majority (157/180) of the newly identified superclades shared a variable degree of sequence homology with existing RdRPs (i.e., BLAST *e*-value $\leq$ 1*E*-3 and/or had HMM model score $\geq$ 10). The exception were 23 superclades that had no detectable homology to viral RdRPs and therefore named as "AI-specific" superclades (Fig. 2a, Extended Data Fig. 6, Supplementary Table 5). To justify the computational prediction, we performed simultaneous DNA and RNA extraction and sequencing to examine whether the viral superclades identified here also exist in DNA form. This analysis revealed that only RNA sequencing reads were mapped to contigs associated with viral RdRPs, whereas both RNA and DNA sequencing reads were mapped to contigs associated with DNA viruses, reverse-transcriptase (RT), and cellular organisms (Fig. 2b, Extended Data Fig. 7-9). These results were further confirmed by a more sensitive RT-PCR approach which showed that none of the sequences encoding viral RdRP were detected in the DNA extractions, suggesting that these viral superclades were *bona fide* RNA organisms (Fig. 2c, Extended Data Fig. 7b). Finally, we performed 3D alignment analysis (newly identified viral RdRPs compared with known viral RdRPs, eukaryotic RdRPs, eukaryotic DdRPs and RT) to determine the degree of structure similarity among them (Fig. 2d). The novel viral RdRP superclades (including AI-specific ones) bore at least three signature motifs that gave them much higher structural similarity to known viral RdRPs than their cellular counterparts.

Genomic structures reveal modularity and flexibility within the RNA virosphere

We next analyzed the composition and structure of potential RNA virus genomes identified in this study. The length of the RdRP-encoding genomes or genome segments differed markedly within and between viral superclades, although most were centered around 2,569 nt (Fig. 3). Notably, our data set contained some extremely long RNA virus genomes identified from soil that belonged to the Nido-like superclade: the length of one of these, at 47.3 kb, exceeded *Planarian secretory cell nidovirus* (41.2 kb)[31] as the longest RNA virus genome identified to date (Fig. 3c, Extended Data Fig. 10 Supplementary Table 6). 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. 11). Importantly, the presence of these additional virus proteins in newly identified supergroups provided further evidence that these were *bona fide* RNA viruses. Furthermore, that the occurrence of these proteins was incongruent with the groupings of RdRPs (Fig. 3e) suggests that RNA virus genomes have a modular-like configuration, transferring proteins across taxonomic groups. This was in line with the dramatic changes in genome structure (genome length, gene organization, ORF numbers, and segmentation) observed among related viruses, such that no prototype genome structures could be defined for each group or supergroup (Fig. 3e).

Expanded phylogenetic diversity of RNA viruses

The enormous expansion in the RNA virosphere described here was also reflected in both the growing size of known virus groups and the addition of entirely new groups (Fig. 4). For existing supergroups, the viruses newly described here were distinguished from those identified previously such that they formed unique clusters at more ancestral positions in the phylogenetic trees (Fig. 4). Interestingly, some previously smaller sized viral groups with limited diversity – the Astro-Poty, Hypo, Yan and Cysto – expanded to become large viral groups comprising substantial genetic diversity (Fig. 4). Several newly identified supergroups were also revealed to have high levels of phylogenetic diversity, including SC022 (8,128 species), SC024 (3,682 species), and SC37 (1,772 species), highlighting the limitations in previous attempts to identify highly divergent groups of RNA viruses. Following our analysis, the supergroups with the greatest number of species were the Narna-Levi (64,667 species), Picorna-Calici (23,430 species), and Tombus-Nada (16,798 species).

In addition to greatly expanding virus genetic diversity, this study identified more virus groups associated with bacterial hosts than the leviviruses, cystoviruses, and the members of Partiti-Picobirna supercluster known previously[7]. Specifically, we identified bacterial viruses within the Narna-Levi, Hepe-Virga, and SC037 supergroups whose sequences were recognized and “recorded” by the bacterial CRISPR system. Furthermore, based on proteins associated with bacterial infection (i.e., Lysis, Prok-E2, and Prok-RING), we inferred potential bacterial RNA viruses in the Tombus-Noda, Yan, and SC022 supergroups

(Supplementary Table 7). As a consequence, those RNA viruses associated with bacteria has expanded to ten supergroups, and these numbers are likely to further increase given our limited knowledge of host associations for most of the viruses in this study.

Ecological structure of the global RNA virome

Our study investigated the RNA virome of 10,487 ecological samples, revealing the ubiquitous presence of RNA viruses across diverse ecological types (48 categories) and in 1,837 locations globally. Despite repeated efforts to uncover the RNA virus diversity from such ecological samples[5-7], a large proportion of the viruses detected here were entirely novel (Fig. 5a). Indeed, the rate of RNA virus discovery did not plateau (Fig. 5b), suggesting that the global space of RNA virus diversity remains largely under-characterized, with a particularly rapid increase in soil (Fig. 5b).

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 ecological types, revealing enormous variation (Fig. 5c, Supplementary Table 8). In general, average alpha diversity was highest in leaf litter, estuary, freshwater, and wetland environments, whereas virus abundance was highest in freshwater, marine sediment, and rhizosphere systems, whose average RPMs were between 12466.9 and 26617.3 (Fig. 5c). In contrast, the lowest average diversity and abundance were observed in halite and subsurface environments (Fig. 5c), which as expected as these samples were particularly low in biomass (i.e., host cells). For extreme ecological types 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 worth noting that the new viral superclades established in this study were mostly identified from aquatic and sediment samples, with few from vertebrate and invertebrate animal samples (Fig. 5c).

Our results further revealed the prevalence and abundance levels of single viral species across different ecological types (Fig. 5d), including some that could be considered ecological generalists. For example, members of the Narna-Levi, Partiti-Picobirna and Picorna superclades as well as Superclade022 were among the prominent generalist RNA viruses and found in more than 45 ecological types (Extended Data Fig. 12). Conversely, the majority (85.9%) of the viruses discovered here only occurred in a single ecological type. Finally, we also identified “marker’ virus species for each ecotype, which appeared at high prevalence and abundance in one ecological type but not in the others (Fig. 5d). Among these, *Partiti-Picobirna sp.* 4207 and *Partiti-Picobirna sp.* 9871 were associated with hot springs and *Tombus-Noda sp.* 2280 and *Superclade026 sp.* 2292 were associated with hydrothermal vents, suggesting their important role in these ecosystems.

Discussion

Our understanding of the genetic diversity of the RNA virosphere, and hence of RNA virus ecology and evolution in general, is greatly hampered by the inability to accurately identify the highly divergent “dark matter” of viruses[32,33]. Indeed, the conventional way to discover RNA viruses has relied heavily on the utility of sequence similarity comparisons and the completeness of sequence databases[11,32]. To address these issues, we developed a data-driven deep learning model (i.e., LucaProt) that overcome these shortcomings while outperforming conventional approaches in accuracy, efficiency, and, most importantly, the scope of diversity. Importantly, LucaProt not only incorporated sequence data but also structural information, which is relevant in predicting protein function (in this case of the RdRP)[34]. Without implementing the structural model, our model had only 41.8% and 94.9% specificity and accuracy, respectively, on the testing 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 larger amino acid sequences[35,36], which can easily capture the relationship between residues from distant parts of sequence space, thereby outperforming the CNN and/or RNN encoders implemented in the CHEER, VirHunter, Virtifier and RNN-VirSeeker RNA virus discovery tools (Extended Data Fig. 13)[20-23]. 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 ecology and evolution of the newly discovered viruses. In particular, nothing is known about the hosts of the viruses identified, including that with the longest virus genome identified to date. It is possible that the viral clades and superclades identified here were largely associated with diverse microbial eukaryotic hosts, given that the majority of current known RNA viruses infect eukaryotes[37,38] and microbial eukaryotes exist in great abundance and diversity in natural environments[39,40]. Nevertheless, it is also likely that a substantial proportion of the novel viruses discovered are associated with bacterial (and perhaps archaeal) hosts[41-43]. Indeed, based on this and previous studies[7], more than ten superclades contained RNA viruses likely associated with bacteria. Importantly, the presence of RNA bacteriophages in multiple RNA viral superclades underlines the evolutionary connection between RNA viruses from bacterial and eukaryotic hosts. If viewed through the lens of virus-host co-divergence[1,2,44], such a link between bacterial and eukaryotic hosts 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 the meta-transcriptomic analysis of 10,487 samples for RNA virus discovery. The majority of the 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. Furthermore, the samples included in this study were subject to high-quality short-read sequencing (i.e., utilizing Illumina sequencing platforms), had between 35.1-204.1 Gbp raw sequencing data output, and were not enriched for any specific types of microbial organisms. 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 obtained 50 samples from Antarctica and China for RNA virus discovery and confirmation. 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, China 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 generated and collected for this study, reads were assembled *de novo* into contigs using MEGAHIT v1.2.8[45] 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 takes into account protein sequence composition and structure information to facilitate the accurate identification of viral RdRPs. The model included five modules: Input, Tokenizer, Encoder, Pooling, and Output (Extended Data Fig. 2e).

Input Layer：Our model uses the amino acid sequence as input.

Tokenizer Layer: This module consists of two components. One used a frequent substring algorithm[46], which generated subwords from the input sequence, treated co-occurring amino acids as a whole (namely, “words”), and resulted in a vocabulary with 20,000 such “words”. The other component broke down each protein sequence into a combination of single amino acid characters which were later used in protein structure modeling.

Encoder Layer: This module processes the two types of input into sequence and structural representation matrices, respectively. 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[47], AlpahFold[15], and ESMFold[48]) to predict 3D protein structure, calculated the distance between the C-atoms (Alpha-C or Beta-C) of all amino acid residues into a Contact Map matrix, and applied Graph Convolutional Network (GCN)[49] 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[48] for structural representation.

Pooling Layer: The previous module obtained the sequence and structure representation matrices. A value-level attention pooling (VLAP) approach[50] 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. 2e) 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, which included 5,979 positive samples of known viral RdRPs (i.e., the well-curated RdRP database described above), and randomly selected 229,434 negative samples of confirmed non-virus RdRPs (as the positive sample accounts 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-like sequences contained proteins from the eukaryotic RNA dependent RNA polymerase (Eu RdRP, N = 2,233), the 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 a wide 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).

LucaProt identified 792,436 putative RdRP signatures from 144,690,558 proteins. These results were first compared with the RdRPs identified based on sequence homology (see below). RdRPs that were identified only by deep learning algorithms were either incorporated into the superclades using the Diamond blastp program v0.9.25.126[51] with an *e*-value threshold of 1*E*-3, or, if they remained unclassified, were subjected to clustering, merging, and manual alignment inspection as described below to form deep learning specific superclades (the case for 23 superclades).

Identification of RNA viruses based on homologous clustered proteins

The first approach to identify RNA viruses was based on sequence and structural similarity to previously known RdRP amino acid sequences (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 from vertebrate and invertebrate hosts[1,2]. The comparisons were performed using the Diamond blastp program v0.9.25.126[51], with the *e*-value threshold set at 1*E*+5 to identify more divergent RdRP proteins (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), which resulted in 3,805,584 clusters. False positives and hits to known RdRP proteins were removed by comparing against the NCBI non-redundant (nr) protein database, the NCBI RefSeq protein database and the virus RdRP database (Extended Data Fig. 2b). The remaining unknown protein clusters were subject to viral RdRP domain search using a hidden Markov models (HMMs) built from a manually reviewed profile of known RdRP clusters using the program hmmscan v3.3.2 (e = 10, hits ≥ 1)[52]. Clusters that contain more than one hmmscan hit were subsequently aligned and inspected for the presence of conserved RdRP motifs. Finally, a total of 713 novel RdRP clusters were retained as a result of our rigorous screening and checking steps.

To further expand the RdRP collection based on the viruses newly discovered here, we updated the RdRP protein database with the 713 novel RdRP clusters identified here and used it to detect additional RdRP sequences from the original 144.6 million amino acid sequences using the Diamond blastp and an *e*-value threshold of 1*E*-3. The newly detected RdRPs were again incorporated into the RdRP database for another round of detection. This process was repeated for ten iterations. The resulting RdRP proteins (21,747,015 in total) were subjected to the homology-based clustering, the removal of false positives, a HMMs-based search, and manual alignment inspection as described above (Extended Data Fig. 2c, Extended Data Fig. 3b).

Finally, the remaining clusters were merged into superclades using a hierarchical method employing the Girvan–Newman algorithm[53], 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 used the following four 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[54] implemented in *R*.

Virus verification

To determine whether the newly discovered viral RdRPs belonged to RNA viruses rather than organisms with DNA genomes, we performed two experiments. First, the 50 environmental samples collected in this study were subject to simultaneous RNA and DNA extraction and sequencing. The reads from the DNA sequencing results were mapped against the RdRP sequences to verify that there was no DNA counterpart. Quality control of viral contigs was performed using bbduk.sh (https://sourceforge.net/projects/bbmap/), and the mapping analyses were performed by Bowtie2 v2.4.2[55] with the "end-to-end" setting. Similarly, from our collection of SRA data, we also searched for those studies that performed both RNA and DNA sequencing, and these data were used for mapping analyses to confirm that the viruses discovered had *bona fide* RNA genomes.

In addition to read mapping, RT-PCR assays were performed to confirm that the detected viral superclades were RNA organisms. Two pairs of validation primers were designed for each of the representative RdRP sequences from 17 RNA viral superclades, gene sequences from two DNA virus families (i.e., *Podoviridae* and *Siphoviridae*), and RT sequences identified in this study, with a product length of 300-550 bp. For each of the 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 of viral RdRPs and homologous proteins

Three-dimensional structures of newly identified viral RdRPs from diverse RNA viral superclades were predicted from primary sequences using AlphaFold 2 v2.3[15] and visualized using the PyMol software v2.5.4 (http://www.pymol.org/pymol). AlphaFold 2 prediction is a relatively reliable source of structure information as the pLDDT socre of more than 2/3 residues it predicted are above 75%. The previously resolved or predicted structures of viral RdRP, eukaryotic RdRP, eukaryotic DdRP and RT proteins were compared using the Super algorithm[56]. Considering that the protein structures have similar molecular weights but substantial variations in their conformations, 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) were predicted from 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 (https://pfam-legacy.xfam.org/) using hmmscan implemented in HMMER[52]. For the remaining ORFs, the annotation was carried out by blastp comparisons against the nr protein database with an *e*-value threshold of 1*E*-3.

Analyses of virome diversity, evolution and ecology

To reveal the diversity of the RNA viruses identified, we used an RdRP identity threshold of 90% to define new virus species. Abundance levels were subsequently estimated for every virus species based on the number of non-rRNA reads per million (RPM) within each sample (i.e. sequencing runs) mapped to viral sequences belonging to that species. Virus alpha diversity (measured with the Shannon index) and overall abundance were subsequently estimated and compared across different geographic locations and ecological types, namely; soil, marine, freshwater, wetland, hot spring, salt marsh, and other types. “Marker virus species” that were greatly enriched in certain ecological types were also identified based on virus mapping results. The marker virus species were defined as present only in one ecological subtype with RPM ≥ 1 and coverage ≥ 20%. To reveal the diversity and evolutionary relationship of RNA viruses within a superclade, RdRP representatives of overall diversity were first selected based on homology-based clustering. These representatives were aligned using L-INS-I algorithm implemented in Mafft v7.475[57]. Phylogenetic analyses were performed based on the alignment using a maximum likelihood algorithm, a LG amino acid substitution model, a Subtree Pruning and Regrafting (SPR) branch swapping algorithm, and a Shimodaira–Hasegawa-like procedure implemented in the

Phyml program v3.1[58].

Identification of CRISPR spacer hits

A CRISPR-Cas spacer database was compiled from 65,703 genomes of bacteria and archaea downloaded from the GTDB database (https://gtdb.ecogenomic.org/)[59] using a modified version of the CRISPR Recognition Tool (CRT)[60]. This database was supplemented with an additional 11.8 million precompiled CRISPR-Cas spacers obtained from the CrisprOpenDB spacer database (http://crispr.genome.ulaval.ca)[61]. All spacers were queried for exact matches against viral contigs using the BLASTn-short function implemented in the NCBI BLAST v2.9.0+ package[62] with parameters "-evalue 1*E*-10, -perc_identity 95, -dust no -word_size 7", allowing only 0-1 mismatches across the entire length of the spacer to minimize the number of false-positive hits.

Data availability

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/, which includes all RNA virus contigs, RdRP CDS, RdRP proteins, RdRP HMM profiles and phylogenetic tree files. Additionally, this website also includes related data sets for model building and validation, and the trained model of LucaProt.

Code availability

The original codes of ClstrSearch and LucaProt are stored at GitHub repository (https://github.com/alibaba/LucaProt), and 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 upon request.

Acknowledgements

This work was supported by Shenzhen Science and Technology Program (KQTD20200820145822023 and JCYJ20210324124414040), National Natural Science Foundation of China (32270160), 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. for their contribution of 15 machines with 128 CPUs and 1T RAM of Elastic High-Performance Computing (EHPC), 48 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 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).

6 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).

7 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).

8 Breitbart, M. & Rohwer, F. Here a virus, there a virus, everywhere the same virus? *Trends Microbiol* **13**, 278-284, doi:10.1016/j.tim.2005.04.003 (2005).

9 Youle, M., Haynes, M. & Rohwer, F. in *Viruses: Essential Agents of Life* (ed Günther Witzany) 61-81 (Springer Netherlands, 2012).

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

11 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).

12 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).

13 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).

14 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).

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

16 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).

17 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).

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

19 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).

20 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).

21 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).

22 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).

23 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).

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

25 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).

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

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

28 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).

29 Dongen, S. v. Graph clustering by flow simulation. (2000).

30 Enright, A. J., Van Dongen, S. & Ouzounis, C. A. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* **30**, 1575-1584 (2002).

31 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).

32 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).

33 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).

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

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

36 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).

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

38 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).

39 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).

40 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).

41 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).

42 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).

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

(2018).

44 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).

45 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).

46 Lennox, M., Robertson, N. & Devereux, B. Expanding the Vocabulary of a Protein: Application of Subword Algorithms to Protein Sequence Modelling. *Annu Int Conf IEEE Eng Med Biol Soc* **2020**, 2361-2367, doi:10.1109/EMBC44109.2020.9176380 (2020).

47 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).

48 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).

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

50 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).

51 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).

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

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

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

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

56 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).

57 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).

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

59 Donovan H Parks, M. C., Christian Rinke, Aaron J Mussig, Pierre-Alain Chaumeil, Philip Hugenholtz. GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalizedand complete genome-based taxonomy. *Nucleic Acids Res* **50**, D785-D794, doi:10.1093/nar/gkab776 (2021).

60 Bland, C. *et al.* CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinformatics* **8**, 209 (2007).

61 Dion, M. B. *et al.* Streamlining CRISPR spacer-based bacterial host predictions to decipher the viral dark matter. *Nucleic Acids Res* **49**, 3127-3138, doi:10.1093/nar/gkab133 (2021).

62 Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *Journal of molecular biology* **215**, 403-410 (1990).

Figure and Table Legends

Fig. 1. Global diverse 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 superclades discovered using two methods (left), and the detection accuracy of RdRP AI modelling (right). **c**, Venn diagram shows the shared representative viral species between available data from Zayed *et al*., Neri *et al*., Chen *et al*., and this study. The bar graph shows the shared viral superclades between the four studies and the unique viral superclades identified in this study. **d**, Diverse clusters of RNA viruses (dark colored small circle) and RNA virus superclades (light colored large circle). The known viral clusters and superclades are denoted in dark grey and light grey, respectively. The novel viral clusters discovered in this study are denoted with dark blue and dark orange circles, while the novel viral superclades are denoted with light blue and light orange circle. The size of the circle reflects the size of the viral cluster.

Fig. 2. Evaluation of authenticity of RNA viral superclades. a, Distribution of BLAST median *e*-value, HMM score and mean AI modeling probabilities of RNA virus superclade grouping by the sensitivity of three methods, with the primary sequence-identified conserved RdRP motif C of each superclade shown on the left. The known viral superclades show high sensitivity for all three methods and are shown in grey. The novel superclades show declining homology but the relative stable AI probability. **b**, The positive libraries and mean RPM (the number of mapped reads per million non-rRNA reads) of representative viral superclades, 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 first pairs of validation primers for representative RdRP sequences from 17 RNA viral superclades, 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 superclades with Eu DdRPs, Eu RdRPs, and RT. Each point stands for a representative structure. The distance between different points represents structure similarity and the greater the distance, the lower the structure similarity. Four RdRP domain structures of the AI-specific superclades are displayed with the A, B and C motifs highlighted.

Fig 3. Genomic features of viral superclades. **a**, Size (the number of contigs) of all novel viral superclades compared to 21 known superclades. **b**, Genome length of all novel viral superclades compared to 21 known superclades. Centre lines in the box plots represent the median bounds. **c**, Histogram of the genome size distribution of RNA viruses from known and novel viral superclades. **d**, The distribution of annotated functional protein in each viral superclade. **e**, Genome structure of representatives from six known superclades, 17 novel superclades and eight AI-specific superclades. Grey stars represent reference virus genomes of known superclades. 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 32 RNA viral superclades. Each phylogenetic tree was

estimated using a maximum likelihood method based on the conserved RdRP domain. Within each phylogenetic tree, the viruses newly identified here are shaded yellow, those reported previously are shaded green and blue. The name of each superclade is shown on the top of each phylogeny and the names of the families within each superclade are shown on right of the tree. The proteins associated with bacterial hosts are denoted with different shapes on the right side of the corresponding viral sequence. All trees are midpoint-rooted for clarity only, and the scale bar indicates 0.5 amino acid substitutions per site.

Fig 5. Ecological dynamics of the global RNA virome. **a**, Global distribution of RNA viruses identified in this study. Species of known virus superclades are shown in gray and species from novel superclades are shown in magenta. Pie size reflects the number of viral species. **b**, Rarefaction curve of all RNA viral species. Inset, Rarefaction curve of RNA viral species at the ecotype level with colors indicating different ecotypes. **c**, Distribution of alpha diversity, RPM, novel viral species and AI-specific species at different ecological subtypes and colored by their ecotype. The ecological subtypes on the y-axis are ordered from the highest to the lowest alpha diversity for each ecotype. **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). Viral species from 11 ecological 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 ecotype level. Pie size is positively correlated to the number of samples. b, Total number of samples at different ecotypes.

Extended Data Fig. 2. Detailed RNA virus discovery pipeline. a, Schematic diagram of homology-based discovery and RdRP AI modeling. **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 RdRP in this process. **d**, RdRP clusters merging into RdRP superclades using BLAST median *e*-value. **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 at three processes (clustering, expand and merge). a, Number of hits using different *e*-values at the test stage. **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 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 AI RdRP modeling. a, The sigmoid function of the AI modeling. **b,** Statistics of the data set for AI model building, including the entire data set, training set, validation set, and testing set. **c,** The distribution of AI modeling probabilities of positive data sets, **d,** The AI distribution of AI modeling probabilities of negative data sets, including RT, Eu DdRP and Eu RdRP.

Extended Data Fig. 5. Comparisons of RNA virus discovery results between three previous studies and the current study. a, The distribution of representative viral RdRPs of four studies at the superclade level and the study-specific level. **b,** Venn diagram shows the number of RdRP superclades found in each study and those shared between and among four studies. **c,** Venn diagram shows the number of representative RdRPs found in each study and those shared between and among four studies. **d,** Bar graph shows the number of known, novel, AI-specific and study-specific RdRPs of four studies.

Extended Data Fig. 6. The distribution of AI modeling probabilities of viral RdRPs. a, Distribution of AI modeling probabilities for all RdRPs from known viral superclades (first left column) and representative RdRP superclades (right four columns). **b,** Distribution of AI modeling probabilities for all RdRPs from novel viral superclades (first left column) and representative RdRP superclades (right four columns) captured by BLAST, HMM and the deep learning model. **c,** The distribution of AI modeling probabilities for all RdRPs from novel viral superclades (first left column) and representative RdRP superclades (right four columns) captured by both HMM and the deep learning model. **d**, Distribution of AI modeling probabilities for all AI-specific RdRPs (first left column) and representative RdRP superclades (right four columns) that could only be captured by the deep learning model.

Extended Data Fig. 7. Expression difference of RNA viruses and DNA organisms in our newly sequenced data. a, Abundance comparisons between 58 RNA viral superclades, 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 superclades, capsid sequences from two DNA virus families (*Podoviridae* and *Siphoviridae*), and RT sequences.

Extended Data Fig. 8. Genome coverage of representative genome for RNA viruses and DNA organisms in our newly sequenced data. For 42 RNA viral superclades, four DNA virus families, RT, and cell organism, genomes with high abundance in RNA libraries were chosen to check reads coverage in DNA libraries.

Extended Data Fig. 9. The coverage and abundance of RNA viruses and DNA organisms. The coverage of viral sequences shown as rising with rpm.

Extended Data Fig. 10. Phylogenetic tree of the Nido-like superclade and the genome structure of representatives. The tree was estimated using a maximum likelihood method based on the conserved RdRP domain. The reference sequences reported previously are shaded grey, 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. 11. Association between RNA viral superclades and other non-RdRP protein clusters. Grey and pink circles denote known and novel superclades, respectively. Blue circles denote non-RdRP protein clusters.

Extended Data Fig. 12. Specificity and shareability of RNA viruses. a, Number of specific viral species ("marker" species) in each ecological subtype. b, Association between RNA viruses and different environmental ecotypes. The size of the colored circles indicates the number of viral species identified by each ecotype, while the thickness of the line indicates the number of viral species shared by each ecotype.

Extended Data Fig. 13. 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 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 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 all the RdRP sequences identified in this study.

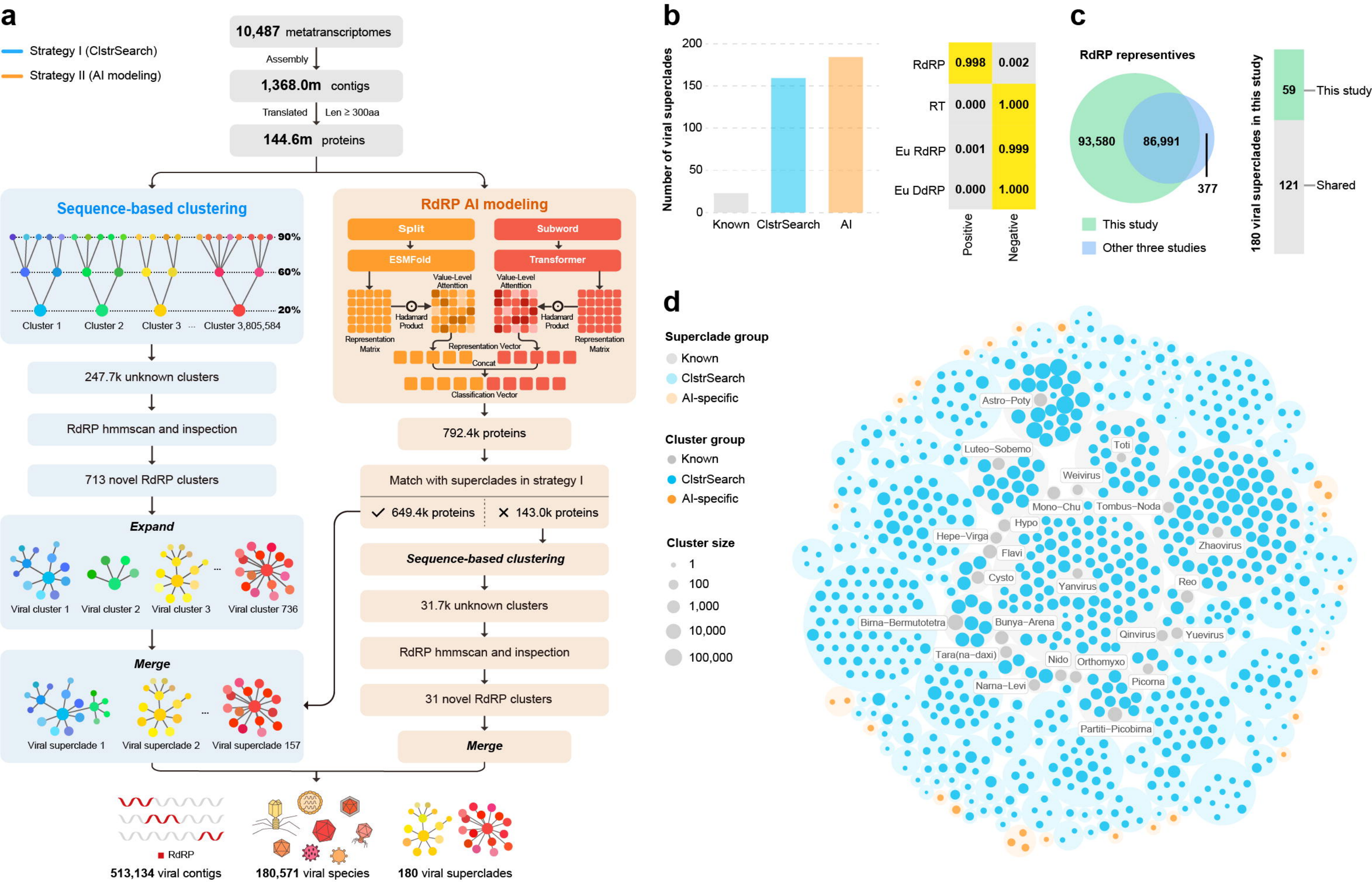
Supplementary Table 4. Taxonomic comparison of known viral superclades between this study and the current RNA virus classification system. New taxonomies that are incompatible with current viral phyla or classes are shown in orange.

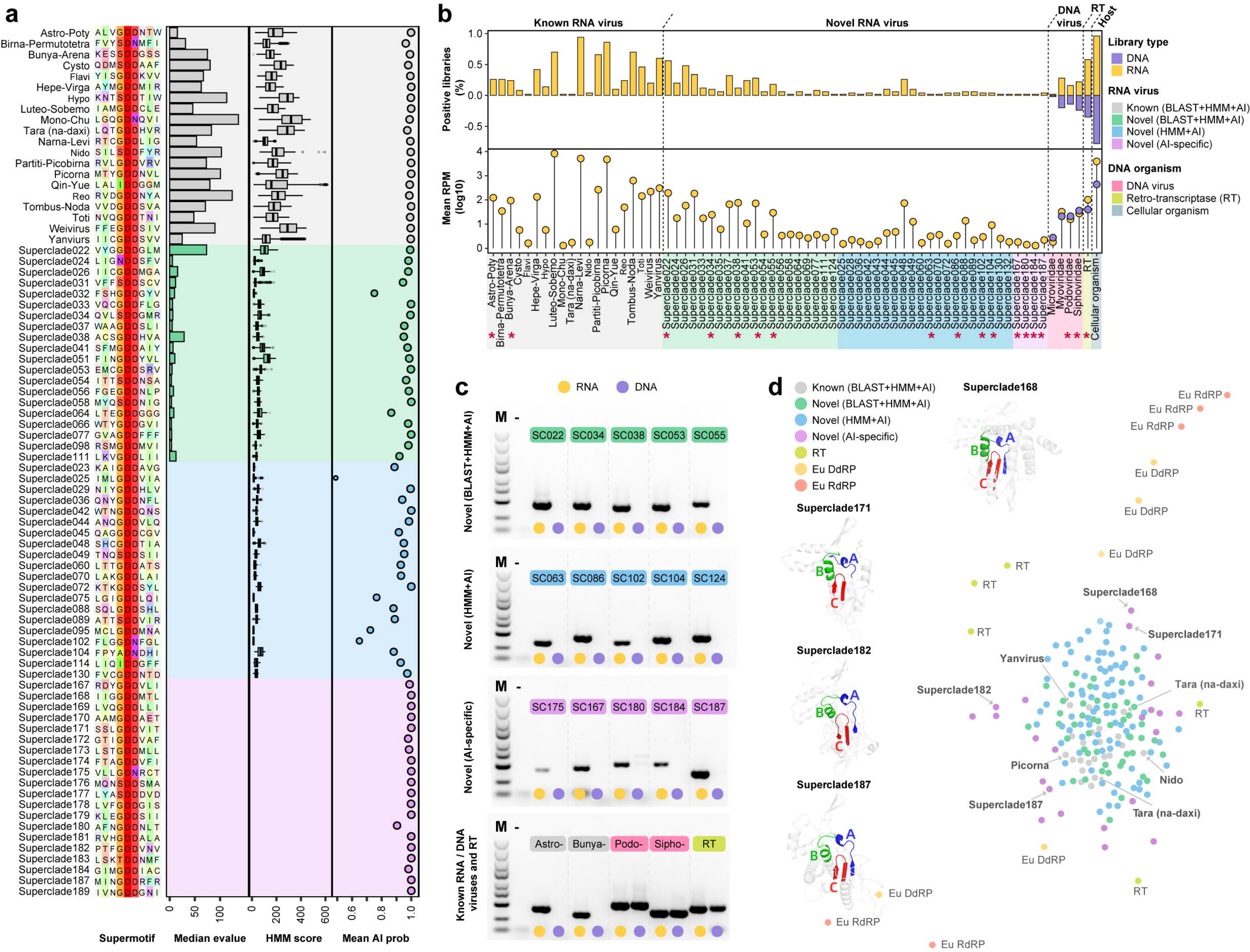
Supplementary Table 5. Predicted results of RdRPs identified in this study using three methods (Threshold: BLAST: e<=1E-3; HMM: score>=10; AI: prob>=0.5).

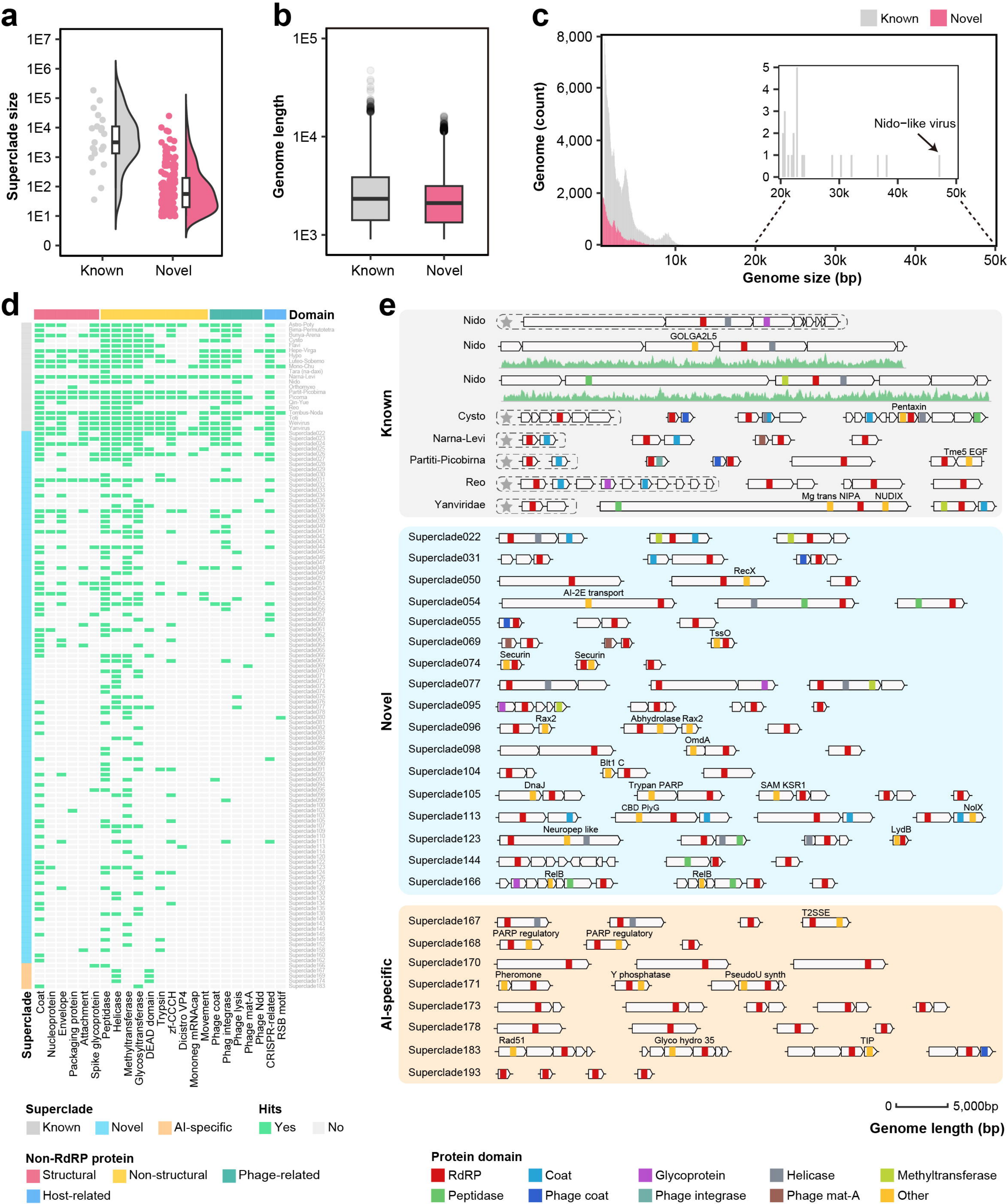
Supplementary Table 6. Size information (the number of contigs) of all viral superclades.

Supplementary Table 7. Distribution of proteins associated with bacterial hosts in viral superclades.

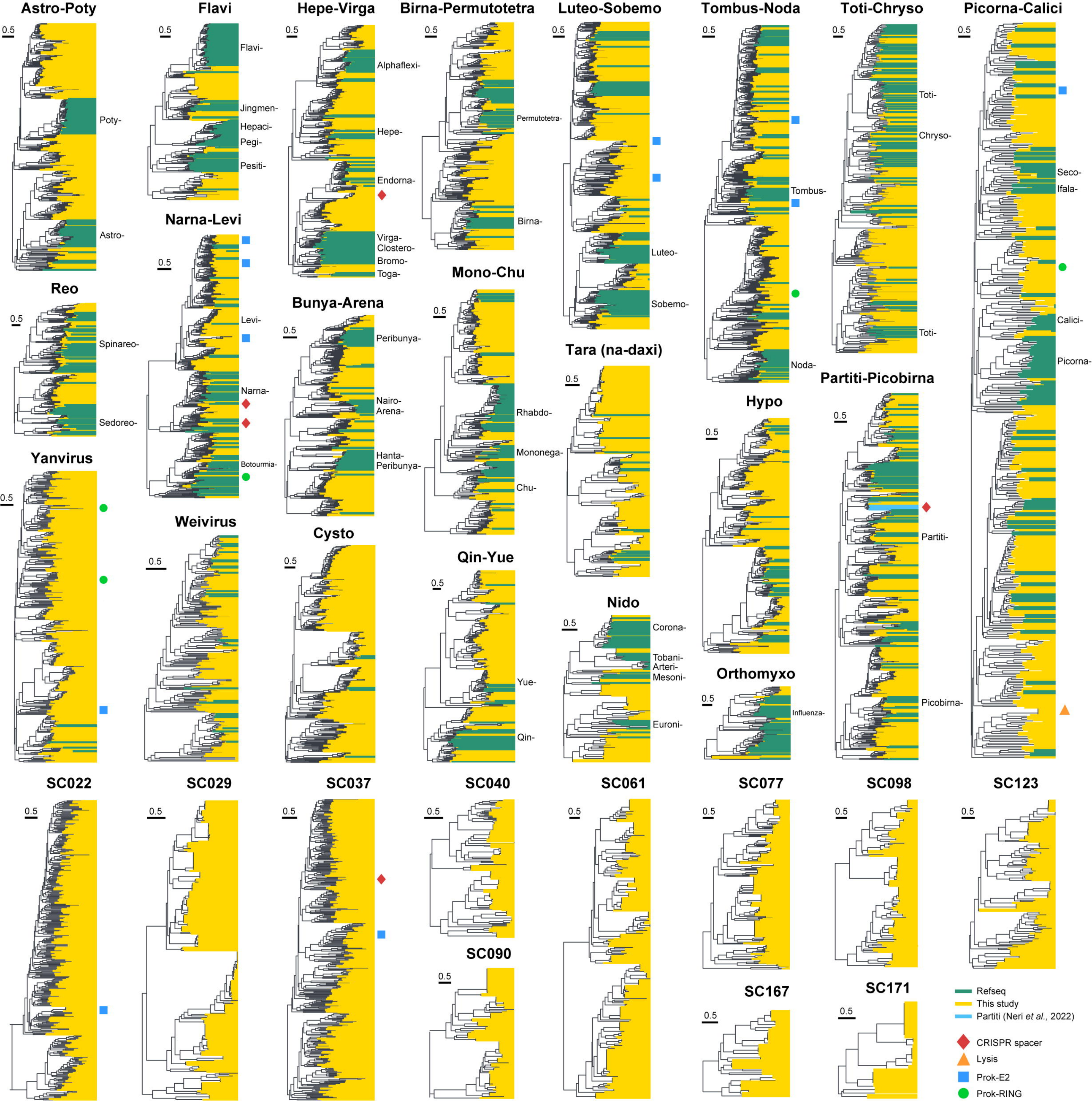
Supplementary Table 8: Normalized abundance levels (measured by RPM) of each viral species in environmental samples (RPM>=1, coverage>=20%).

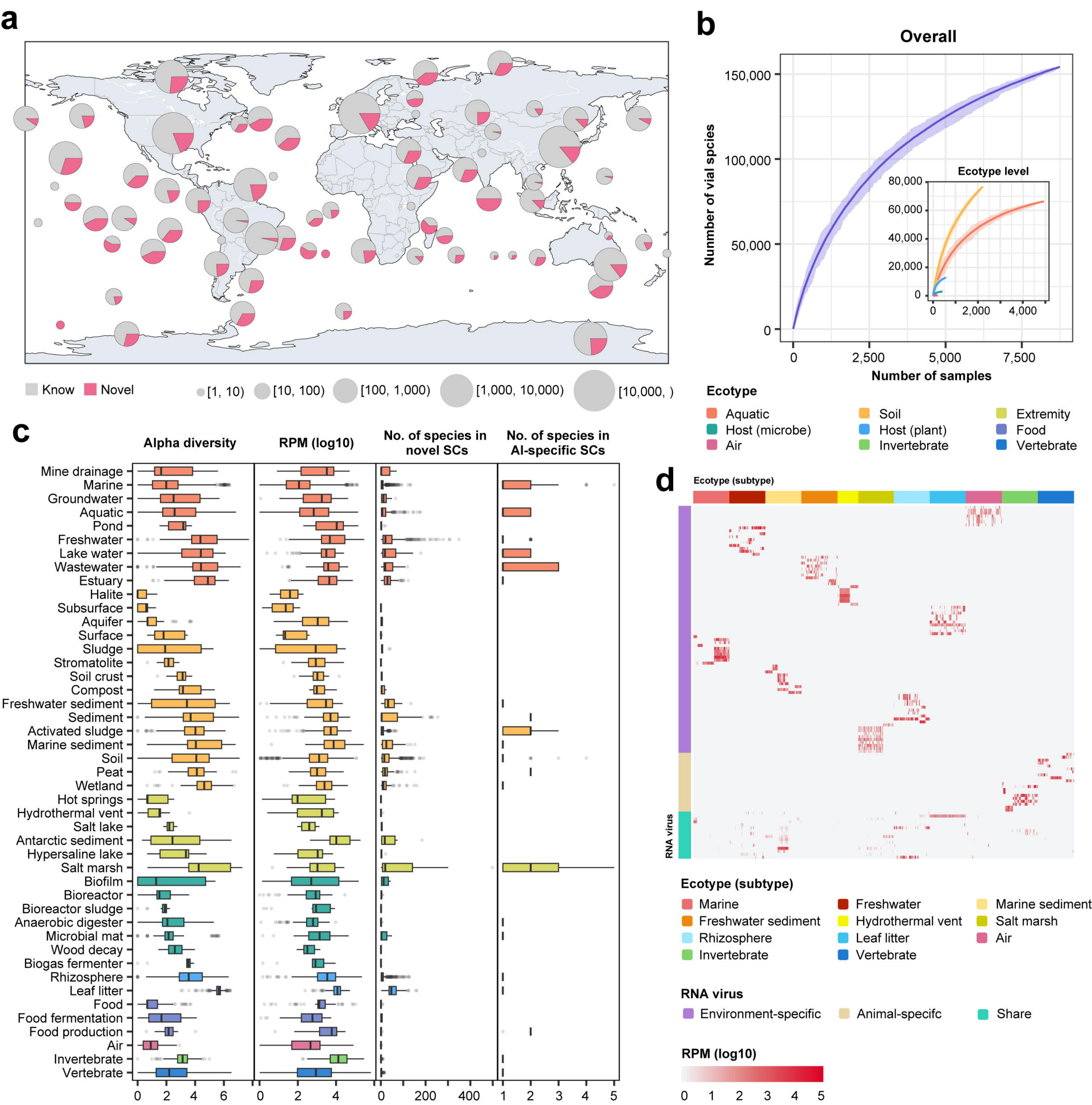

a
Strategy I (ClstrSearch)
Strategy II (AI modeling)
10,487 metatranscriptomes
Assembly
1,368.0m contigs
Translated
Len ≥ 300aa
144.6m proteins
Sequence-based clustering
90%
60%
20%
Cluster 1
Cluster 2
Cluster 3
···
Cluster 3,805,584
247.7k unknown clusters
RdRP hmmscan and inspection
713 novel RdRP clusters
Expand
Viral cluster 1
Viral cluster 2
Viral cluster 3
···
Viral cluster 736
Merge
Viral superclade 1
Viral superclade 2
···
Viral superclade 157
RdRP AI modeling
Split
ESMFold
Subword
Transformer
Value-Level Attenttion
Hadamard Product
Representation Matrix
Representation Vector
Concat
Classification Vector
792.4k proteins
Match with superclades in strategy I
✓ 649.4k proteins
✕ 143.0k proteins
Sequence-based clustering
31.7k unknown clusters
RdRP hmmscan and inspection
31 novel RdRP clusters
Merge
RdRP
513,134 viral contigs
180,571 viral species
180 viral superclades
b
Number of viral superclades
200
150
100
50
0
Known
ClstrSearch
AI
RdRP 0.998 0.002
RT 0.000 1.000
Eu RdRP 0.001 0.999
Eu DdRP 0.000 1.000
Positive
Negative
c
RdRP representives
93,580
86,991
377
This study
Other three studies
180 viral superclades in this study
59
This study
121
Shared
d
Superclade group
Known
ClstrSearch
AI-specific
Cluster group
Known
ClstrSearch
AI-specific
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
Tara(na-daxi)
Nido
Orthomyxo
Picorna
Narna-Levi
Partiti-Picobirna


a
Astro-Poty ALVGDDNTW
Birna-Permutotetra FVYSDNMFI
Bunya-Arena KESSDDGSS
Cysto QDMSDDAAF
Flavi YISGDDKVV
Hepe-Virga AYMGDDMIR
Hypo KNTSDDTIW
Luteo-Sobemo IAMGDDCLE
Mono-Chu LGQGDNQVI
Tara (na-daxi) LQTGDDHVR
Narna-Levi RTCGDDLIG
Nido SILSDDFYR
Partiti-Picobirna RVLGDDVRV
Picorna MTYGDDNVL
Qin-Yue LALIDDGGM
Reo RVDGDDNYA
Tombus-Noda VVDGDDSVA
Toti NVQGDDTNI
Weivirus FFEGDDSIV
Yanvirus IICGDDSVV
Superclade022 VYGGDDGLM
Superclade024 LIGNDDSFV
Superclade026 IICGDDMGA
Superclade031 VFFSDDSCV
Superclade032 FSHGDDGYV
Superclade033 VQCGDDFLG
Superclade034 QVLSDDGMR
Superclade037 WAAGDDSLI
Superclade038 ACSGDDHVA
Superclade041 SFMGDDAIY
Superclade051 FINGDDYVL
Superclade053 EMCGDDSRV
Superclade054 ITTSDDNSA
Superclade056 FGEGDDNLP
Superclade058 MYQSDDNIG
Superclade064 LTEGDDGGG
Superclade066 WTYGDDGVI
Superclade077 GVAGDDFFF
Superclade098 RSMGDDMVI
Superclade111 LKVGDDLII
Superclade023 KAIGDDAVG
Superclade025 IMLGDDVIA
Superclade029 NIYGDDHLV
Superclade036 QNYGDDNFL
Superclade042 WTNGDDQNS
Superclade044 ANQGDDVLQ
Superclade045 QAGGDDCGV
Superclade048 SHCGDDTIA
Superclade049 TNQSDDSII
Superclade060 LTTGDDATS
Superclade070 LAKGDDLAI
Superclade072 KTKGDDSYL
Superclade075 LGIGDDLQI
Superclade088 SQLGDDSHL
Superclade089 ATTSDDVIR
Superclade095 MCLGDDMNA
Superclade102 FLGGDNFGL
Superclade104 FPYADNDHI
Superclade114 LIQIDDGFF
Superclade130 FVCGDDNTD
Superclade167 RDYGDDVLI
Superclade168 IIGGDDMTL
Superclade169 LVQGDDLLI
Superclade170 AAMGDDAET
Superclade171 SSLGDDVIT
Superclade172 GTIGDDVAF
Superclade173 LSTGDDMLL
Superclade174 FTAGDDVFI
Superclade175 VLLGDNRCT
Superclade176 MQNSDDSMA
Superclade177 LYASDDDVD
Superclade178 LVFGDDGIS
Superclade179 KLEGDDSII
Superclade180 AFNGDDNLT
Superclade181 RVHGDDALA
Superclade182 PTFGDDVNV
Superclade183 LSKTDDNMF
Superclade184 GIMGDDIAC
Superclade187 MINGDDRFR
Superclade189 IVNGDDGNI
0 50 100
0 200 400 600
0.6 0.8 1.0
Supermotif
Median evalue
HMM score
Mean AI prob
b
Known RNA virus
Novel RNA virus
DNA virus
RT
Host
Positive libraries (%)
1.0 0.5 0.0 -0.5
Mean RPM (log10)
4 3 2 1 0
Library type
DNA
RNA
RNA virus
Known (BLAST+HMM+AI)
Novel (BLAST+HMM+AI)
Novel (HMM+AI)
Novel (AI-specific)
DNA organism
DNA virus
Retro-transcriptase (RT)
Cellular organism
Astro-Poty
Birna-Permutotetra
Bunya-Arena
Cysto
Flavi
Hepe-Virga
Hypo
Luteo-Sobemo
Mono-Chu
Tara (na-daxi)
Narna-Levi
Nido
Partiti-Picobirna
Picorna
Qin-Yue
Reo
Tombus-Noda
Toti
Weivirus
Yanvirus
Superclade022
Superclade024
Superclade026
Superclade031
Superclade033
Superclade034
Superclade035
Superclade037
Superclade038
Superclade041
Superclade053
Superclade054
Superclade055
Superclade056
Superclade058
Superclade064
Superclade069
Superclade077
Superclade111
Superclade124
Superclade025
Superclade028
Superclade036
Superclade042
Superclade043
Superclade044
Superclade045
Superclade048
Superclade049
Superclade060
Superclade063
Superclade070
Superclade072
Superclade086
Superclade088
Superclade089
Superclade102
Superclade104
Superclade130
Superclade132
Superclade167
Superclade180
Superclade184
Superclade187
Microviridae
Myoviridae
Podoviridae
Siphoviridae
RT
Cellular organism
*
c
RNA
DNA
Novel (BLAST+HMM+AI)
M -
SC022 SC034 SC038 SC053 SC055
Novel (HMM+AI)
M -
SC063 SC086 SC102 SC104 SC124
Novel (AI-specific)
M -
SC175 SC167 SC180 SC184 SC187
Known RNA / DNA viruses and RT
M -
Astro- Bunya- Podo- Sipho- RT
d
Known (BLAST+HMM+AI)
Novel (BLAST+HMM+AI)
Novel (HMM+AI)
Novel (AI-specific)
RT
Eu DdRP
Eu RdRP
Superclade168
Superclade171
Superclade182
Superclade187
A
B
C
Eu RdRP
Eu DdRP
RT
Superclade168
Superclade171
Yanvirus
Tara (na-daxi)
Superclade182
Picorna
Nido
Superclade187
Tara (na-daxi)



a
Superclade size
1E7
1E6
1E5
1E4
1E3
1E2
1E1
0
Known
Novel
b
Genome length
1E5
1E4
1E3
Known
Novel
c
Known
Novel
Genome (count)
8,000
6,000
4,000
2,000
0
10k
20k
30k
40k
50k
Genome size (bp)
5
4
3
2
1
0
20k
30k
40k
50k
Nido−like virus
d
Domain
Astro-Poty
Birna-Permutotetra
Bunya-Arena
Cysto
Flavi
Hepe-Virga
Hypo
Luteo-Sobemo
Mono-Chu
Tara (na-daxi)
Narna-Levi
Nido
Orthomyxo
Partit-Picobirna
Picorna
Qin-Yue
Reo
Tombus-Noda
Toti
Weivirus
Yanvirus
Superclade022
Superclade023
Superclade024
Superclade025
Superclade026
Superclade027
Superclade028
Superclade029
Superclade030
Superclade031
Superclade032
Superclade033
Superclade034
Superclade035
Superclade036
Superclade037
Superclade038
Superclade039
Superclade040
Superclade041
Superclade042
Superclade043
Superclade044
Superclade045
Superclade046
Superclade047
Superclade048
Superclade049
Superclade050
Superclade051
Superclade052
Superclade053
Superclade054
Superclade055
Superclade056
Superclade057
Superclade058
Superclade060
Superclade061
Superclade062
Superclade063
Superclade064
Superclade065
Superclade066
Superclade067
Superclade069
Superclade070
Superclade071
Superclade072
Superclade073
Superclade074
Superclade075
Superclade076
Superclade077
Superclade078
Superclade080
Superclade081
Superclade082
Superclade083
Superclade084
Superclade085
Superclade086
Superclade087
Superclade089
Superclade090
Superclade091
Superclade092
Superclade093
Superclade094
Superclade095
Superclade098
Superclade099
Superclade100
Superclade102
Superclade103
Superclade105
Superclade107
Superclade109
Superclade110
Superclade111
Superclade113
Superclade114
Superclade120
Superclade122
Superclade123
Superclade124
Superclade126
Superclade127
Superclade128
Superclade130
Superclade132
Superclade134
Superclade135
Superclade138
Superclade140
Superclade143
Superclade144
Superclade145
Superclade148
Superclade152
Superclade158
Superclade160
Superclade162
Superclade166
Superclade167
Superclade169
Superclade174
Superclade183
Superclade
Coat
Nucleoprotein
Envelope
Packaging protein
Attachment
Spike glycoprotein
Peptidase
Helicase
Methyltransferase
Glycosyltransferase
DEAD domain
Trypsin
zf-CCCH
Dicistro VP4
Mononeg mRNAcap
Movement
Phage coat
Phag integrase
Phage lysis
Phage mat-A
Phage Ndd
CRISPR-related
RSB motif
e
Known
Nido
Nido
GOLGA2L5
Nido
Cysto
Pentaxin
Narna-Levi
Partiti-Picobirna
Tme5 EGF
Reo
Yanviridae
Mg trans NIPA
NUDIX
Novel
Superclade022
Superclade031
Superclade050
RecX
Superclade054
AI-2E transport
Superclade055
Superclade069
TssO
Superclade074
Securin
Securin
Superclade077
Superclade095
Superclade096
Rax2
Abhydrolase Rax2
Superclade098
OmdA
Superclade104
Blt1 C
Superclade105
DnaJ
Trypan PARP
SAM KSR1
Superclade113
CBD PlyG
NolX
Superclade123
Neuropep like
LydB
Superclade144
Superclade166
RelB
RelB
AI-specific
Superclade167
T2SSE
Superclade168
PARP regulatory
PARP regulatory
Superclade170
Superclade171
Pheromone
Y phosphatase
PseudoU synth
Superclade173
Superclade178
Superclade183
Rad51
Glyco hydro 35
TIP
Superclade193
0
5,000bp
Genome length (bp)
Superclade
Known
Novel
AI-specific
Hits
Yes
No
Non-RdRP protein
Structural
Non-structural
Phage-related
Host-related
Protein domain
RdRP
Coat
Glycoprotein
Helicase
Methyltransferase
Peptidase
Phage coat
Phage integrase
Phage mat-A
Other


Astro-Poty
0.5
Poty-
Astro-
Flavi
0.5
Flavi-
Jingmen-
Hepaci-
Pegi-
Pesiti-
Hepe-Virga
0.5
Alphaflexi-
Hepe-
Endorna-
Virga-
Clostero-
Bromo-
Toga-
Birna-Permutotetra
0.5
Permutotetra-
Birna-
Luteo-Sobemo
0.5
Luteo-
Sobemo-
Tombus-Noda
0.5
Tombus-
Noda-
Toti-Chryso
0.5
Toti-
Chryso-
Toti-
Picorna-Calici
0.5
Seco-
Ifala-
Calici-
Picorna-
Reo
0.5
Spinareo-
Sedoreo-
Narna-Levi
0.5
Levi-
Narna-
Botourmia-
Bunya-Arena
0.5
Peribunya-
Nairo-
Arena-
Hanta-
Peribunya-
Mono-Chu
0.5
Rhabdo-
Mononega-
Chu-
Tara (na-daxi)
0.5
Hypo
0.5
Partiti-Picobirna
0.5
Partiti-
Picobirna-
Yanvirus
0.5
Weivirus
0.5
Cysto
0.5
Qin-Yue
0.5
Yue-
Qin-
Nido
0.5
Corona-
Tobani-
Arteri-
Mesoni-
Euroni-
Orthomyxo
0.5
Influenza-
SC022
0.5
SC029
0.5
SC037
0.5
SC040
0.5
SC090
0.5
SC061
0.5
SC077
0.5
SC167
0.5
SC098
0.5
SC171
0.5
SC123
0.5
Refseq
This study
Partiti (Neri et al., 2022)
CRISPR spacer
Lysis
Prok-E2
Prok-RING


a
Know
Novel
[1, 10)
[10, 100)
[100, 1,000)
[1,000, 10,000)
[10,000,)
b
Overall
Nunmber of vial spcies
150,000
100,000
50,000
0
0
2,500
5,000
7,500
Number of samples
Ecotype level
80,000
60,000
40,000
20,000
0
0
2,000
4,000
Ecotype
Aquatic
Host (microbe)
Air
Soil
Host (plant)
Invertebrate
Extremity
Food
Vertebrate
c
Alpha diversity
RPM (log10)
No. of species in novel SCs
No. of species in AI-specific SCs
Mine drainage
Marine
Groundwater
Aquatic
Pond
Freshwater
Lake water
Wastewater
Estuary
Halite
Subsurface
Aquifer
Surface
Sludge
Stromatolite
Soil crust
Compost
Freshwater sediment
Sediment
Activated sludge
Marine sediment
Soil
Peat
Wetland
Hot springs
Hydrothermal vent
Salt lake
Antarctic sediment
Hypersaline lake
Salt marsh
Biofilm
Bioreactor
Bioreactor sludge
Anaerobic digester
Microbial mat
Wood decay
Biogas fermenter
Rhizosphere
Leaf litter
Food
Food fermentation
Food production
Air
Invertebrate
Vertebrate
0
2
4
6
0
2
4
0
200
400
1
2
3
4
5
d
Ecotype (subtype)
RNA virus
Ecotype (subtype)
Marine
Freshwater
Marine sediment
Freshwater sediment
Hydrothermal vent
Salt marsh
Rhizosphere
Leaf litter
Air
Invertebrate
Vertebrate
RNA virus
Environment-specific
Animal-specifc
Share
RPM (log10)
0
1
2
3
4
5