VirID: Beyond Virus Discovery - An Integrated Platform for Comprehensive RNA Virus Characterization

Ziyue Yang[#,1,2,3], Yongtao Shan[#,1,2,3], Xue Liu[1,2,3], Guowei Chen[4], Yuanfei Pan[5], Qinyu Gou[1,2,3], Jie Zou[1,2,3], Zilong Chang[1,2,3], Qiang Zeng[1,2,3], Chunhui Yang[1,2,3], Jianbin Kong[1,2,3], Yanni Sun[4], Shaochuan Li[6], Xu Zhang[6], Wei-Chen Wu[1,2,3], Chunmei Li [1,2,3], Hong Peng[1,2,3], Edward C. Holmes[7,8], Deyin Guo*[9,10], Mang Shi*[1,2,3,11]

[1]State Key Laboratory for Biocontrol, School of Medicine, Shenzhen Campus of Sun Yat-sen University, Sun Yat-sen University, Shenzhen, China.

[2]National Key Laboratory of Intelligent Tracking and Forecasting for Infectious Diseases, Sun Yat-sen University, Shenzhen, China.

[3]Shenzhen Key Laboratory for Systems Medicine in Inflammatory Diseases, Shenzhen Campus of Sun Yat-sen University, Sun Yat-sen University, Shenzhen, China.

[4]Department of Electrical Engineering, City University of Hong Kong, 83 Tat Chee Avenue, Kowloon, Hong Kong (SAR), China.

[5]Ministry of Education Key Laboratory of Biodiversity Science and Ecological Engineering, School of Life Sciences, Fudan University, Shanghai, China.

[6]Goodwill Institute of Life Sciences, Guangzhou, China.

[7]School of Medical Sciences, The University of Sydney, Sydney, New South Wales, Australia.

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

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

[10]State Key Laboratory of Respiratory Disease, National Clinical Research Center. for Respiratory Disease, Guangzhou Institute of Respiratory Health, The First Affiliated Hospital of Guangzhou Medical University, Guangzhou, Guangdong, China.

[11]Guangdong Provincial Center for Disease Control and Prevention, Guangzhou, China.

[#]These authors contributed equally to this work

Corresponding author: guodeyin@mail.sysu.edu.cn

Corresponding author: shim23@mail.sysu.edu.cn

ABSTRACT

RNA viruses exhibit vast phylogenetic diversity and can significantly impact public health and agriculture. However, current bioinformatics tools for viral discovery from metagenomic data frequently generate false positive virus results, overestimate viral diversity, and misclassify virus sequences. Additionally, current tools often fail to determine virus-host associations, which hampers investigation of the potential threat posed by a newly detected virus. To address these issues, we developed VirID, a software tool specifically designed for the discovery and characterization of RNA viruses from metagenomic data. The basis of VirID is a comprehensive RNA-dependent RNA polymerase (RdRP) database to enhance a workflow that includes RNA virus discovery, phylogenetic analysis, and phylogeny-based virus characterization. Benchmark tests on a simulated data set demonstrated that VirID had high accuracy in profiling viruses and estimating viral richness. In evaluations with real-world samples, VirID was able to identity RNA viruses of all type, but also provide accurate estimations of viral genetic diversity and virus classification, as well as comprehensive insights into associations with humans, animals, and plants. VirID therefore offers a robust tool for virus discovery and serves as a valuable resource in basic virological studies, pathogen surveillance, and early warning systems for infectious disease outbreaks.

KEYWORDS

RNA virus; Virus discovery; Phylogenetic analysis; Virome; Emerging pathogens

53 **INTRODUCTION**

54 RNA viruses are renowned for their genetic and phenotypic diversity and ability to
55 infect hosts ranging from animals, plants, fungi, to microbial organisms, sometimes
56 with devastating health and economic consequences (1). RNA viruses have caused
57 human epidemics for millenia, with notable recent examples including human
58 immunodeficiency virus (HIV; 1981) (2), SARS-CoV (2002) (3), pandemic H1N1
59 influenza (2009) (4), MERS-CoV (2012) (5), Ebola virus (Western Africa, 2013) (6),
60 Zika virus (2015) (7) and most recently SARS-CoV-2 (2019) (8). Importantly,
61 approximately 70% of these pandemic-causing viral pathogens originate from wildlife
62 animals or utilize arthropod vectors (9), and many have evaded surveillance systems
63 before their emergence in human populations (10-13). Additionally, RNA viruses pose
64 significant threats to agriculture, particularly as epidemics in domestic animals and
65 crops can jeopardize global food security (14-19). Understanding the diversity of RNA
66 viruses and enhancing surveillance of those that pose threats to humans and
67 economically important species are therefore endeavors of utmost importance (20,21).

68 Meta-transcriptomics (i.e., total RNA sequencing) provides a potentially unbiased
69 survey of the genetic information from all types of organisms in biological samples and
70 has transformed the detection and characterization of RNA viruses (22-26). This
71 method provides efficiency and breadth in virus discovery compared to traditional
72 cultivation techniques (27-30), which are often restricted by their reliance on cell
73 culture growth (31), and to PCR-based approaches that depend on prior knowledge of
74 existing viral diversity (32). As a consequence, meta-transcriptomics has become the
75 primary tool for discovering RNA viruses (26), with the highly conserved RNA-
76 dependent RNA polymerase (RdRP) that is essential for RNA virus replication, serving
77 a powerful universal genetic marker (33-38).

78 As well as sequencing, the bioinformatics tools for detecting and characterizing
79 virus sequences within metagenomic data sets have experienced major improvements.
80 Initial workflow tools like ViromeScan (39) and Taxonomer (40) detected viruses in
81 samples by analyzing sequencing reads. Subsequent tools, such as Vipie (41), ID-seq

82 (42), Lazypipe2 (43), and ViWrap (44), use contigs from *de novo* assemblies for virus
83 identification, enhancing the length of query sequences and aiding the discovery of
84 divergent viruses. With the advent of machine learning and deep learning, tools such as
85 PhiSpy (45), VirSorter (46), VirSorter2 (47), and VIBRANT (48) were developed to
86 differentiate viral from microbial sequences by analyzing gene and/or genomic
87 sequences and features, although with relatively high false positive rates. Methods like
88 VirFinder (49) and DeepVirFinder (50) utilize the frequency of consecutive nucleotides
89 (k-mers) in known viral and cellular genomes to identify viral sequences. More recently,
90 specialized RNA virus discovery software such as VirBot (51) have been developed.
91 VirBot constructs databases of RNA virus protein families and employs profile Hidden
92 Markov Models (pHMM) to identify distantly related viral sequences, exhibiting
93 superior performance to other virus classification tools.

94 Despite these advancements, the bioinformatic tools for virus discovery continue
95 to confront major challenges, such as a high false positive rate, the overestimation of
96 viral diversity, and inaccurate virus classification (52,53). These issues are particularly
97 problematic for RNA viruses due to their highly divergent genomes and distinct
98 genomic structures which complicate identification using basic annotation methods
99 (54,55). Additionally, current bioinformatic tools often lack the capability for the in-
100 depth characterization of viruses, including the accurate identification of host
101 associations, which impedes an understanding of their potential threat to public health
102 and agricultural systems.

103 In response to these challenges, we developed VirID – a comprehensive and user-
104 friendly software tool tailored for the discovery and characterization of RNA viruses
105 from metagenomic data. VirID features a robust database of RdRP sequences and
106 comprises three core modules: (i) RNA virus discovery, (ii) phylogenetic analysis, and
107 (iii) phylogeny-based virus characterization. Benchmark tests on simulated data set
108 revealed VirID's high accuracy in profiling and classifying RNA viruses. In practical
109 applications with real-world data sets, VirID demonstrated its capacity for virus
110 discovery and for conducting thorough virome analyses.
111

112 **MATERIALS AND METHODS**

113 **RdRP database**

114 We established a database of 7080 representative RdRP protein sequences for RNA
115 virus discovery and characterization. The RdRP sequences were derived from three
116 independent sources, including a "backbone RdRP data set" (*N* = 5,384)(28,56), the
117 NCBI RefSeq database (viral RdRP associated, *N* = 19,574), and the NCBI GenBank
118 database (viral RdRP associated, *N* = 5,710,331). RdRPs from the RefSeq and GenBank
119 databases were initially identified based on key annotation terms, including "RdRP",
120 "RNA-dependent RNA polymerase" and "polymerase" under the taxonomy "Ribovira".
121 Highly divergent RdRPs derived from environmental samples (29,37,57) were
122 excluded due to lack of confirmation. The RdRPs from NCBI were then compared to
123 the backbone RdRP data set using Diamond v2.0.15.153 (58) to remove false positives.
124 The presence of key RdRP domains, specifically the highly conserved A, B and C
125 sequence motifs, were examined using palmscan v1.0.i86linux64 (59). The remaining
126 sequences were clustered using CD-HIT v4.8.1 (60) at 80% amino acid identity, with a
127 representative then selected from each cluster. This resulted in a final reference database
128 of 7080 sequences than were then systematically organized into 24 RNA virus
129 "superclades" based on broad phylogenetic relationships (28,56), and subsequently into
130 5 phyla, 20 classes, 28 orders, and 112 families based on both the phylogenetic
131 relationships and ICTV taxonomy (Figure 1A, Supplementary Table S1).

132 For each of the RdRP sequences, information on host organism was initially
133 retrieved from the Virus-Host Database (61) and confirmed by checking the original
134 publications. Based on the host information and phylogenetic relationships, we
135 categorized the RdRP sequences into four broad groups: (i) those infecting humans
136 (including vector-borne viruses, *N* = 188), (ii) those infecting vertebrate animals (*N*
137 =363), (ii) those infecting plants (*N* = 325), and (iv) all other host associations (Figure
138 1B). This information, together virus phylogenetic relationships, was used to define
139 specific host groups on phylogenetic trees (Figure 1C, Supplementary Figure S1) from

which the host-association of the newly identified RdRP sequences could be inferred (See below).

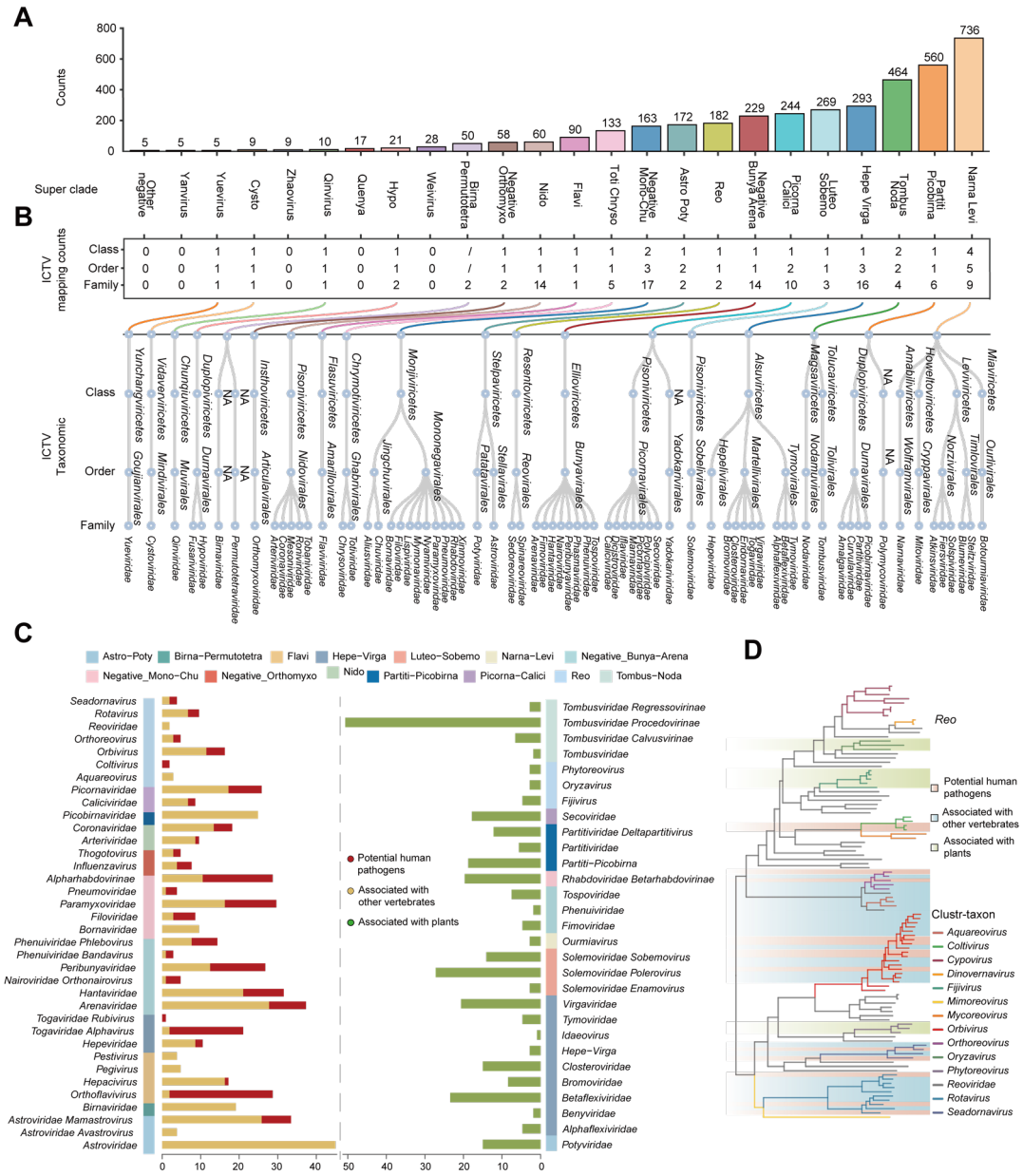


Figure 1. Overview of the RdRP Reference Database. (**A**) Compilation of representative RdRP sequences into 24 RNA virus superclades, categorized based on phylogenetic analysis. (**B**) Alignment of viral superclades with the ICTV taxonomic system, spanning virus class, order, and family levels. Numbers reflect the count of ICTV classifications mapped to each RdRP superclade, with specific ICTV representative mapping details provided. (**C**) The distribution of viral sequences across supergroups and ICTV taxonomy, highlighting 551 vertebrate-related and 325 plant-related sequences. Data points are color-coded to represent host associations, and bar plots show the proportions of human pathogens and plant-associated viruses in each taxonomic group. (**D**) Using the Reoviridae as a case study, the diagram presents a phylogenetic tree in which branch colors indicate taxonomy and background hues denote host association. Reference sequences are clearly annotated based on their positions within the phylogenetic tree.

154 **Processing and assembly of sequencing reads**

155 VirID performs an initial processing of the input read sequences. The quality
156 control of reads was conducted using bbduk.sh (62). To remove host and microbial
157 ribosomal RNA (rRNA) sequences, the reads were then mapped, using Bowtie2 v 2.5.1
158 (63), against a reference rRNA database that contained a total of 505,405 rRNA non-
159 repetitive sequences obtained from SILVA database v138.1 (64) and the RDP-II
160 database (65). The remaining reads were assembled *de novo* into contigs using
161 MEGAHIT v1.2.9 (66) and under default parameters. Only contigs longer than 600bp
162 were retained for subsequent virus discovery and characterization.

163

164 **Discovery and quality control of RNA virus contigs**

165 VirID discovers RNA viral contigs through a homology-based search approach.
166 The assembled contigs were first compared against the RdRP reference database using
167 the Diamond blastx program v2.0.15.153 (58) to identify potential RNA viral sequences.
168 To remove false positives, potential viral contigs were subsequently compared against
169 the NCBI Non-Redundant Protein Database (NR) and the hits were annotated using
170 TaxonKit v0.14.2(67). Contigs with only non-viral hits were removed, and the top hits
171 of the remaining contigs were used to provide an initial taxonomic annotation for the
172 viral contigs.

173 In addition to the standard quality control steps described above, VirID has extra
174 procedures to identify and remove false positives and erroneous sequences. To filter out
175 potential endogenous virus elements (EVEs) with disrupted ORFs, we removed contigs
176 whose RdRP-associated open reading frames (ORFs) were predicted to contain less
177 than 200 amino acids by scanning the viral sequence in six possible reading frames. To
178 identify and control for misassembled contigs that contained both viral and non-viral
179 sequences, the sequence was first compared using BLASTn (68) against a "non-viral"
180 database, specifically a sub-set of NCBI Nucleotide Sequence Database (NT) that
181 excludes RNA viral sequences. Based on the position and length of hit, the target
182 contigs were either partially corrected (i.e., in scenarios when "non-viral" regions

183 appear at either end of the contigs and were shorter than the "viral" region) or
184 completely removed (all other cases).
185
186 **RNA virus species identification and quantification**
187 For viral species identification, the RdRP-associated contigs were further clustered
188 based on 80% sequence identity and 40% sequence length to form species-level
189 taxonomic units. For each species, the predicted RdRPs were compared against the NR
190 database, and those with >90% identity and 70% coverage to existing RNA viral species
191 were denoted "known" viral species, whereas those below these thresholds were
192 considered as "new" viral species.
193 The sequence Reads Per Million (RPM) metric was used to evaluate the relative
194 abundance of virus species within the sample. Clean reads are first mapped to all contigs
195 associated with a virus species using Bowtie2 v2.5.1(63). The number of mapped reads
196 is multiplied by 10^6 and divided by the total number of reads to give the RPM value.
197
198 $$RPM = \frac{MappedReads * 10^6}{TotalReads}$$
199
200 **Viral sequence alignment at the superclade level**
201 For newly identified viral sequences, multiple alignment was performed before
202 subsequent phylogenetic placement. Viral RdRP sequences were first classified into 24
203 superclades for sequence alignment and phylogenetic analysis, based on BLASTx
204 comparisons (68) against the RdRP reference database. VirID utilizes the Amino Acid
205 Consistency (AAC) index to quantify the similarity between two amino acid sequences.
206 The AAC is calculated by dividing the number of identical, non-empty amino acids by
207 the total length of the amino acid sequence (Extended Data Figure 1).
208 To ensure the accuracy of multiple sequence alignments, we employed an iterative
209 approach with Mafft v7.520 (69), followed by the removal of ambiguously aligned
210 regions using trimAl v1.4 (70). In each iteration, for each reference sequence (R^i), its
211 amino acid identity r^i is determined by the highest amino acid identity value

compared to other reference sequences. For each potential viral sequence (Q^i), its amino acid identity q^i is set as the maximum value when compared to all reference sequences. If q^i falls below the minimum value of R across the entire reference sequence set, the sequence Q^i is removed as a poor-quality sequence, and the alignment process is repeated. The pseudocode for eliminating false positives through multiple sequence alignment is provided in Supplementary Figure S2.

$$r^i = \max_{i \neq j} AAC(R^i, R^j)$$

$$q^i = \max_{k \in RefList} AAC(Q^i, R^K)$$

Phylogenetic placement

To optimize phylogenetic tree inference, we implemented an approach based on reference trees. We first estimated reference maximum likelihood trees using the PROTGAMMALGF model in RaxML v7 (71). Subsequently, new sequences were integrated into these pre-built reference trees based on the alignment using the LG model in pplacer v1.1 (72), which accurately identifies their most likely positions. The output from pplacer, provided in JSON format, is then converted into the Newick tree format using guppy v1.1 (72), facilitating further phylogenetic analyses.

To evaluate the reliability of our phylogenetic placement, we performed leave-one-out cross-validation, focusing on tree recall, which measures the proportion of correctly recovered branches. For this analysis, we only included branches supported by a bootstrap confidence value exceeding 90%. We observed that the distribution of recall rates varied across superclades, likely influenced by the number of reference sequences and their similarity. In superclusters containing more than 50 sequences, recall rates for most clusters tended to concentrate around 0.75, with few deviations (Supplementary Figure S3). Additionally, in virus clades with a high density of sequences, such as Picorna-Calici and Partiti-Picobirna, this placement method maintained the fundamental stability of the tree structure, demonstrating robustness despite the addition of numerous new sequences.

$$Tree\ recall = \frac{Edge_{both}}{Edge_{\text{reference}}}$$

Classification and host association inference based on phylogenetic analysis

The classification and host association of a query genome were inferred based on its placement within each RdRP superclade tree. We systematically annotated the evolutionary tree by traversing from the leaf nodes upwards, adhering to the principle that sibling nodes share the same labels. The labeling process for each leaf node followed the following criteria: (i) if the node corresponds to a known sequence, its label is assigned based on the existing annotation; (ii) if its sibling node is a known sequence, the leaf node inherits the label of its sibling; and (iii) if its sibling node is an internal node, we iteratively determined the label of that internal node before assigning it to the leaf node. For internal nodes, we conducted a systematic top-down traversal of all their daughter nodes. If all daughter nodes carry identical labels, the internal node adopts the same label as its descendants. However, if the daughter nodes have differing labels, the internal node is categorized under the lowest common taxonomic label for all those leaf nodes in this superclade. This structured approach ensures consistent and logical assignment of labels throughout each superclade tree.

Each potential viral sequence is assigned two types of labels: a taxonomic label that spans superclade-taxon to ICTV hierarchical classification and eventually down to the species level, and a host association label that categorizes the virus as related to humans, vertebrates, or plants. During the phylogenetic tree annotation, host association tags (represented as 0 or 1) are assigned to sequences. Specifically, if a viral sequence's sibling node is identified as a human pathogen, the sequence is considered potentially relevant to humans only if the similarity exceeds 80%. For the ICTV classification within the phylogenetic tree, the label of each internal node is determined by identifying the longest common prefix among the ICTV labels of all its daughter nodes. For sequences that remain unclassified, a default taxonomic hierarchy is assigned within the corresponding superclade. This classification system effectively categorizes potential viral sequences up to the genus level. For well-documented

viruses, species-level classification is determined using the NCBI annotation file derived from comparisons within the NR database. This rigorous approach ensures accurate and detailed categorization of viral sequences across various levels of taxonomic and host associations.

Benchmarking Preliminary Viral Contig Screening on Simulated Data set

To evaluate the performance of VirID in the preliminary screening of RNA viral sequences, we constructed a simulated short-read shotgun metagenomic data set for benchmarking using CAMISIM (73). This data set included 15 samples from three categories: 'known RNA viruses', 'new RNA viruses', and 'others', with each category comprising 15 species. The 'known RNA viruses' and 'others' data sets were derived from 463 complete RNA virus genomes and 55 complete non-RNA virus genomes available in the NCBI RefSeq database, respectively (Supplementary Figure S4). The 'new RNA viruses' were subdivided into three levels based on their sequence similarity to the reference database and containing five species each: 'new RNA virus L1', 'new RNA virus L2', and 'new RNA virus L3', with similarity thresholds set at $0.70 \leq$ similarity (L1) ≤ 0.90, $0.40 \leq$ similarity (L2) < 0.70, and $0.20 \leq$ similarity (L3) < 0.40, respectively. The sequences used to simulate 'new RNA viruses' were sourced from 299 high-quality RNA virus sequences obtained from recent publications (57,74-76), ensuring a comprehensive and challenging test environment.

The individual simulated data samples are provided in pair-end fastq format, with each file approximately 20 GB in size and each read 150 base pairs long. The distribution of the three data types within each sample is maintained at a consistent ratio of 1:1:1. Unlike other tools that require contigs, VirID accepts reads directly as input. To ensure consistent data standards across all tools, the contigs generated from VirID's intermediate outputs are used as input for the other tools. Additionally, for uniformity in evaluation metrics, all tools have adopted VirID's criterion of using contigs longer than 600bp.

298 We assessed the performance of six different viral detection tools—VirFinder,
299 DeepVirFinder, VirSorter, VirSorter2, VirBot, and VirID—on a simulated data set using
300 accuracy, precision, recall, and F1 score. These metrics focus on the number of correctly
301 identified species in the tool's output, rather than the number of sequences. The
302 accuracy of each tool reflects the proportion of RNA viral species correctly identified
303 from the total number of species. Recall measures the tool's sensitivity by calculating
304 the proportion of RNA viral species detected relative to the total number of RNA viral
305 species in the data set. Precision indicates the specificity of the tool, defined as the
306 proportion of RNA viral species detected out of the total number of species identified
307 by the tool. The F1 score, the harmonic mean of precision and recall, provides a
308 balanced measure of a tool's overall effectiveness.
309

310 **Metagenomic data sets**

311 To demonstrate VirID's visualization capabilities, we analyzed 20 SRA libraries
312 from a large scale animal study conducted in China (76) that comprised 14 species
313 across five provinces. In addition, we assessed the versatility of VirID through the
314 analysis of 192 public libraries from seven studies, which included a wide range of
315 samples from human swabs (77) and diverse animal tissues (78-80), as well as
316 arthropod (81), plant (82), and soil samples (83). The public metagenomes utilized in
317 this analysis were sourced from NCBI, with details provided in Supplementary Table
318 S2, S3.
319

320 **Parameters and visualizations**

321 VirID offers three user-selectable modes: 'end to end', 'assembly and basic
322 annotation', and 'phylogenetic analysis'. The 'end to end' mode executes both 'assembly
323 and basic annotation' and 'phylogenetic analysis' sequentially. The 'assembly and basic
324 annotation' mode processes sequencing data up to the point of producing annotated
325 contigs, but it does not include phylogenetic analyses. Conversely, the 'phylogenetic
326 analysis' mode begins with contigs and carries out subsequent processing from there.

327 During the 'assembly and basic annotation' stage, users can opt to remove mis-
328 assembled portion of a virus contig, a process that may require significant memory
329 resources. Additionally, an ultrasensitive mode is available for analyzing RdRP libraries
330 that enhances the detection of a broader range of potential viral sequences. In the
331 'phylogenetic analysis' phase, users can choose to eliminate redundant sequences and
332 customize the amino acid length thresholds for specific virus of interest.
333 VirID employs several tools for visualizations: sankey diagrams, stacked
334 histograms, and phylogenetic tree diagrams are generated using the R packages
335 networkD3 (84), ggtree (85), and ggplot2 (86), respectively. Further visualizations are
336 created using Matplotlib v3.7.4 (87) and Seaborn v0.13.0 (88), enabling comprehensive
337 graphical representations of the data.
338
339
340 **RESULTS**
341 **VirID workflow**
342 VirID is a user-friendly command-line tool designed for the detection and
343 detailed characterization of RNA viruses from metagenomic data. The workflow is
344 divided into three stages: RNA virus discovery, phylogenetic analysis, and phylogeny-
345 based virus characterization (Figure 2). In the first stage, putative RNA virus genomes
346 containing the RdRP gene are identified through a homology approach, checked for
347 false positives and contamination, and clustered based on sequence similarity at the
348 species level. This process yields high-quality genomic sequences and related
349 information such as sequence length, closest relatives, and abundance levels. In the
350 second stage, VirID conducts high-quality sequence alignment and comprehensive
351 evolutionary analysis across the full diversity of the RNA virosphere: this further
352 removes low-quality sequences and reveals the phylogenetic position of all newly
353 discovered viruses. In the final stage, based on these evolutionary analyses, VirID
354 provides precise classification information and predicts potential viral associations
355 with humans, vertebrates, and plants.

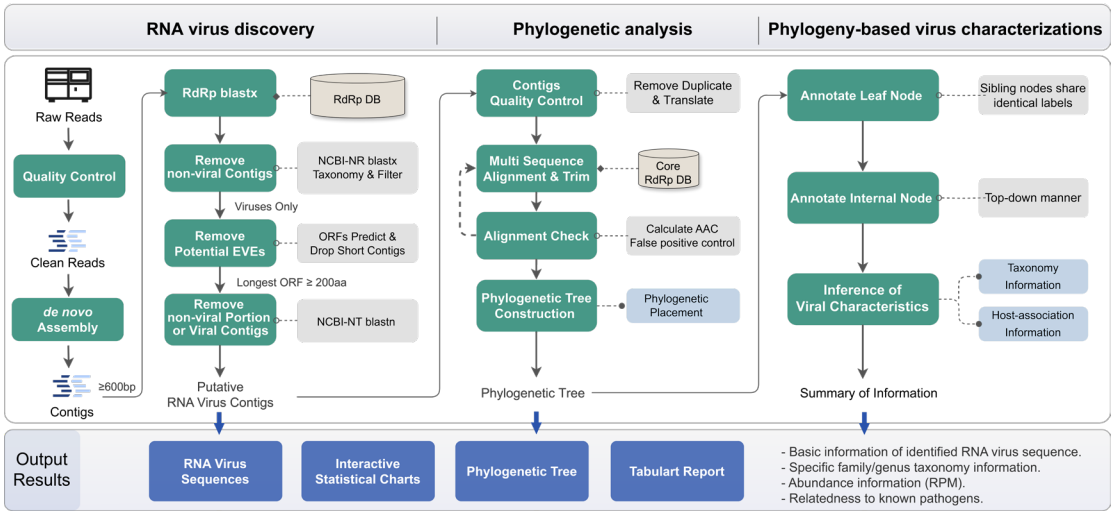


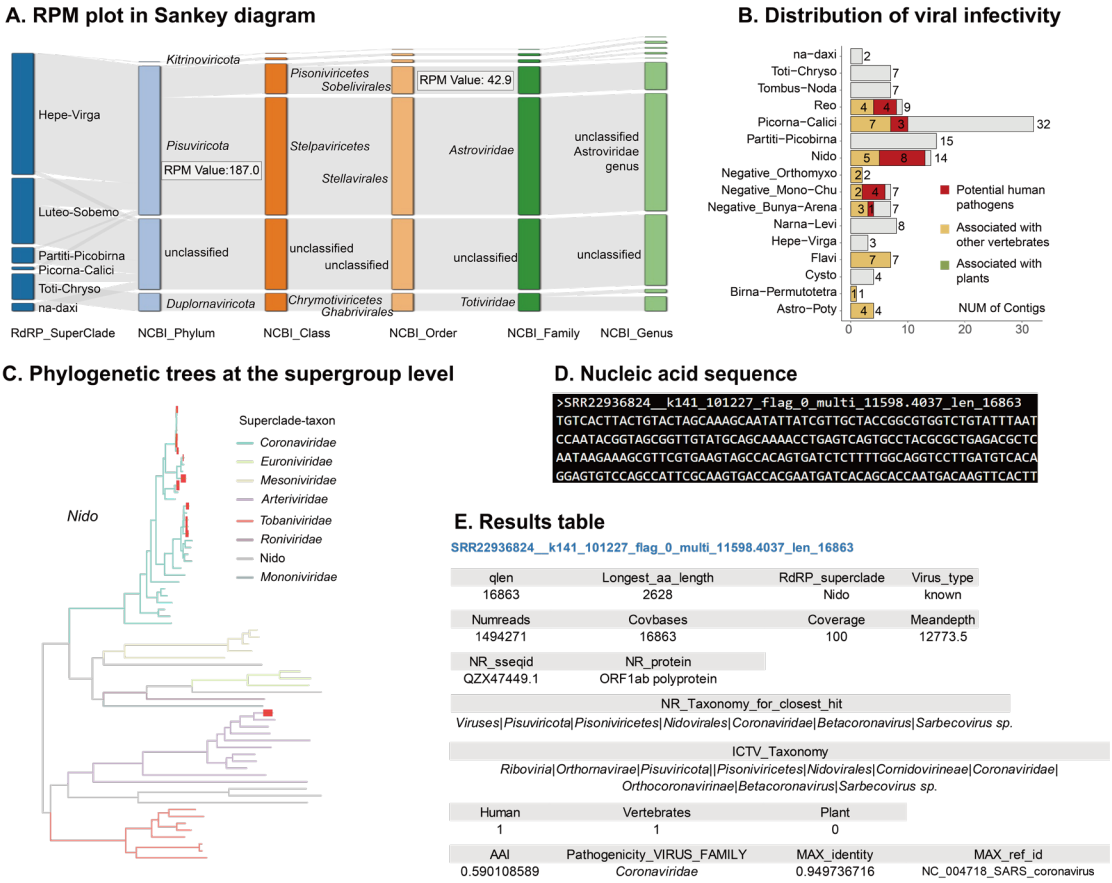
Figure 2. Framework for Identifying RNA Viruses from Sequencing Samples. The VirID process for automated RNA virus detection, which comprises three main stages: (**i**) RNA virus discovery, (**ii**) phylogenetic analysis, and (**iii**) phylogeny-based virus characterization. It pr oduces outputs that include viral sequences, phylogenetic trees, and comprehensive information including sequence length, best match of BLASTx comparison, virus classification, and host association.

Analysis Results Output in VirID

VirID conducts thorough data analyses and offers robust visualization tools, enabling an intuitive and comprehensive presentation of results. VirID organizes its output into three main folders: (i) ' assembly and basic annotation ' for intermediate files related to RNA virus discovery, (ii) 'phylogenetic analysis' containing well-labelled trees for various superclades, and (iii) 'results' for other figures and tables. It generates tables, fasta files, web files, and figure files (pdf format) to suit various analytical needs. We demonstrated these features here using a data set from 20 wildlife sequencing libraries that includes bats, insectivores, pangolins, pika and rodent samples (76) (Figure 3, Supplementary Table S2). In the RNA virus discovery segment, VirID estimates Reads Per Million (RPM) values for each potential viral species to assess their relative abundance. These values are displayed in a sankey plot on an HTML webpage, illustrating RPM distributions across RdRP superclades and NCBI taxonomy at multiple taxonomy levels—phylum, class, order, family, genus, and species (Figure 3A). VirID also categorizes potential viral sequences into one or two of four host-

379 association groups: human, vertebrate, plants, and others (Figure 3B). This information
380 is provided in color-coded trees that contain information on both phylogenetic
381 relationships and potential host associations (Figure 3C). Additionally, VirID outputs
382 high-quality viral genome sequences in fasta format (Figure 3D) and all other relevant
383 information, including sequence length, highest BLASTx match, virus classification,
384 association with principal hosts, in multidimensional tables (Figure 3E, Supplementary
385 Table S4, S5).

386



387

388 **Figure 3.** Visualization of Outputs from VirID. (**A**) Distribution of Reads Per Million (RPM) of potential viral
389 sequences from SRA sample SRR22936818, categorized by RdRP superclade and corresponding NCBI taxonomic
390 ranks. (**B**) The host associations of all identified virus sequences based on their placement in the phylogenetic tree.
391 (**C**) A phylogenetic tree is used to distinguish different viral lineages and colour-coded by taxonomic group. In this
392 case, newly identified virus sequences in the Nido superclade are highlighted in red at the end of the branches,
393 emphasizing their significance. (**D**) Complete nucleic acid sequences for detailed analysis and verification. (**E**)
394 Detailed information for each identified viral sequence, including host association, genomic details, and taxonomic
395 classification.

396

397 In the example of the 20 wildlife sequencing libraries VirID identified 129 potential
398 RNA viral contigs belonging to 107 various species, including 87 new virus species,
399 spanning 26 families and 15 superclades (Supplementary Figure S5A). Overall, 54 of
400 these species can be assigned at the family level, with the remainder falling outside
401 known families. Eleven of these species are potentially relevant to humans and include
402 Wufeng Niviventer niviventer orthohantavirus 1, Severe acute respiratory syndrome-
403 related coronavirus, Rotavirus A, and Pangolin respirovirus (Supplementary Figure
404 S5B, Supplementary Figure S6). Additionally, several viral species associated with
405 vertebrate infections were identified, including those within the genera
406 *Alphacoronavirus*, *Alphainfluenzavirus*, and *Mammarenavirus*. The remaining 88 viral
407 species are most likely derived from diet, parasites, and other microbial cellular
408 organisms within the principal host, because they were identified as plant viruses (e.g.
409 members of the *Tombunsviridae*) or closely related to arthropod (e.g. *Iflaviridae*) or
410 fungal viruses (e.g. *Mitoviridae*).
411
412 **Benchmarking VirID on simulated data set**
413 We initially evaluated VirID alongside five other tools—VirBot, VirSorter2,
414 VirSorter, DeepVirFinder, and VirFinder—using a collection of simulated samples (N
415 = 15) that included a diverse mix of RNA viruses and other species. These read data
416 samples were generated using the CAMISIM simulator, drawing on source genomes
417 and sequences from both the NCBI RefSeq database and recent studies. The RNA
418 viruses in the CAMISIM pool were comprised of four categories: 'known RNA viruses',
419 'new RNA virus L1', 'new RNA virus L2', and 'new RNA virus L3', representing
420 different similarity levels to the reference database sequences.
421 Six tools were compared across 15 data samples, including the average accuracy,
422 precision, recall, and F1 score metrics (Figure 4A). VirID achieved the highest F1 score
423 of 0.9334, with a standard error of ±0.0180. Its recall, at 0.8822 with a standard error
424 of ±0.0301, was close to VirBot (0.8822 ± 0.0247). Notably, VirID's precision was
425 perfect across all 15 samples, with an absence of false positives. Of all the tools, the

426 three with the best performance—VirBot, VirSorter2, and VirID—and were further
427 evaluated for their performance on 'new RNA viruses' which shared less than 90%
428 amino acid identity with those described in NCBI database (Figure 4B) and 'known
429 RNA viruses'. VirID continued to perform well in identifying 'new RNA viruses', again
430 with zero false positives (Supplementary Figure S7).

431 We also analyzed whether the number of viral sequences or species revealed by
432 each tool matched that of expected species count in the simulated data (Figure 4C).
433 VirID's estimation of contigs (39, average) and species (27, average) were the mostly
434 closely aligned with the 30 true species samples, followed by VirSorter2 (average of 45
435 contigs and 26 species) and then VirBot (53 contigs and 27 species). VirID achieved a
436 ratio of 1.47 for the number of estimated sequences relative to the number of true
437 species. This ratio was the lowest among all tools, indicating that VirID's output most
438 accurately reflect the actual number of virus species.

439 The integrity of the RNA viral sequences revealed and processed by VirID was
440 assessed using the CAMISIM simulator's 'gold standard' contigs as a benchmark. These
441 standard contigs represent idealized assembly results of the simulated reads, providing
442 a baseline for evaluating the detected virus sequences. We aligned the putative viral
443 contigs with these 'gold standard' contigs, selecting the best matches based on coverage
444 and identity. The average completeness for contigs in 10 samples exceeded 90%, with
445 an overall average completeness of 90.80% across all 15 samples (Figure 4D). This
446 meets CheckV's criterion for 'high quality completeness' (89).

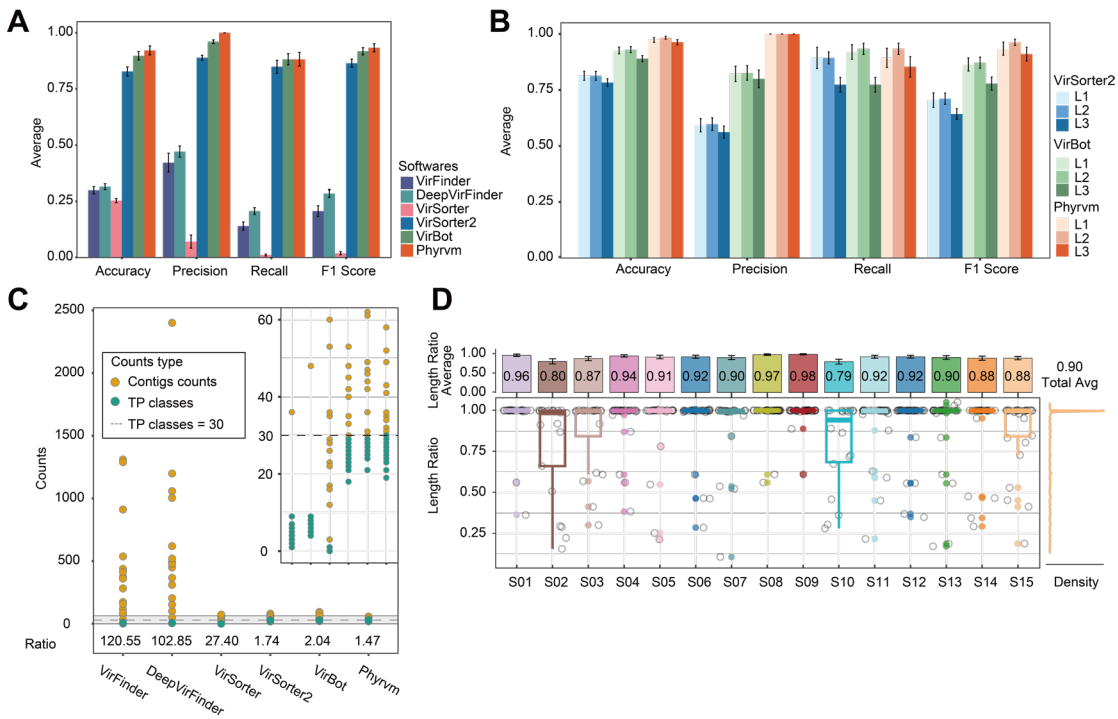


Figure 4. Benchmarking of Various Tools using Simulated Data. (**A**) Performance of six bioinformatic tools across 15 simulated samples, evaluating average accuracy, precision, recall, and F1 score. Error bars represent standard errors. (**B**) Three tools – VirID, VirBot, VirSorter2 – had superior performance and were further assessed for their specific ability to identify 'new RNA viruses' at various levels of classification. (**C**) Effectiveness of each tool in identifying RNA virus sequences across 15 simulated samples. Yellow dots represent the number of RNA virus sequences identified, while green dots indicate the number of corresponding species-level taxa. The inset graph in the upper right corner adjusts the y-axis to range from 0-60 for a focused view. (**D**) Ratio of the lengths of sequences identified by VirID in various simulated data samples compared to the length of the corresponding gold standard sequence. Only sequences that uniquely correspond to gold standard sequences are considered.

A re-analyses of previously published data sets

We next assessed the performance of VirID by analyzing 192 meta-transcriptomic SRA libraries from seven distinct virome studies, covering a range of sample types including clinical (human), wild birds, Malayan pangolins, Qinghai voles, seabird ticks, soybean fields, and peat soil samples (Supplementary Table S3). VirID significantly improved viral discovery. From these data sets, a total of 1,283 RNA viral species were identified, with the soil samples displaying the highest average virus species richness at 88.5 species, followed by Qinghai vole at 2 and seabird tick at 1.9. Importantly, 1,202 novel RNA viral species were identified, among which 1,198 species were not described in the original study. Among these, the highest number of novel viruses (N =

740) was found in the soil samples, with 56% of these viruses categorized within the Narna-Levi superclade. At the family level, the *Picobirnaviridae* contained the most viral species (*N* = *164*), followed by the *Tombusviridae* (*N* = *83*) and *Mitoviridae* (*N* = *69*), although many virus species (54.5%) could not be assigned at family level (Figure 5A).

We next evaluated the number of families identified and compared these with previous studies, omitting any with incomplete classification information. Our analysis revealed an average increase of five viral families, with notable differences particularly observed in the vole and soybean data sets (Figure 5B). In addition, VirID provided a more precise classification system for the target viruses. For instance, viruses that were previously classified only at the kingdom level, such as Bulatov virus, Fennes virus, Ronne virus, and Vovk virus from the seabird tick data set, were now classified at the genus level (Supplementary Table S6).

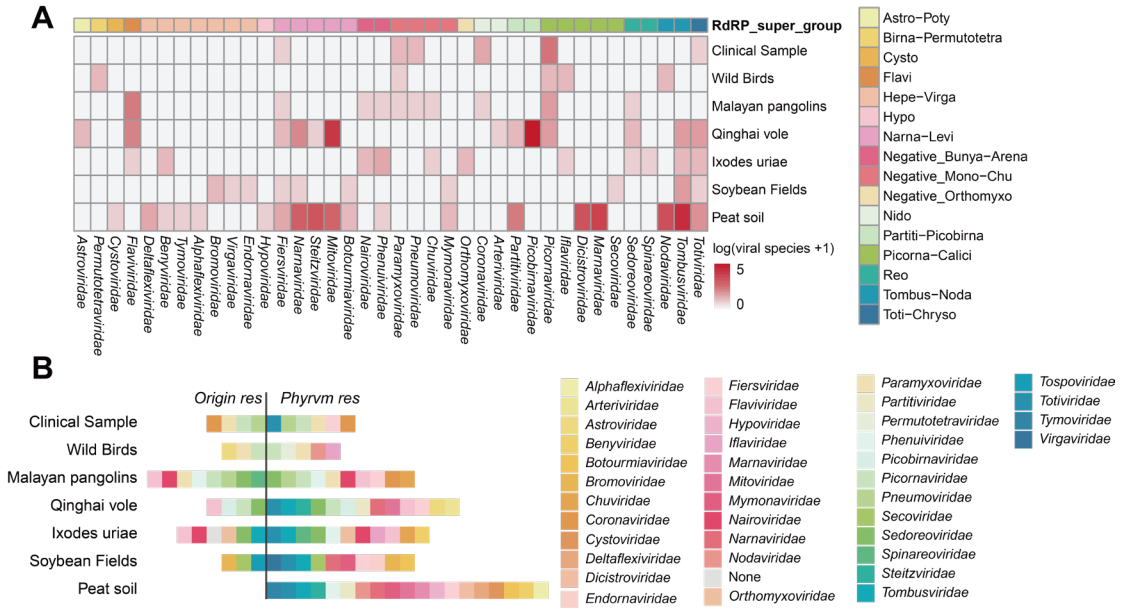


Figure 5. Reclassification of Viruses Using VirID using Real-World Data. (**A**) Distribution of virus species across different superclades and families identified by VirID in previously published data sets, including human swabs (77), wild birds (78), Malayan pangolins (79) Qinghai voles (80), arthropods (81), plant (82), and soil samples (83). (**B**) The utility of VirID across a broader spectrum of RNA virus families. Virus families documented in the original publications are displayed on the left, while those identified in this study are shown on the right.

489 VirID also provides information on whether the viruses observed are associated
490 with vertebrates, plants, or have the potential to infect humans. For example, in the
491 Qinghai vole data set, we identified 180 species that were likely associated with voles,
492 including *Mamastrovirus* (N = 1), *Hepacivirus* (*N = 5*), *Pegivirus* (*N = 2*), *Arteriviridae*
493 (*N = 1*), and *Picobirnaviridae* (*N =164*). Additionally, 250 virus species were likely
494 associated with food sources, parasites, and symbionts within these hosts, including
495 Narna-Levi (*N = 163*), Partiti-Picobirna (*N = 74*) and Toti-Chryso (*N = 6*) (Figure 6A).
496 Overall, vertebrate-associated viruses accounted for a median of 66% of the total RNA
497 in host samples, ranging from 30.2-73.6%, while other viruses accounted for 23.3%,
498 with a range of 10.5-67.9% (Figure 6B). Notably, we identified two species,
499 Hepatovirus D and Rotavirus A, that are potential human pathogens. Hepatovirus D
500 was described in the original publication (80), while Rotavirus A was newly identified
501 here. These viruses were discovered because they cluster with known human pathogens
502 at over 80% sequence identity with these viral contigs.
503 Similarly, we identified 17 potential human pathogens, primarily within the clinical
504 sample data set. This included potential human pathogens such as Pangolin coronavirus
505 HKU4, Rotavirus A, and Pangolin Respirovirus identified in pangolins, as well as the
506 vector-borne Gadgets Gully virus, Okhotskiy virus, and Neke harbour virus from the
507 seabird tick samples. This is consistent with previous publications. Additionally, plant-
508 infecting viruses were identified in the vole (*N = 5*, likely diet-related), soybean (*N =*
509 *6*), and soil (*N = 24*) samples (Figure 6C, Supplementary Table S7). In the Soybean
510 field data set, we detected 79 contigs from six viral species associated with plants,
511 constituting about 74.5% of the total RNA, including Tobacco streak virus, Soybean
512 dwarf virus, Pepper mild mottle virus, Bean pod mottle virus, and Alfalfa mosaic virus
513 (Figure 6D).
514
515

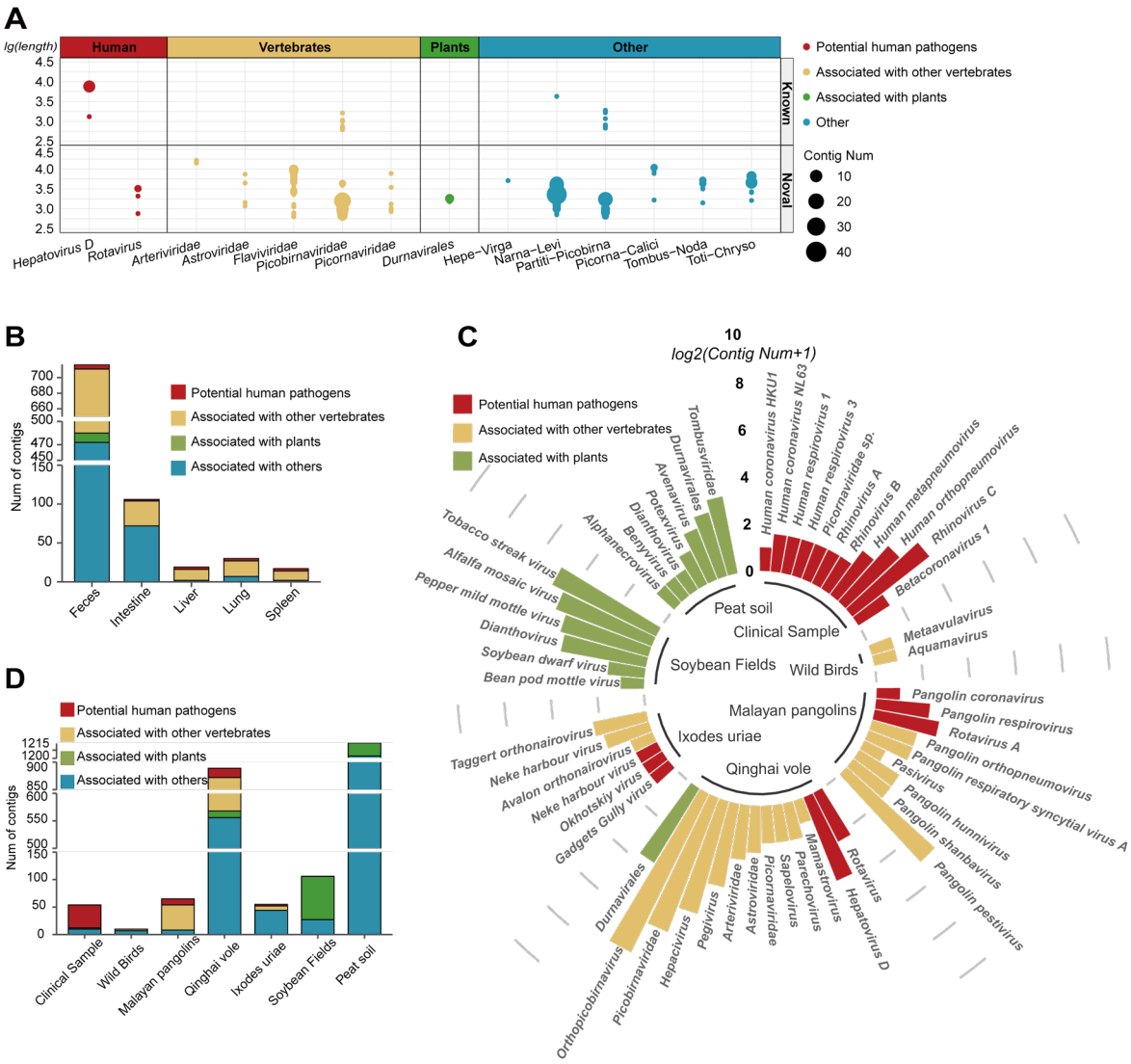


Figure 6. Virus-Host Associations Revealed by VirID. (**A**) Number of known and unknown virus sequences related to humans, vertebrates, plants, and other hosts as identified by VirID in the Qinghai vole case study. (**B**) Count of virus sequences associated with different hosts in various tissue and fecal samples from voles identified by VirID. (**C**) Number and taxonomy of virus sequences related to different hosts in various real-world samples identified by VirID (**D**) Number of virus sequences related to different hosts in a variety of real-world samples identified by VirID.

DISCUSSION

Herein, we present VirID, an RNA virus discovery platform specifically designed for Linux servers. VirID integrates phylogenetic analysis into the identification and characterization of RNA viruses, thereby substantially improving virus discovery. The platform employs an iterative scoring strategy in its alignment processes to minimize false positives and incorporates a phylogenetic placement algorithm for the rapid and

530 stable phylogenetic tree inference. As a result, users can accurately estimate RNA virus
531 diversity, classify each identified virus with high precision, at the same time obtaining
532 insights into host associations which enable the potential threat to human, animal and
533 plant health to be evaluated. The wealth of information generated by this platform will
534 be invaluable for researchers involved in early warning of infectious diseases across
535 diverse settings.

536 Our results show that VirID provides the most precise estimation of viral species
537 composition, while other programs tend to overestimate genetic diversity. This
538 overestimation may occur because some methods consider all virus-associated contigs,
539 including those with low coverage and fragmented genomes, potentially counting them
540 as multiple species. To depict diversity more accurately, it is essential to focus on
541 regions of the genome shared by the majority of viruses, thereby minimizing
542 overestimation (90,91). The RdRP gene is an excellent candidate for several reasons.
543 First, all RNA viruses, with the exception of RNA satellite viruses, contain the RdRP
544 (27-29). Second, the RdRP protein is the most conserved gene in the RNA virus genome,
545 and often comparable across different virus classes or even phyla (92,93). Therefore,
546 focusing solely on RdRP contigs is key to accurately determining viral diversity within
547 a sample.

548 Compared to other virus discovery programs, the phylogenetic analysis feature of
549 VirID stands provides a central analytical component. In particular, it provides an
550 automated and reliable method for accurately classifying previously undescribed RNA
551 viruses, especially those that are highly divergent in sequence. Traditionally, the
552 approach to classifying new viruses involved BLAST analyses and assigning
553 taxonomic position based on the closest related hits (41,43,94). However, this method
554 may counter the guidelines laid down by the International Committee on Taxonomy of
555 Viruses (ICTV) who apply varying criteria for virus taxonomy at the species and genus
556 levels (24,95,96). Additionally, the lack of a clear definition of what levels of genetic
557 similarity differentiate higher taxonomic ranks complicates the assignment of viruses
558 to new families or orders when protein identity is below 40%. More importantly, some

559 reference sequences are incorrectly classified, which can introduce errors into the
560 classification process. For instance, a sequence from soil metagenomic data
561 (MN035928), from a divergent member within the order *Bunyavirales*, was mistakenly
562 labeled as belonging to the genus *Arenavirus* within the *Arenaviridae* (97). As a
563 consequence, subsequent discoveries of similar viruses (OQ715420)(98) were also
564 mislabeled as *Arenaviridae*. Such misclassifications could be avoided with robust and
565 reliable phylogenetic analyses.

566 Phylogenetic analysis is also central to the inference of host associations. Viruses
567 from similar host categories tend to cluster together, forming what is known as a
568 phylogenetic monophyly, indicative of host structure in virus phylogeny (27,28,99,100).
569 This pattern holds true across different types of viruses, such as vertebrate-specific
570 viruses, arthropod-borne viruses, and plant viruses, in which host-associated
571 phylogenetic monophyletic groups are identifiable at various taxonomic levels—
572 ranging from family level (e.g., *Picornaviridae*), to genus level (e.g., *Alphavirus*), and
573 even within specific genetic lineages in a genus (e.g., mosquito-borne and tick-borne
574 virus groups within genus *Flavivirus*). Due to the varying scales and complexity of
575 these host-associated groups, phylogenetic analysis serves as a valuable tool for
576 exploring these relationships (75,76,79). However, assigning viruses to other host
577 groups such as arthropods, nematodes, and even basal eukaryotes (such as parasites)
578 remains challenging due to the scarcity of relevant virus data on these hosts. In addition,
579 a limitation to all virus discovery tools, including VirID, is that they may not identify
580 vertebrate-associated viruses if they occupy phylogenetic positions not previously
581 linked to vertebrate hosts.

582 The ability to infer host associations is highly relevant in disease monitoring
583 programs in which emerging pathogens are continuously identified and evaluated
584 (21,101,102). Numerous viruses are discovered in sequencing efforts, but not all are
585 relevant to the infection of principal host or impact health. Indeed, those relevant to
586 disease often constitute only a small fraction (103). For example, in a survey of over
587 1941 game animals across China, more than 1000 viruses were identified, yet only 102

588 were linked to mammalian infections, and even fewer (N = 21) were considered to pose
589 a significant risk of infecting humans or other animal species (104). Similarly, a recent
590 meta-transcriptomics study of 2438 mosquitoes in China revealed that among the 564
591 RNA virus species detected, 393 were likely associated with mosquitoes, but only 7
592 were linked to mammalian infections (i.e., arboviruses) (105). Thus, discerning
593 potential host associations is crucial for assessing the public health or economic impact
594 of discovered viruses (20,106,107).

595 Our study is subject to several limitations. First, while the phylogenetic placement
596 method has accelerated the estimation of evolutionary trees, the self-iterative multi-
597 sequence alignment process remains time-consuming. Second, VirID primarily relies
598 on the RdRP for homology evidence and is therefore unable to identify genomes that
599 do not contain RdRP sequences or that are too divergent to be detected in homology-
600 based searches (108,109). Consequently, for segmented viruses, the full genome
601 assembly may be incomplete. Third, VirID does not provide strain or genotype level
602 typing which often depend on more variable parts of viral genomes (110,111). Finally,
603 the RdRP database is continuously expanding, particularly with the addition of data
604 from environmental samples (29,30,37,57,112). However, since these environmental
605 RdRP data are unverified, and this study focuses on human, animal, and plant health,
606 they are currently excluded from our database. Nevertheless, given the flexibility of our
607 RdRP database structure, these data can be readily incorporated to facilitate future
608 broader scale analyses.

609

610 **DATA AVAILABILITY**

611 VirID is freely available as an open-source Python code at
612 https://github.com/ZiyueYang01/VirID

613

614 **SUPPLEMENTARY DATA**

615 Supplementary Data are available at NAR Online.

616

AUTHOR CONTRIBUTIONS

Conceptualization, EC Holmes, DY Guo and M Shi; Methodology, ZY Yang, YT Shan, X Liu, GW Chen, YF Pan, DY Guo and M Shi; Investigation, ZY Yang and YT Shan; Data Collection and Processing, QY Gou, J Zou, ZL Chang, Q Zeng, CH Yang, JB Kong, WC Wu, DY Guo and M Shi; Writing – Original Draft, ZY Yang and YT Shan; Writing – Review and Editing, All authors. Funding Acquisition, CM Li, H Peng, DY Guo and M Shi; Resources (computational), SC Li, X Zhang, DY Guo and M Shi; Supervision, CM Li , H Peng, DY Guo and M Shi.

FUNDING

This study is funded by National Natural Science Foundation of China (82341118, 32270160), Natural Science Foundation of Guangdong Province of China (2022A1515011854), Shenzhen Science and Technology Program (JCYJ20210324124414040, KQTD20200820145822023), Hong Kong Innovation and Technology Fund (ITF) (MRP/071/20X), Major Project of Guangzhou National Laboratory (GZNL2023A01001), Guangdong Province “Pearl River Talent Plan” Innovation, Entrepreneurship Team Project (2019ZT08Y464), and the Fund of Shenzhen Key Laboratory (ZDSYS20220606100803007). E.C. Holmes was supported by an NHMRC (Australia) Investigator Award (GNT2017197) and by AIR@InnoHK administered by the Innovation and Technology Commission, Hong Kong Special Administrative Region, China.

REFERENCES

1. Nicaise, V. (2014) Crop immunity against viruses: outcomes and future challenges. *Frontiers in plant science*, **5**, 660.
2. Fauci, A.S. (1988) The human immunodeficiency virus: infectivity and mechanisms of pathogenesis. *Science*, **239**, 617-622.
3. Zhong, N., Zheng, B., Li, Y., Poon, L., Xie, Z., Chan, K., Li, P., Tan, S., Chang, Q. and Xie, J. (2003) Epidemiology and cause of severe acute respiratory syndrome (SARS) in Guangdong, People's Republic of China, in February, 2003. *The Lancet*, **362**, 1353-1358.
4. Smith, G.J., Vijaykrishna, D., Bahl, J., Lycett, S.J., Worobey, M., Pybus, O.G., Ma, S.K., Cheung, C.L., Raghwani, J. and Bhatt, S. (2009) Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. *Nature*, **459**, 1122-1125.

5. Assiri, A., McGeer, A., Perl, T.M., Price, C.S., Al Rabeeah, A.A., Cummings, D.A., Alabdullatif, Z.N., Assad, M., Almulhim, A. and Makhdoom, H. (2013) Hospital outbreak of Middle East respiratory syndrome coronavirus. *New England Journal of Medicine*, **369**, 407-416.
6. Team, W.E.R. (2014) Ebola virus disease in West Africa—the first 9 months of the epidemic and forward projections. *New England Journal of Medicine*, **371**, 1481-1495.
7. Musso, D. and Gubler, D.J. (2016) Zika virus. *Clinical microbiology reviews*, **29**, 487-524.
8. Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., Zhao, X., Huang, B., Shi, W. and Lu, R. (2020) A novel coronavirus from patients with pneumonia in China, 2019. *New England journal of medicine*, **382**, 727-733.
9. Chan, J.F.-W., To, K.K.-W., Tse, H., Jin, D.-Y. and Yuen, K.-Y. (2013) Interspecies transmission and emergence of novel viruses: lessons from bats and birds. *Trends in microbiology*, **21**, 544-555.
10. Claas, E.C., Osterhaus, A.D., Van Beek, R., De Jong, J.C., Rimmelzwaan, G.F., Senne, D.A., Krauss, S., Shortridge, K.F. and Webster, R.G. (1998) Human influenza A H5N1 virus related to a highly pathogenic avian influenza virus. *The Lancet*, **351**, 472-477.
11. Peiris, J.M., De Jong, M.D. and Guan, Y. (2007) Avian influenza virus (H5N1): a threat to human health. *Clinical microbiology reviews*, **20**, 243-267.
12. Martina, B.E., Koraka, P. and Osterhaus, A.D. (2009) Dengue virus pathogenesis: an integrated view. *Clinical microbiology reviews*, **22**, 564-581.
13. Ergönül, Ö. (2006) Crimean-Congo haemorrhagic fever. *The Lancet infectious diseases*, **6**, 203-214.
14. He, S. and Krainer, K.M.C. (2020) Pandemics of people and plants: which is the greater threat to food security? *Molecular plant*, **13**, 933-934.
15. Lee, C. (2015) Porcine epidemic diarrhea virus: an emerging and re-emerging epizootic swine virus. *Virology journal*, **12**, 1-16.
16. Scholthof, K.B.G., Adkins, S., Czosnek, H., Palukaitis, P., Jacquot, E., Hohn, T., Hohn, B., Saunders, K., Candresse, T. and Ahlquist, P. (2011) Top 10 plant viruses in molecular plant pathology. *Molecular plant pathology*, **12**, 938-954.
17. Untiveros, M., Fuentes, S. and Salazar, L.F. (2007) Synergistic interaction of Sweet potato chlorotic stunt virus (Crinivirus) with carla-, cucumo-, ipomo-, and potyviruses infecting sweet potato. *Plant disease*, **91**, 669-676.
18. Robilotti, E., Deresinski, S. and Pinsky, B.A. (2015) Norovirus. *Clinical microbiology reviews*, **28**, 134-164.
19. Mackenzie, J.S., Gubler, D.J. and Petersen, L.R. (2004) Emerging flaviviruses: the spread and resurgence of Japanese encephalitis, West Nile and dengue viruses. *Nature medicine*, **10**, S98-S109.
20. Lefrançois, T., Malvy, D., Atlani-Duault, L., Benamouzig, D., Druais, P.-L., Yazdanpanah, Y., Delfraissy, J.-F. and Lina, B. (2023) After 2 years of the COVID-19 pandemic, translating One Health into action is urgent. *The Lancet*, **401**, 789-794.
21. Carroll, D., Daszak, P., Wolfe, N.D., Gao, G.F., Morel, C.M., Morzaria, S., Pablos-Méndez, A., Tomori, O. and Mazet, J.A. (2018) The global virome project. *Science*, **359**, 872-874.
22. Zhang, Y.-Z., Shi, M. and Holmes, E.C. (2018) Using metagenomics to characterize an expanding virosphere. *Cell*, **172**, 1168-1172.

23. Shi, M., Zhang, Y.-Z. and Holmes, E.C. (2018) Meta-transcriptomics and the evolutionary biology of RNA viruses. *Virus research*, **243**, 83-90.
24. Simmonds, P., Adams, M.J., Benkő, M., Breitbart, M., Brister, J.R., Carstens, E.B., Davison, A.J., Delwart, E., Gorbalenya, A.E. and Harrach, B. (2017) Virus taxonomy in the age of metagenomics. *Nature Reviews Microbiology*, **15**, 161-168.
25. Zhang, Y.-Z., Chen, Y.-M., Wang, W., Qin, X.-C. and Holmes, E.C. (2019) Expanding the RNA virosphere by unbiased metagenomics. *Annual Review of Virology*, **6**, 119-139.
26. Greninger, A.L. (2018) A decade of RNA virus metagenomics is (not) enough. *Virus Research*, **244**, 218-229.
27. Shi, M., Lin, X.-D., Tian, J.-H., Chen, L.-J., Chen, X., Li, C.-X., Qin, X.-C., Li, J., Cao, J.-P. and Eden, J.-S. (2016) Redefining the invertebrate RNA virosphere. *Nature*, **540**, 539-543.
28. Shi, M., Lin, X.-D., Chen, X., Tian, J.-H., Chen, L.-J., Li, K., Wang, W., Eden, J.-S., Shen, J.-J. and Liu, L. (2018) The evolutionary history of vertebrate RNA viruses. *Nature*, **556**, 197-202.
29. Zayed, A.A., Wainaina, J.M., Dominguez-Huerta, G., Pelletier, E., Guo, J., Mohssen, M., Tian, F., Pratama, A.A., Bolduc, B. and Zablocki, O. (2022) Cryptic and abundant marine viruses at the evolutionary origins of Earth's RNA virome. *Science*, **376**, 156-162.
30. Chen, Y.-M., Sadiq, S., Tian, J.-H., Chen, X., Lin, X.-D., Shen, J.-J., Chen, H., Hao, Z.-Y., Wille, M. and Zhou, Z.-C. (2022) RNA viromes from terrestrial sites across China expand environmental viral diversity. *Nature Microbiology*, **7**, 1312-1323.
31. Huhtamo, E., Moureau, G., Cook, S., Julkunen, O., Putkuri, N., Kurkela, S., Uzcategui, N.Y., Harbach, R.E., Gould, E.A. and Vapalahti, O. (2012) Novel insect-specific flavivirus isolated from northern Europe. *Virology*, **433**, 471-478.
32. Culley, A.I., Lang, A.S. and Suttle, C.A. (2003) High diversity of unknown picorna-like viruses in the sea. *Nature*, **424**, 1054-1057.
33. Lam, T.T.-Y., Jia, N., Zhang, Y.-W., Shum, M.H.-H., Jiang, J.-F., Zhu, H.-C., Tong, Y.-G., Shi, Y.-X., Ni, X.-B. and Liao, Y.-S. (2020) Identifying SARS-CoV-2-related coronaviruses in Malayan pangolins. *Nature*, **583**, 282-285.
34. Mifsud, J.C., Gallagher, R.V., Holmes, E.C. and Geoghegan, J.L. (2022) Transcriptome mining expands knowledge of RNA viruses across the plant kingdom. *Journal of Virology*, **96**, e00260-00222.
35. Li, C.-X., Shi, M., Tian, J.-H., Lin, X.-D., Kang, Y.-J., Chen, L.-J., Qin, X.-C., Xu, J., Holmes, E.C. and Zhang, Y.-Z. (2015) Unprecedented genomic diversity of RNA viruses in arthropods reveals the ancestry of negative-sense RNA viruses. *elife*, **4**, e05378.
36. Holmes, E.C. (2009) *The evolution and emergence of RNA viruses*. Oxford University Press.
37. Edgar, R.C., Taylor, B., Lin, V., Altman, T., Barbera, P., Meleshko, D., Lohr, D., Novakovsky, G., Buchfink, B. and Al-Shayeb, B. (2022) Petabase-scale sequence alignment catalyses viral discovery. *Nature*, **602**, 142-147.
38. Shi, M., Hou, X., He, Y., Fang, P., Mei, S.-Q., Xu, Z., Wu, W.-C., Tian, J.-h., Zhang, S. and Zeng, Z.-Y. (2023) Artificial intelligence redefines RNA virus discovery.
39. Rampelli, S., Soverini, M., Turroni, S., Quercia, S., Biagi, E., Brigidi, P. and Candela, M. (2016) ViromeScan: a new tool for metagenomic viral community profiling. *Bmc*

Genomics, **17**, 1-9.

40. Flygare, S., Simmon, K., Miller, C., Qiao, Y., Kennedy, B., Di Sera, T., Graf, E.H., Tardif, K.D., Kapusta, A. and Rynearson, S. (2016) Taxonomer: an interactive metagenomics analysis portal for universal pathogen detection and host mRNA expression profiling. *Genome biology*, **17**, 1-18.
41. Lin, J., Kramna, L., Autio, R., Hyöty, H., Nykter, M. and Cinek, O. (2017) Vipie: web pipeline for parallel characterization of viral populations from multiple NGS samples. *BMC genomics*, **18**, 1-11.
42. Kalantar, K.L., Carvalho, T., de Bourcy, C.F., Dimitrov, B., Dingle, G., Egger, R., Han, J., Holmes, O.B., Juan, Y.-F. and King, R. (2020) IDseq—An open source cloud-based pipeline and analysis service for metagenomic pathogen detection and monitoring. *Gigascience*, **9**, giaa111.
43. Plyusnin, I., Vapalahti, O., Sironen, T., Kant, R. and Smura, T. (2023) Enhanced viral metagenomics with lazypipe 2. *Viruses*, **15**, 431.
44. Zhou, Z., Martin, C., Kosmopoulos, J.C. and Anantharaman, K. (2023) ViWrap: A modular pipeline to identify, bin, classify, and predict viral–host relationships for viruses from metagenomes. *Imeta*, **2**, e118.
45. Akhter, S., Aziz, R.K. and Edwards, R.A. (2012) PhiSpy: a novel algorithm for finding prophages in bacterial genomes that combines similarity-and composition-based strategies. *Nucleic acids research*, **40**, e126-e126.
46. Roux, S., Enault, F., Hurwitz, B.L. and Sullivan, M.B. (2015) VirSorter: mining viral signal from microbial genomic data. *PeerJ*, **3**, e985.
47. Guo, J., Bolduc, B., Zayed, A.A., Varsani, A., Dominguez-Huerta, G., Delmont, T.O., Pratama, A.A., Gazitúa, M.C., Vik, D. and Sullivan, M.B. (2021) VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. *Microbiome*, **9**, 1-13.
48. Kieft, K., Zhou, Z. and Anantharaman, K. (2020) VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. *Microbiome*, **8**, 1-23.
49. Ren, J., Ahlgren, N.A., Lu, Y.Y., Fuhrman, J.A. and Sun, F. (2017) VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. *Microbiome*, **5**, 1-20.
50. Ren, J., Song, K., Deng, C., Ahlgren, N.A., Fuhrman, J.A., Li, Y., Xie, X., Poplin, R. and Sun, F. (2020) Identifying viruses from metagenomic data using deep learning. *Quantitative Biology*, **8**, 64-77.
51. Chen, G., Tang, X., Shi, M. and Sun, Y. (2023) VirBot: an RNA viral contig detector for metagenomic data. *Bioinformatics*, **39**, btad093.
52. Hegarty, B., Riddell V, J., Bastien, E., Langenfeld, K., Lindback, M., Saini, J.S., Wing, A., Zhang, J. and Duhaime, M.J.M. (2024) Benchmarking informatics approaches for virus discovery: caution is needed when combining in silico identification methods. **9**, e01105-01123.
53. Dutilh, B.E., Varsani, A., Tong, Y., Simmonds, P., Sabanadzovic, S., Rubino, L., Roux, S., Munoz, A.R., Lood, C. and Lefkowitz, E.J.J.C.o.i.v. (2021) Perspective on taxonomic classification of uncultivated viruses. **51**, 207-215.

54. Simon-Loriere, E. and Holmes, E.C. (2011) Why do RNA viruses recombine? *Nature Reviews Microbiology*, **9**, 617-626.
55. Drake, J.W. (1993) Rates of spontaneous mutation among RNA viruses. *Proceedings of the National Academy of Sciences*, **90**, 4171-4175.
56. Shi, M., Lin, X.-D., Tian, J.-H., Chen, L.-J., Chen, X., Li, C.-X., Qin, X.-C., Li, J., Cao, J.-P., Eden, J.-S. *et al.* (2016) Redefining the invertebrate RNA virosphere. *Nature*, **540**, 539-+.
57. Hou, X., He, Y., Fang, P., Mei, S.-Q., Xu, Z., Wu, W.-C., Tian, J.-H., Zhang, S., Zeng, Z.-Y. and Gou, Q.-Y. (2023) Artificial intelligence redefines RNA virus discovery. *bioRxiv*, 2023.2004. 2018.537342.
58. Buchfink, B., Xie, C. and Huson, D.H. (2015) Fast and sensitive protein alignment using DIAMOND. *Nature methods*, **12**, 59-60.
59. Babaian, A. and Edgar, R. (2022) Ribovirus classification by a polymerase barcode sequence. *PeerJ*, **10**, e14055.
60. Fu, L., Niu, B., Zhu, Z., Wu, S. and Li, W. (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, **28**, 3150-3152.
61. Mihara, T., Nishimura, Y., Shimizu, Y., Nishiyama, H., Yoshikawa, G., Uehara, H., Hingamp, P., Goto, S. and Ogata, H. (2016) Linking virus genomes with host taxonomy. *Viruses*, **8**, 66.
62. Bushnell, B. (2014). Lawrence Berkeley National Lab.(LBNL), Berkeley, CA (United States).
63. Langmead, B. and Salzberg, S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nature methods*, **9**, 357-359.
64. Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J. and Glöckner, F.O. (2012) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic acids research*, **41**, D590-D596.
65. Cole, J.R., Chai, B., Farris, R.J., Wang, Q., Kulam-Syed-Mohideen, A., McGarrell, D.M., Bandela, A., Cardenas, E., Garrity, G.M. and Tiedje, J.M. (2007) The ribosomal database project (RDP-II): introducing myRDP space and quality controlled public data. *Nucleic acids research*, **35**, D169-D172.
66. Li, D., Liu, C.-M., Luo, R., Sadakane, K. and Lam, T.-W. (2015) MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*, **31**, 1674-1676.
67. Shen, W. and Ren, H. (2021) TaxonKit: A practical and efficient NCBI taxonomy toolkit. *Journal of Genetics and Genomics*, **48**, 844-850.
68. Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. and Madden, T.L. (2009) BLAST+: architecture and applications. *BMC bioinformatics*, **10**, 1-9.
69. Katoh, K. and Standley, D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution*, **30**, 772-780.
70. Capella-Gutiérrez, S., Silla-Martínez, J.M. and Gabaldón, T. (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, **25**, 1972-1973.
71. Stamatakis, A., Hoover, P. and Rougemont, J. (2008) A rapid bootstrap algorithm for the RAxML web servers. *Systematic biology*, **57**, 758-771.
72. Matsen, F.A., Kodner, R.B. and Armbrust, E.V. (2010) pplacer: linear time maximum-

826 likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference
827 tree. *BMC bioinformatics*, **11**, 1-16.
828 73. Fritz, A., Hofmann, P., Majda, S., Dahms, E., Dröge, J., Fiedler, J., Lesker, T.R., Belmann, P.,
829 DeMaere, M.Z. and Darling, A.E. (2019) CAMISIM: simulating metagenomes and
830 microbial communities. *Microbiome*, **7**, 1-12.
831 74. Feng, Y., Gou, Q.-y., Yang, W.-h., Wu, W.-c., Wang, J., Holmes, E.C., Liang, G. and Shi, M.
832 (2022) A time-series meta-transcriptomic analysis reveals the seasonal, host, and gender
833 structure of mosquito viromes. *Virus Evolution*, **8**, veac006.
834 75. Wang, J., Pan, Y.-f., Yang, L.-f., Yang, W.-h., Lv, K., Luo, C.-m., Wang, J., Kuang, G.-p.,
835 Wu, W.-c. and Gou, Q.-y. (2023) Individual bat virome analysis reveals co-infection and
836 spillover among bats and virus zoonotic potential. *Nature Communications*, **14**, 4079.
837 76. Cui, X., Fan, K., Liang, X., Gong, W., Chen, W., He, B., Chen, X., Wang, H., Wang, X. and
838 Zhang, P. (2023) Virus diversity, wildlife-domestic animal circulation and potential
839 zoonotic viruses of small mammals, pangolins and zoo animals. *Nature*
840 *Communications*, **14**, 2488.
841 77. Graf, E.H., Simmon, K.E., Tardif, K.D., Hymas, W., Flygare, S., Eilbeck, K., Yandell, M. and
842 Schlaberg, R. (2016) Unbiased detection of respiratory viruses by use of RNA
843 sequencing-based metagenomics: a systematic comparison to a commercial PCR panel.
844 *Journal of clinical microbiology*, **54**, 1000-1007.
845 78. Chang, W.-S., Eden, J.-S., Hall, J., Shi, M., Rose, K. and Holmes, E.C. (2020)
846 Metatranscriptomic analysis of virus diversity in urban wild birds with paretic disease.
847 *Journal of Virology*, **94**, 10.1128/jvi. 00606-00620.
848 79. Shi, W., Shi, M., Que, T.-C., Cui, X.-M., Ye, R.-Z., Xia, L.-Y., Hou, X., Zheng, J.-J., Jia, N.
849 and Xie, X. (2022) Trafficked Malayan pangolins contain viral pathogens of humans.
850 *Nature microbiology*, **7**, 1259-1269.
851 80. He, X., Wang, X., Fan, G., Li, F., Wu, W., Wang, Z., Fu, M., Wei, X., Ma, S. and Ma, X. (2022)
852 Metagenomic analysis of viromes in tissues of wild Qinghai vole from the eastern
853 Tibetan Plateau. *Scientific Reports*, **12**, 17239.
854 81. Pettersson, J.H.-O., Ellström, P., Ling, J., Nilsson, I., Bergström, S., González-Acuña, D.,
855 Olsen, B. and Holmes, E.C. (2020) Circumpolar diversification of the Ixodes uriae tick
856 virome. *PLoS Pathogens*, **16**, e1008759.
857 82. Elmore, M.G., Groves, C.L., Hajimorad, M., Stewart, T.P., Gaskill, M.A., Wise, K.A., Sikora,
858 E., Kleczewski, N.M., Smith, D.L. and Mueller, D.S. (2022) Detection and discovery of
859 plant viruses in soybean by metagenomic sequencing. *Virology Journal*, **19**, 149.
860 83. Bender, K.M., Svenning, M.M., Hu, Y., Richter, A., Schückel, J., Jørgensen, B., Liebner, S.
861 and Tveit, A.T. (2021) Microbial responses to herbivory-induced vegetation changes in a
862 high-Arctic peatland. *Polar Biology*, **44**, 899-911.
863 84. Allaire, J., Ellis, P., Gandrud, C., Kuo, K., Lewis, B., Owen, J., Russell, K., Rogers, J., Sese, C.
864 and Yetman, C. (2017) Package 'networkD3'. *D3 JavaScript network graphs from R*.
865 85. Yu, G., Smith, D.K., Zhu, H., Guan, Y. and Lam, T.T.Y. (2017) ggtree: an R package for
866 visualization and annotation of phylogenetic trees with their covariates and other
867 associated data. *Methods in Ecology and Evolution*, **8**, 28-36.
868 86. Wickham, H. (2011) ggplot2. *Wiley interdisciplinary reviews: computational statistics*, **3**,
869 180-185.

87. Hunter, J.D. (2007) Matplotlib: A 2D graphics environment. *Computing in science & engineering*, **9**, 90-95.
88. Waskom, M.L. (2021) Seaborn: statistical data visualization. *Journal of Open Source Software*, **6**, 3021.
89. Nayfach, S., Camargo, A.P., Schulz, F., Eloe-Fadrosh, E., Roux, S. and Kyrpides, N.C. (2021) CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nature biotechnology*, **39**, 578-585.
90. Beerenwinkel, N., Günthard, H.F., Roth, V. and Metzner, K.J. (2012) Challenges and opportunities in estimating viral genetic diversity from next-generation sequencing data. *Frontiers in microbiology*, **3**, 31515.
91. García-López, R., Vázquez-Castellanos, J.F. and Moya, A. (2015) Fragmentation and coverage variation in viral metagenome assemblies, and their effect in diversity calculations. *Frontiers in bioengineering and biotechnology*, **3**, 141.
92. Venkataraman, S., Prasad, B.V. and Selvarajan, R. (2018) RNA dependent RNA polymerases: insights from structure, function and evolution. *Viruses*, **10**, 76.
93. Mönttinen, H.A., Ravantti, J.J. and Poranen, M.M. (2021) Structure unveils relationships between RNA virus polymerases. *Viruses*, **13**, 313.
94. Zhao, G., Wu, G., Lim, E.S., Droit, L., Krishnamurthy, S., Barouch, D.H., Virgin, H.W. and Wang, D. (2017) VirusSeeker, a computational pipeline for virus discovery and virome composition analysis. *Virology*, **503**, 21-30.
95. Lefkowitz, E.J., Dempsey, D.M., Hendrickson, R.C., Orton, R.J., Siddell, S.G. and Smith, D.B. (2018) Virus taxonomy: the database of the International Committee on Taxonomy of Viruses (ICTV). *Nucleic acids research*, **46**, D708-D717.
96. King, A.M., Adams, M.J., Carstens, E.B. and Lefkowitz, E.J. (2012) Virus taxonomy. *Ninth report of the International Committee on Taxonomy of Viruses*, **9**.
97. Starr, E.P., Nuccio, E.E., Pett-Ridge, J., Banfield, J.F. and Firestone, M.K. (2019) Metatranscriptomic reconstruction reveals RNA viruses with the potential to shape carbon cycling in soil. *Proceedings of the National Academy of Sciences*, **116**, 25900-25908.
98. Chen, Y.-M., Hu, S.-J., Lin, X.-D., Tian, J.-H., Lv, J.-X., Wang, M.-R., Luo, X.-Q., Pei, Y.-Y., Hu, R.-X. and Song, Z.-G. (2023) Host traits shape virome composition and virus transmission in wild small mammals. *Cell*, **186**, 4662-4675. e4612.
99. French, R.K., Anderson, S.H., Cain, K.E., Greene, T.C., Minor, M., Miskelly, C.M., Montoya, J.M., Wille, M., Muller, C.G. and Taylor, M.W. (2023) Host phylogeny shapes viral transmission networks in an island ecosystem. *Nature Ecology & Evolution*, **7**, 1834-1843.
100. Kitchen, A., Shackelton, L.A. and Holmes, E.C. (2011) Family level phylogenies reveal modes of macroevolution in RNA viruses. *Proceedings of the National Academy of Sciences*, **108**, 238-243.
101. Carlson, C.J., Albery, G.F., Merow, C., Trisos, C.H., Zipfel, C.M., Eskew, E.A., Olival, K.J., Ross, N. and Bansal, S. (2022) Climate change increases cross-species viral transmission risk. *Nature*, **607**, 555-562.
102. Ko, K.K., Chng, K.R. and Nagarajan, N. (2022) Metagenomics-enabled microbial surveillance. *Nature Microbiology*, **7**, 486-496.

914 103. Liang, G. and Bushman, F.D. (2021) The human virome: assembly, composition and host
915 interactions. *Nature Reviews Microbiology*, **19**, 514-527.
916 104. He, W.-T., Hou, X., Zhao, J., Sun, J., He, H., Si, W., Wang, J., Jiang, Z., Yan, Z. and Xing, G.
917 (2022) Virome characterization of game animals in China reveals a spectrum of
918 emerging pathogens. *Cell*, **185**, 1117-1129. e1118.
919 105. Pan, Y.-F., Zhao, H., Gou, Q.-Y., Shi, P.-B., Tian, J.-H., Feng, Y., Li, K., Yang, W.-H., Wu, D.
920 and Tang, G. (2024) Metagenomic analysis of individual mosquito viromes reveals the
921 geographical patterns and drivers of viral diversity. *Nature Ecology & Evolution*, **8**, 947-
922 959.
923 106. Bernstein, A.S., Ando, A.W., Loch-Temzelides, T., Vale, M.M., Li, B.V., Li, H., Busch, J.,
924 Chapman, C.A., Kinnaird, M. and Nowak, K. (2022) The costs and benefits of primary
925 prevention of zoonotic pandemics. *Science Advances*, **8**, eabl4183.
926 107. Rahman, M.T., Sobur, M.A., Islam, M.S., Ievy, S., Hossain, M.J., El Zowalaty, M.E., Rahman,
927 A.T. and Ashour, H.M. (2020) Zoonotic diseases: etiology, impact, and control.
928 *Microorganisms*, **8**, 1405.
929 108. Hu, W.-S. and Hughes, S.H. (2012) HIV-1 reverse transcription. *Cold Spring Harbor*
930 *perspectives in medicine*, **2**, a006882.
931 109. Telesnitsky, A. and Goff, S. (2011) Reverse transcriptase and the generation of retroviral
932 DNA.
933 110. Yang, C.-R., King, C.-C., Liu, L.-Y.D. and Ku, C.-C. (2020) FluConvert and IniFlu: a suite of
934 integrated software to identify novel signatures of emerging influenza viruses with
935 increasing risk. *BMC bioinformatics*, **21**, 1-14.
936 111. Liao, H., Cai, D. and Sun, Y. (2022) VirStrain: a strain identification tool for RNA viruses.
937 *Genome Biology*, **23**, 38.
938 112. Wolf, Y.I., Silas, S., Wang, Y., Wu, S., Bocek, M., Kazlauskas, D., Krupovic, M., Fire, A.,
939 Dolja, V.V. and Koonin, E.V. (2020) Doubling of the known set of RNA viruses by
940 metagenomic analysis of an aquatic virome. *Nature microbiology*, **5**, 1262-1270.

941

942 **TABLE AND FIGURES LEGENDS**

943

944 **Figure 1. Overview of the RdRP Reference Database.**

945 (**A**) Compilation of representative RdRP sequences into 24 RNA virus superclades,
946 categorized based on phylogenetic analysis.

947 (**B**) Alignment of viral superclades with the ICTV taxonomic system, spanning virus
948 class, order, and family levels. Numbers reflect the count of ICTV classifications
949 mapped to each RdRP superclade, with specific ICTV representative mapping details
950 provided.

951 (**C**) The distribution of viral sequences across supergroups and ICTV taxonomy,
952 highlighting 551 vertebrate-related and 325 plant-related sequences. Data points are
953 color-coded to represent host associations, and bar plots show the proportions of human
954 pathogens and plant-associated viruses in each taxonomic group.
955 (**D**) Using the *Reoviridae* as a case study, the diagram presents a phylogenetic tree in
956 which branch colors indicate taxonomy and background hues denote host association.
957 Reference sequences are clearly annotated based on their positions within the
958 phylogenetic tree.
959
960 **Figure 2. Framework for Identifying RNA Viruses from Sequencing Samples.**
961 The VirID process for automated RNA virus detection, which comprises three main
962 stages: (i) RNA virus discovery, (ii) phylogenetic analysis, and (iii) phylogeny-based
963 virus characterization. It produces outputs that include viral sequences, phylogenetic
964 trees, and comprehensive information including sequence length, best match of
965 BLASTx comparison, virus classification, and host association.
966
967 **Figure 3. Visualization of Outputs from VirID.**
968 (**A**) Distribution of Reads Per Million (RPM) of potential viral sequences from SRA
969 sample SRR22936818, categorized by RdRP superclade and corresponding NCBI
970 taxonomic ranks.
971 (**B**) The host associations of all identified virus sequences based on their placement in
972 the phylogenetic tree.
973 (**C**) A phylogenetic tree is used to distinguish different viral lineages and colour-coded
974 by taxonomic group. In this case, newly identified virus sequences in the Nido
975 superclade are highlighted in red at the end of the branches, emphasizing their
976 significance.
977 (**D**) Complete nucleic acid sequences for detailed analysis and verification.
978 (**E**) Detailed information for each identified viral sequence, including host association,
979 genomic details, and taxonomic classification.

980

981 **Figure 4. Benchmarking of Various Tools using Simulated Data.**

982 (**A**) Performance of six bioinformatic tools across 15 simulated samples, evaluating
983 average accuracy, precision, recall, and F1 score. Error bars represent standard errors.

984 (**B**) Three tools – VirID, VirBot, VirSorter2 – had superior performance and were
985 further assessed for their specific ability to identify 'new RNA viruses' at various levels
986 of classification.

987 (**C**) Effectiveness of each tool in identifying RNA virus sequences across 15 simulated
988 samples. Yellow dots represent the number of RNA virus sequences identified, while
989 green dots indicate the number of corresponding species-level taxa. The inset graph in
990 the upper right corner adjusts the y-axis to range from 0-60 for a focused view.

991 (**D**) Ratio of the lengths of sequences identified by VirID in various simulated data
992 samples compared to the length of the corresponding gold standard sequence. Only
993 sequences that uniquely correspond to gold standard sequences are considered.

994

995 **Figure 5. Reclassification of Viruses Using VirID using Real-World Data.**

996 (**A**) Distribution of virus species across different superclades and families identified by
997 VirID in previously published data sets, including human swabs (77), wild birds (78),
998 Malayan pangolins (79) Qinghai voles (80), arthropods (81), plant (82), and soil
999 samples (83).

1000 (**B**) The utility of VirID across a broader spectrum of RNA virus families. Virus
1001 families documented in the original publications are displayed on the left, while those
1002 identified in this study are shown on the right.

1003

1004 **Figure 6. Virus-Host Associations Revealed by VirID.**

1005 (**A**) Number of known and unknown virus sequences related to humans, vertebrates,
1006 plants, and other hosts as identified by VirID in the Qinghai vole case study.

1007 (**B**) Count of virus sequences associated with different hosts in various tissue and fecal
1008 samples from voles identified by VirID.

1009 (**C**) Number and taxonomy of virus sequences related to different hosts in various real-
1010 world samples identified by VirID.
1011 (**D**) Number of virus sequences related to different hosts in a variety of real-world
1012 samples identified by VirID.
1013
1014 **SUPPLEMENTARY INFORMATION**
1015 **Supplementary Figure S1. Host Associations of Reference Viruses in 24**
1016 **Superclades.**
1017 A total of 24 phylogenetic trees where background shading signifies host type (i.e.,
1018 human, vertebrates, plants). Reference sequences are annotated according to their
1019 phylogenetic positions within the tree.
1020
1021 **Supplementary Figure S2. Pseudocode for Eliminating False Positives.**
1022 The ACC function is used to calculate the consistency between two amino acid
1023 sequences. After each multiple sequence comparison, VirID calculates the consistency
1024 between query sequence and the reference sequence to obtain false positive sequences.
1025
1026 **Supplementary Figure S3. Accuracy of the VirID Placement Method for Different**
1027 **Superclades.**
1028 The performance of superclades performing leave-one-out cross-validation, evaluating
1029 tree recall. Each point represents one experiment.
1030
1031 **Supplementary Figure S4. Composition of the Simulation Data set Candidate Pool.**
1032 The simulated pool includes 'known RNA viruses', 'new RNA viruses', and 'others'. The
1033 'known RNA viruses' were derived from the NCBI RefSeq database. The 'new RNA
1034 viruses' were subdivided into three levels based on their sequence similarity to the
1035 reference database and were sourced from recent research publications. The 'others'
1036 includes DNA viruses, eukaryotes, bacteria, and archaea.
1037

1038 **Supplementary Figure S5. Viruses And Host Associations Revealed by VirID.**
1039 (**A**) The distribution of virus species across different superclades and families identified
1040 by VirID in 20 SRA libraries.
1041 (**B**) The number of virus sequences related to humans, vertebrates, and other hosts as
1042 identified by VirID in 20 SRA libraries
1043
1044 **Supplementary Figure S6. Phylogeny of Viruses Associated with Human or**
1045 **Vertebrates.**
1046 Phylogenetic tree shows the evolutionary relationships of the *Hantaviridae,*
1047 *Coronaviridae, Paramyxoviridae,* and *Sedoreoviridae* that contain potential human
1048 pathogens. Viral sequences identified by VirID. Those labelled in red are potential
1049 human pathogens, while sequences labelled in orange are viruses associated with
1050 vertebrates.
1051
1052 **Supplementary Figure S7. Performance of Three Tools – VirBot, VirSorter2, and**
1053 **VirID – in Identifying 'new RNA viruses' and 'known RNA viruses'.**
1054 VirID recorded zero false positives in the identification of 'new RNA viruses'.
1055
1056 **Supplementary Table S1.** Mapping of the RdRp protein reference sequence database
1057 of RNA viruses to the ICTV taxonomic system.
1058
1059 **Supplementary Table S2.** Detailed information on 20 meta-transcriptomic from one
1060 previous study of bat, rodent, pangolin and zoo animal in China.
1061
1062 **Supplementary Table S3.** Detailed information on 192 public libraries from 7 studies
1063 retrieved from the SRA database.
1064
1065 **Supplementary Table S4.** Output information from the assembly and basic annotation
1066 phase.

1067

1068 **Supplementary Table S5.** Output information of the phylogenetic analysis section.

1069

1070 **Supplementary Table S6.** Detailed information on the classification results identified

1071 by VirID in seabird ticks.

1072

1073 **Supplementary Table S7.** Detailed information on possible host associations of viral

1074 sequences predicted by VirID in seven real-world cases.

1075