Hidden viral sequences in public sequencing data and warning for future emerging diseases

Junna Kawasaki[a,b#], Shohei Kojima[a*], Keizo Tomonaga[a,b,c], Masayuki Horie[a,d,e#]

[a]Laboratory of RNA Viruses, Department of Virus Research, Institute for Frontier Life and Medical Sciences, Kyoto University, Kyoto 606-8507, Japan

[b]Laboratory of RNA Viruses, Department of Mammalian Regulatory Network, Graduate School of Biostudies, Kyoto University, Kyoto 606-8507, Japan

[c]Department of Molecular Virology, Graduate School of Medicine, Kyoto University, Kyoto 606-8507, Japan

[d]Hakubi Center for Advanced Research, Kyoto University, Kyoto 606-8507, Japan

[e]Division of Veterinary Sciences, Graduate School of Life and Environmental Sciences, Osaka Prefecture University, Osaka, 599-8531, Japan

Running Head: Public data reusability to identify viral infections

#Address correspondence to Junna Kawasaki, jrt13mpmuk@gmail.com and Masayuki Horie, mhorie@vet.osakafu-u.ac.jp

*Present address: Shohei Kojima, Genome Immunology RIKEN Hakubi Research Team, RIKEN Cluster for Pioneering Research, Yokohama 230-0045, Japan

Abstract

RNA viruses cause numerous emerging diseases, mostly due to transmission from mammalian and avian reservoirs. Large-scale surveillance of RNA viral infections in these animals is a fundamental step for controlling viral infectious diseases. Metagenomic analysis is a powerful method for virus identification with low bias and has substantially contributed to the discovery of novel viruses. Deep sequencing data have been collected from diverse animals and accumulated in public databases, which can be valuable resources for identifying unknown viral sequences. Here, we screened for infections of 33 RNA viral families in publicly available mammalian and avian sequencing data and found approximately 900 hidden viral infections. We also discovered six nearly complete viral genomes in livestock, wild, and experimental animals: hepatovirus in a goat, hepeviruses in blind mole-rats and a galago, astrovirus in macaque monkeys, parechovirus in a cow, and pegivirus in tree shrews. Some of these viruses were phylogenetically close to human pathogenic viruses, suggesting the potential risk of causing disease in humans upon infection. Furthermore, the infections of five novel viruses were identified in several different individuals, indicating that their infections may have already spread in the natural host population. Our findings demonstrate the reusability of public sequencing data for surveying viral infections and identifying novel viral sequences, presenting a warning about a new threat of viral infectious disease to public health.

45 **Importance**

46 Monitoring the spread of viral infections and identifying novel viruses capable of
47 infecting humans through animal reservoirs are necessary to control emerging viral
48 diseases. Massive amounts of sequencing data collected from various animals are
49 publicly available, and these data may contain sequences originating from a wide variety
50 of viruses. Here, we analyzed more than 46,000 public sequencing data and identified
51 approximately 900 hidden RNA viral infections in mammalian and avian samples. Some
52 viruses discovered in this study were genetically similar to pathogens that cause hepatitis,
53 diarrhea, or encephalitis in humans, suggesting the presence of new threats to public
54 health. Our study demonstrates the effectiveness of reusing public sequencing data to
55 identify known and unknown viral infections, indicating that future continuous
56 monitoring of public sequencing data by metagenomic analyses would help prepare and
57 mitigate future viral pandemics.

58 **Introduction**

59 RNA viruses have caused numerous emerging diseases; for example, it was reported that
60 94 % of zoonoses from 1990 to 2010 were caused by RNA viruses (1). Mammalian and
61 avian species are especially high-risk transmission sources for zoonotic viruses because
62 of their frequent contact with humans as livestock, bushmeat, companion, or laboratory
63 animals (2). Additionally, the spread of viral infectious diseases in livestock animals
64 impacts sustainable food security and economic growth (3). Thus, large-scale
65 surveillance of RNA viral infections in these animals would help monitor infections of
66 known and unknown viruses that can cause outbreaks in humans and domestic animals.

67 Metagenomic analysis can identify viruses with low bias and has substantially
68 contributed to elucidating virus diversity for more than a decade (4). With the increase in
69 research using metagenomic analysis, new virus species, genera, and families have been
70 successively established by the International Committee on Taxonomy of Viruses (ICTV)
71 (5). However, a previous study estimated the existence of at least 40,000 mammalian viral
72 species (6), which far exceeds the number of viral species classified by the ICTV to date
73 (5, 7). Therefore, further research is needed to understand viral diversity and prepare for
74 future viral pandemics. The amount of RNA-seq data in public databases is growing
75 exponentially (8); however, only limited studies have examined for viral infections in
76 publicly available sequencing data (9-11). The public data are derived from samples with
77 various research backgrounds and may contain a wide variety of viral sequences.
78 Therefore, analyzing publicly available RNA-seq data can be an effective way to assess
79 the spread of viral infections and discover novel viruses.

80 In this study, we analyzed more than 46,000 RNA-seq data to screen hidden
81 RNA virus infections in mammalian and avian species and identified approximately 900

infections. We also discovered six nearly complete viral genomes in livestock, wild, and laboratory animals. Phylogenetic analyses showed some of the novel viruses were closely related to human pathogenic viruses, suggesting the potential risk of causing disease in humans. Furthermore, the viral infections were identified in several individuals collected by independent studies, indicating that their infections may have already spread in the natural host population. Our findings demonstrate the reusability of public sequencing data for surveying viral infections that may present a threat to public health.

91 **Results**

92 **Detection of RNA viral infections hidden in public sequence data**

93 To detect RNA viral infections in mammalian and avian RNA-seq data, we first
94 performed *de novo* sequence assembly (**Fig. 1A**). We then performed BLASTX screening
95 using contigs to extract RNA virus-derived sequences. Among 422,615,819 contigs, we
96 identified 17,060 RNA virus-derived sequences. The median length of the viral contigs
97 was 821 bp (**Fig. 1B**), which was shorter than the genomic size of RNA viruses (**Fig. 1C**).
98 These results indicate that most viral contigs were detected as partial sequences of the
99 viral genome, and several contigs may have originated from the same viral infection event.
100 Therefore, we sought to determine the viral infections in each sequencing data by the
101 alignment coverage-based method to avoid double counting (**Fig. 1A and details in**
102 **Materials and Methods**). Briefly, we constructed sequence alignments by TBLASTX
103 using the viral contigs in each RNA-seq data and reference viral genomes, and then
104 calculated the alignment coverage between the viral contigs and each viral reference
105 sequence. Here, we defined a viral infection when the alignment coverage exceeded the
106 threshold (more than 20 %). This threshold was determined using sequencing data
107 obtained from viral infection experiments (**Fig. S1 and details in Materials and**
108 **Methods**). Finally, we totalized the infections at the virus family level after excluding the
109 viruses inoculated experimentally

110 More than 46,000 mammalian and avian RNA-seq data were used to investigate
111 infections by 33 RNA virus families reported to infect vertebrates. Consequently, we
112 identified 882 infections of 22 RNA virus families in 695 sequencing data from 53 host
113 species (**Fig. 2A**). These results indicate that analyzing public sequencing data by
114 metagenomic analysis is useful for identifying hidden viral infections.

115

116 **Frequent detection of diverse virus families in bird samples**

117 Many viral infectious diseases associated with birds have been reported so far (12), such
118 as influenza A virus (13, 14) and West Nile virus (15). In this study, we frequently
119 detected viral infections in bird samples (**Fig. 2B**). The odds ratio of RNA virus detection
120 in birds compared with that in mammals was 3.3. Furthermore, among the investigated
121 species, we found relatively high viral detection rates in Gallus and Anas species at 5.4 %
122 and 7.6 %, respectively (**Fig. 2C**). We also found infections of 12 and 8 virus families in
123 Gallus and Anas species, respectively (**Fig. 2D**). These results indicate that birds,
124 especially Gallus and Anas species, are frequently infected with various virus families,
125 suggesting that these species are reservoirs for a wide variety of viruses (**see Discussion**).

126

127 **Identification of unknown reservoir hosts at virus family levels**

128 To identify novel virus-host relationships at virus family levels, we compared our data
129 with known virus-host relationships provided in the Virus-Host DB (16) (**Fig. 3A**). This
130 database lists virus-host relationships based on the identification of viral sequences from
131 a host animal. We found 50 newly identified virus-host relationships using this database
132 for comparison, and 17 of them were identified with more than 70 % alignment coverage.
133 Notably, we identified nearly complete genomic sequences classified into the family
134 *Hepeviridae* in Spalax and Galago species for the first time. These discoveries expanded
135 our understanding of hepeviral host ranges (details of the viral characteristics are
136 described in the section: "*Hepeviruses in blind mole-rats and a galago: expanding*
137 *understanding of the hepatitis E virus host range*"). A novel relationship was also
138 identified between the family *Rhabdoviridae* and Recurvirostra species. We did not

139 perform further investigations because the complete rhabdovirus genome could not be
140 obtained, although the alignment coverage was more than 70 %. Additionally, novel
141 virus-host relationships were also found in the families *Dicistroviridae*, *Iflaviridae*,
142 *Marnaviridae*, and *Nodaviridae*, suggesting that these viral host ranges are broader than
143 previously expected. It should be noted that these relationships might be due to
144 contamination from environmental viruses, because few species in these viral families
145 have been reported to infect mammals or birds (17-20) (**see Discussion**).
146

147 **Investigation of novel viruses with complete genomic sequences**

148 To identify novel sequences comparable to a complete viral genome, we simultaneously
149 analyzed sequence similarity with known viruses and the alignment coverages with
150 reference viral genomes (**Figs. 3B-C**). We found some viral sequences showing low
151 sequence similarity with known viruses and high alignment coverage, which were
152 expected to be novel viruses with a nearly complete genome. Therefore, we further
153 characterized these viral sequences by phylogenetic analyses, annotations of viral
154 genomic features, and quantification of viral reads in RNA-seq data (**Figs. 4-5 and S2-**
155 **3**). Consequently, we discovered six viruses: hepatovirus in a goat, hepeviruses in blind
156 mole-rats and a galago, astrovirus in macaque monkeys, parechovirus in a cow, and
157 pegivirus in tree shrews.
158

159 **Goat hepatovirus: the first report on hepatoviral infections in livestock animals**

160 Hepatitis A virus (HAV), belonging to the genus *Hepatovirus* of the family
161 *Picornaviridae*, can cause acute and fulminant hepatitis and is typically transmitted via
162 fecal-oral routes, including contaminated water or foods (21). The World Health

163 Organization (WHO) reported that HAV infections resulted in the death of over 7,000
164 people in 2016 (https://www.who.int/news-room/fact-sheets/detail/hepatitis-a). Here, we
165 identified a hepatoviral infection in a goat sample (**Fig. 4A**). To our knowledge, this is
166 the first report of hepatoviral infection in livestock animals.

167 We further analyzed the hepatovirus prevalence in a natural host population by
168 quantifying the viral reads in other goat RNA-seq data because this virus was initially
169 identified in only one goat sample. Among 1,593 samples, we found the viral infections
170 in nine samples from four independent studies with > 1.0 read per million reads (RPM)
171 (**Fig. 5A and Dataset S7**). The goat hepatoviral infections were detected in liver and lung
172 samples, suggesting that the goat hepatovirus can infect tissues other than the liver.
173 Although the lungs are not considered preferential tissues for hepatoviral replication, a
174 previous report also detected hepatoviral RNAs in the seal lungs (22). The infected goat
175 samples were collected in East Asia, including China and Mongolia. Therefore, goat
176 hepatoviruses may be prevalent in the natural host population, suggesting this virus can
177 be a new threat to public health through the contamination of water and foods by infected
178 animals.

179

180 **Hepeviruses in blind mole-rats and a galago: expanding understanding of the**
181 **hepatitis E virus host range**

182 Several million infections of hepatitis E virus (HEV) are estimated to occur worldwide;
183 the WHO reported approximately 44,000 deaths due to HEV infection in 2015
184 (https://www.who.int/news-room/fact-sheets/detail/hepatitis-e). Here, we found
185 hepeviruses, classified into the same viral family as HEV, in blind mole-rats and a galago
186 for the first time (**Fig. 3A**). Phylogenetic analysis indicated that these hepeviruses formed

187 a single cluster with moose HEV (23) and members of Orthohepevirus A that infect
188 humans, pigs, rabbits, and camels (24) (**Fig. 4B**). However, the hepeviruses identified in
189 this study appeared to have an early divergence from the HEV common ancestor. These
190 results suggest a high diversity and broader host range of HEV-like viruses.

191 The blind mole-rat hepevirus was identified in the host livers, which coincides
192 with the tissue tropism of HEV (25). Additionally, we found that the 3'-portion of the
193 blind mole-rat hepevirus genome was highly transcribed (**Fig. S3B**), suggesting the
194 transcription of subgenomic RNAs (26). In contrast, we could not determine the tissues
195 infected by the galago hepevirus because the relevant metadata were not available.
196 Further, we did not observe a clear read-mapping pattern that suggests any subgenomic
197 RNA transcription in the galago sample (**Fig. S3C**).

198 We also investigated the spread of these hepeviruses in a natural population
199 using RNA-seq data from blind mole-rats and galagos. Among 91 RNA-seq data from
200 blind mole-rats, we detected the hepeviral infections in six samples (**Fig. 5B**). These
201 infected individuals were captured and kept as laboratory animals in Israel by the same
202 research (**Dataset S8**). There were two possibilities about when the hepeviruses have
203 infected blind mole-rats: the hepeviruses had already infected these blind mole-rats when
204 they were captured, or the viral infections had spread during the maintenance of these
205 individuals in the laboratory. To explore these possibilities, we investigated the inter-
206 individual diversity of the hepevirus sequences and found that these blind mole-rats were
207 infected with relatively diverse hepeviruses representing nucleotide sequence identities
208 ranging from 83.6 % to 99.5 % (**Fig. 5C**). These results suggested that several individuals
209 had already been infected with distinct hepeviruses in the wild before being captured. The
210 galago hepeviral infections were detected in only two samples originating from a study

211 in which we first identified the virus (**Dataset S9**). This may be simply because only four
212 galago RNA-seq data obtained from the same study were available. Taken together, we
213 suggest that these hepeviruses can become a new threat to public health, similar to HEV.
214

215 **MLB-like astrovirus detected in macaque monkeys with chronic diarrhea**

216 We found astroviruses genetically similar to human astrovirus MLB (HAstV-MLB) in
217 fecal samples of macaque monkeys (**Fig. 4C**). Although HAstV-MLB infections are
218 typically asymptomatic (27, 28), several studies have reported the viral detection in cases
219 with diarrhea (29), encephalitis (30), or meningitis (31). Interestingly, the macaque MLB-
220 like astrovirus was found in macaque monkeys with chronic diarrhea. We analyzed the
221 viral read amounts in the patient (n = 12) and control (n = 12) monkeys to assess the
222 association between MLB-like astroviral infections and symptom prevalence (**Fig. 5D**
223 **and Dataset S10**). Abundant MLB-like astroviral reads were detected in two patients,
224 suggesting that the viral infections are associated with host symptoms. However, we did
225 not observe the viral infection in other patients; further, we found the infection in a control
226 individual, although the viral read amount was approximately 100 times less than those
227 of the patients. Additionally, a previous study reported that monkeys, in which partial
228 sequences of MLB-like astroviruses were detected, had no obvious clinical signs,
229 including diarrhea (32). Thus, further experiments are needed to clarify the pathogenesis
230 of MLB-like astrovirus. Considering that there is no current experimental system for
231 examining HAstV-MLB infections (28), our findings suggest that macaque monkeys can
232 be used as animal model systems for researching MLB-like astroviruses.
233

234 **Silent infections of bovine parechovirus having a broad tissue tropism**

235 Human parechovirus infection is especially problematic in infants and young children.
236 Although most parechovirus infections are considered asymptomatic, their infections
237 have been reported in patients with respiratory, digestive, and central nervous system
238 disorders (33). In this study, we identified a parechovirus, classified into the family
239 *Picornaviridae*, in the lower digestive tract of a cow (**Fig. 4D**). Despite the broad host
240 range of parechovirus, including mammals, birds, and reptiles (34), to our knowledge,
241 this is the first report on parechovirus infections in livestock animals.

242 Phylogenetic analysis indicated that this parechovirus was closely related to the
243 falcon parechovirus, a member of Parechovirus E. Next, we compared the bovine
244 parechovirus with the ICTV species demarcation criteria (34) to investigate whether this
245 virus is a novel species (**Fig. S2B**). Consequently, we found that the bovine parechovirus
246 was distant enough from other known parechovirus species and could be considered a
247 separate species based on the following criteria: divergence of amino acid sequences in
248 polyprotein (37.8 %), P1 protein (37.8 %), and 2C+3CD (29.9 %) protein. Therefore, we
249 propose that this virus belongs to a new species in the genus *Parechovirus*.

250 We also investigated the prevalence of this parechovirus infection in a natural
251 host population using public cow RNA-seq data (**Fig. 5E and Dataset S11**). Among
252 8,284 samples, we detected the parechovirus infections in 944 samples from eight
253 independent studies with > 1.0 RPM. The viral infections were detected in various tissues,
254 such as the digestive, lymphatic, and central nervous systems. These results suggest a
255 broad tissue tropism of the bovine parechovirus. To assess the parechovirus pathogenicity,
256 we analyzed the viral prevalence among 36 or 44 samples with a diagnosis for a
257 gastrointestinal disorder or respiratory lesion, respectively. We did not observe a
258 significant association between the viral infections and the presence/absence of

259 symptoms in these two studies (**Fig. 5F**). These results indicate that bovine parechovirus
260 infections may be asymptomatic, similar to the typical outcome of human parechoviral
261 infections. Furthermore, this also suggests that infected cows can spread parechoviral
262 infections as silent reservoirs.
263

264 **Geographical expansion of tree shrew pegivirus infection associated with host**
265 **migration**

266 We found a pegivirus belonging to the genus *Pegivirus* of the family *Flaviviridae* in tree
267 shrew liver samples. Phylogenetic analysis indicated that this pegivirus was closely
268 related to the Pegivirus G identified in various bat species (**Fig. 4E**). According to the
269 ICTV species demarcation criteria (35), this virus appeared to be the same species as
270 Pegivirus G because the amino acid sequence identity in the NS5B gene was 70.9 % (**Fig.**
271 **S2C**). These results indicate that Pegivirus G can infect distinct host lineages: tree shrews
272 and bats.

273 We also investigated the pegiviral infections in other tree shrew samples by read
274 mapping analysis. Among the 59 samples, the pegiviral infections were detected in four
275 samples collected from a research colony in the United Kingdom (**Dataset S12**). A recent
276 report partially identified a pegiviral sequence (GenBank: MT085214.1) in tree shrews
277 collected in Southeast Asia (36), which showed 84.9 % nucleotide sequence identity to
278 the pegivirus identified in this study (**Fig. 4E**). These results indicate that tree shrew
279 pegivirus infections were found in both Asia and Europe, suggesting an expanding
280 geographic distribution of Pegivirus G along with host animal transportation as
281 experimental resources. Thus, the global trade of host animals may lead to spreading
282 pegiviral infections hidden in tree shrews.

283 **Discussion**

284 Metagenomic analysis is a powerful approach for surveying viral infections (4, 5).
285 Extensive deep sequencing data have been accumulated in public databases, which can
286 be used for identifying viral infections. In this study, we analyzed the publicly available
287 RNA-seq data to search for hidden RNA viral infections in mammalian and avian species
288 and subsequently identified approximately 900 infections by 22 RNA virus families (**Figs.**
289 **1 and 2**). These results indicate that reusing public sequencing data is a cost-effective
290 approach for identifying viral infections. Furthermore, we discovered six novel viral
291 genomes in livestock, wild, and experimental animals (**Fig. 4**). Some of these viruses
292 were detected in different individuals, suggesting that the viral infections may have
293 already spread in the natural host population (**Fig. 5**). Overall, our work demonstrates the
294 reusability of public sequencing data for surveying infections by both known and
295 unknown viruses.

296 In this study, we determined viral infections by a combination of sequence
297 assembly and the alignment coverage-based method to solve several issues in viral
298 metagenomic analysis (**Fig. 1A**). One of the problems is detecting infections in data with
299 a small number of viral reads because almost all public sequencing data were collected
300 without using virus enrichment strategies. The result that most virus contigs were shorter
301 than the reference viral genomes reflects this difficulty (**Figs. 1B-C**). To resolve this issue,
302 we determined viral infections by the alignment coverage-based method, which uses
303 relatively short viral sequences as clues (**Figs. 1A and S1**). Consequently, we succeeded
304 in detecting approximately 900 RNA viral infections in public sequencing data (**Fig. 2A**).
305 Another problem in viral metagenomic analysis is the viral detectability depending on
306 sequence similarity with known viruses. We here discovered six nearly complete viral

307 genomes (**Fig. 4**) by sequence assembly and BLAST screening (**Fig. 1A**). Notably, these
308 viral infections were undetectable in almost all samples, even at the virus family and
309 genus levels, by the NCBI SRA Taxonomy Analysis Tool (https://github.com/ncbi/ngs-
310 tools/tree/tax/tools/tax), which determines the taxonomic composition of reads in the
311 RNA-seq data without sequence assembly (**Dataset S7-S12**). These results indicate that
312 our method can identify novel viruses with full-length genomes, which would effectively
313 elucidate virus diversity. Taken together, our strategy using sequence assembly and the
314 alignment coverage-based method can efficiently detect known and unknown viral
315 infections in publicly available sequencing data.

316 However, there are still several challenges for identifying viral infections in
317 public sequencing data. First, we could not determine complete viral sequences mostly
318 (**Figs. 3B and 3C**). Further improvement in sequence assembly efficiency (37) or
319 integrative analysis using short- and long-read sequence datasets (38) can solve this
320 problem. Second, there may be a bias in virus detection using public sequencing data
321 depending on their genomic types. Among the 882 viral infections identified in this study,
322 77.0% were positive-sense single-stranded RNA (ssRNA(+)) viral infections, whereas
323 11.5% and 11.5% were double-stranded RNA and negative-sense single-stranded RNA
324 viral infections (**Fig. 2A**). The RNA-seq procedure, such as enrichment of polyadenylated
325 (poly-A) transcripts, can be relevant to this bias because many ssRNA(+) viruses have a
326 poly-A tract at the 3'-end of their genome (39). Alternatively, this bias may result from a
327 repertoire of reference viral genomes used for the viral search (**Fig. 1C**), which can be
328 solved in the future by database expansion. Thirdly, our method demands relatively
329 abundant computational resources, including operation time, for determining viral
330 infections in each RNA-seq data. We reconstructed viral sequences from RNA-seq data

331 according to several steps: mapping analysis for excluding host transcripts, de novo
332 sequence assembly using unmapped reads, and BLASTX screening for identifying viral
333 sequences (**Fig. 1A**). In contrast, another study performed a search for viral RNA-
334 dependent RNA polymerase proteins in translated nucleotide sequences, which enabled
335 the authors to screen for viral infections in approximately 5.7 million public sequencing
336 data within 11 days (11). Considering that the number of public sequencing data will keep
337 increasing, platform development and maintenance, which can save computational
338 resources, would be necessary for continuing such viral surveillance.

339 Another challenge in viral metagenomic analysis using public data is
340 distinguishing true viral infections from contamination. To address this issue, we
341 performed integrative analyses using sample metadata and sequence information,
342 including sequence similarity and alignment coverage with known viruses (**details in**
343 **Materials and Methods**). Consequently, we found several possible contamination cases:
344 influenza A virus in a Myotis bat, vesicular stomatitis Indiana virus (VSV) in cultured
345 chicken cells, mammalian rubulavirus 5 (PIV5) in cultured cells and quail egg samples,
346 and Kadipiro virus (KDV) in rat samples (**Fig. 3A and Dataset S3**). For example, the
347 influenza A viral nucleotide sequence identified in a bat sample showed 100 % similarity
348 to a laboratory strain of influenza A virus (A/WSN/1933(H1N1)). Considering that the
349 bat sample was collected in 2012, it is difficult to expect that such a highly similar
350 influenza A virus was maintained for approximately 80 years. Likewise, the infections of
351 VSVs and PIVs were also identified with approximately 100 % sequence similarity to the
352 reference viral sequences (**Dataset S3**). VSV is frequently used as experimental tools; for
353 example, as a pseudotype virus (40). Previous studies also have reported possible
354 contamination of PIV5 in cultured cells (41, 42). Additionally, it was reported that KDV

355 RNA might be a contaminant in the nucleic acid extraction kit (43). Therefore, we
356 excluded these viral infections to avoid counting false positives. These cases emphasize
357 the importance of multilayered validations for viral infections found by viral
358 metagenomic analysis alone.

359 Further research efforts to elucidate viral diversity are necessary to prepare for a
360 possible future viral pandemic (1, 5). A strategic approach, such as determining the host
361 samples used for virus search based on the expectation of viral infection frequency or
362 viral diversity, would be necessary. It has been discussed that birds may be high-risk viral
363 hosts of zoonoses because of their high species diversity and wide habitat range (12). In
364 this study, we found that viral infections were more frequently detected in birds,
365 especially Gallus and Anas species (**Figs. 2B-D**). Furthermore, among 217 viral
366 infections identified in Gallus and Anas samples, 78 infections (35.9 %) showed less than
367 95 % amino acid sequence similarity with known viruses, suggesting that these sequences
368 may be derived from unknown viruses. Therefore, further viral metagenomic analyses
369 targeting bird samples may effectively detect viral infections, including unknown ones.

370 In conclusion, we demonstrated the reusability of public sequencing data for
371 monitoring viral infections and discovering novel viral sequences, and elucidated diverse
372 RNA viruses hidden in animal samples. Our findings also emphasize the necessity of
373 continuous surveillance for viral infections using public sequencing data to prepare for
374 future viral pandemics, as well as the importance of developing a fundamental
375 bioinformatics platform for surveillance (11, 44).

376

377 **Materials and Methods**

378 **Sequence assembly using publicly available RNA-seq data**

379 We collected RNA-seq data of 41,332 mammals (169 genera and 228 species) and 5,027
380 birds (70 genera and 83 species) from the NCBI Sequence Read Archive (SRA) database
381 (8) during June and July 2019 according to the following search conditions:
382 [("Mammalia"[Organism] OR "Mammals"[All Fields]) AND ("biomol rna"[Properties]
383 AND "library layout paired"[Properties] AND "filetype fastq"[Properties]) NOT ("Homo
384 sapiens"[Organism]) NOT ("Mus musculus"[Organism])] and [("Aves"[Organism] OR
385 "Aves"[All Fields]) AND ("biomol rna"[Properties] AND "library layout
386 paired"[Properties] AND "filetype fastq"[Properties])]. The RNA-seq data were
387 downloaded from the NCBI SRA database by pfastq-dump
388 (https://github.com/inutano/pfastq-dump) and preprocessed using fastp (version 0.20.0)
389 (45) with options “-l 35”, “-y -3”, “-W 3”, “-M 15”, and “-x”.

390 Sequence assembly was conducted by 1) mapping reads to the host or sister
391 species genome and 2) *de novo* assembly of sequences using unmapped reads. First, we
392 performed a mapping analysis to exclude the reads originating from host transcripts and
393 endogenous viral elements. We mapped the reads in each RNA-seq data to the host
394 genome by HISAT2 (version 2.1.0) (46) with the default parameters or used the sister
395 species genomes of the host in the same genus when the host genome data were not
396 available. Unmapped reads were extracted by Samtools (version 1.9) (47) and Picard
397 (version 2.20.4) (http://broadinstitute.github.io/picard). When the relevant genome data
398 were unavailable, the preprocessed reads were directly used for sequence assembly.
399 Sequence assembly was conducted by SPAdes (version 3.13.0) (48) or metaSPAdes
400 (version 3.13.0) (49) with *k*-mers of 21, 33, 55, 77, and 99. Finally, we excluded contigs

401 with lengths shorter than 500 bp by Seqkit (version 0.9.0) (50) and then clustered the
402 contigs showing 95.0 % nucleotide sequence similarity by cd-hit-est (version 4.8.1) (51).
403 Consequently, we obtained 422,615,819 contigs and used them for subsequent analyses.
404 We listed the SRA Run accession numbers, genome files used for mapping analysis, and
405 sequence assembly tools in **Dataset S1**.
406

407 **Identification of contigs originating from RNA viruses**

408 To determine the origins of the contigs, we analyzed the sequence similarity between the
409 contigs and known sequences in BLASTX screening (version 2.9.0) (52). First, we
410 performed BLASTX searches with the options "-word_size 2", "-evalue 1E-3", and
411 "max_target_seqs 1" using a custom database consisting of RNA viral proteins. We
412 constructed the custom database by downloading the viral protein sequences of the realm
413 *Riboviria* from the NCBI GenBank (version: 20190102) (53) and clustering the sequences
414 showing 98.0 % similarity by cd-hit (version 4.8.1). Second, to confirm that the contigs
415 are not derived from organisms other than viruses, we further performed BLASTX
416 searches with the options "-word_size 2", "-evalue 1E-4", and "-max_target_seqs 10"
417 using the NCBI nr database (versions: 20190825-20190909 were used for screening
418 contigs in mammalian data and versions: 20190330-20190403 were used for screening
419 contigs in avian data). We determined the contig origins by comparing the bitscores in
420 the first and second BLASTX screening. Consequently, we obtained 17,060 contigs that
421 were deduced to encode RNA viral proteins.
422

423 **Totalization of RNA viral infections in public RNA-seq data**

424 Since most viral contigs were shorter than the reference viral genomic sizes (**Figs. 1B-C**),
425 we sought to determine viral infections based on the alignment coverage-based method
426 (**Fig. 1A**). First, we performed sequence alignment by TBLASTX (version 2.9.0) using
427 viral contigs from the RNA-seq data and complete viral genomes in the NCBI RefSeq
428 genomic viral database (version 20200824). Next, we calculated the alignment coverage
429 with the genome of each viral species: the proportion of aligned sites in the entire
430 reference viral genome. In this study, we considered that an infection of the viral family
431 is present if the alignment coverage was greater than 20 %. Validation of this totalization
432 method and evaluation of the criteria are described in the next section (**Fig. S1**).
433 Furthermore, we manually checked sequences with more than 70 % alignment coverage
434 and more than 95 % identity with known viruses in the TBLASTX alignment to examine
435 possible contamination with laboratory viral strains, as well as experimentally inoculated
436 viruses. We excluded experimentally inoculated viral infections (**Dataset S2**) and
437 possible contamination (**Dataset S3**) from the final totalization (**Fig. 2A**).
438

439 **Validation of the procedure used to totalize viral infections**

440 Using samples obtained from viral infection experiments, we first compared the
441 alignment coverage-based method with that based on viral read amounts in order to
442 validate the detection rate of viral infections of our method (**Fig. S1 and Dataset S2**). We
443 obtained the read amounts derived from experimentally infected viruses from the NCBI
444 SRA Taxonomy Analysis Tool results (https://github.com/ncbi/ngs-
445 tools/tree/tax/tools/tax). The calculation procedure for alignment coverage between viral
446 contigs in each RNA-seq data and viral reference genomes is described in the previous
447 section. We observed a positive correlation between the alignment coverage and viral

448 read amounts (Pearson's correlation coefficient: 0.19, p-value: 1.87E-6) (**Fig. S1A**).
449 Among the samples collected from experiments of viral infections, the true-positive rate
450 (the detection rate of experimentally inoculated viruses) was 88.3 %, and the false-
451 positive rate (the rate that mock samples were determined to be infected samples) was
452 62.5 % when we used 20 % alignment coverage as the criterion for determining viral
453 infections (**Fig. S1B**). The relatively high false-positive rate may be due to similar
454 amounts of viral reads in some mock samples as those in infected samples (**Fig. S1A**).
455 Next, we analyzed the association between alignment coverages and viral genome size
456 (**Fig. S1C**) because the detectability of viral infections in our method may depend on the
457 reference viral genome size. As expected, we observed a tendency for viruses with small
458 genomes to be detected with relatively high alignment coverage. However, more than
459 80 % of experimentally infected viral infections were detected with more than 20 %
460 alignment coverage, regardless of the viral genome size. Based on these results, we
461 established the alignment coverage of 20 % to totalize the viral infections. Consequently,
462 we identified a total of 1,410 RNA viral infections, including 503 infections in samples
463 from viral infectious experiments (**Fig. S1D**).
464

465 **Collection of information on experimentally infected viruses**

466 To exclude experimentally infected viruses from the final totalization, we investigated
467 the experimental background of RNA-seq data. We first collected the experimental
468 descriptions of RNA-seq data: title and abstract from the NCBI BioProject database (54).
469 Then, we manually checked the terms relevant to viral infections in the descriptions,
470 focusing on viral name abbreviations and viral vector usage. We listed the obtained
471 information about viral infection experiments in **Dataset S2**.

472

473 **Summarization of virus-host relationships**

474 To identify novel reservoir hosts at the viral family levels, we compared the virus-host
475 relationships identified in this study with the dataset provided by the Virus-Host DB
476 (version: 20200629) (16). We defined a "novel virus-host relationship" as one in which
477 the viral sequence has not been reported in the host. The virus-host relationships at the
478 viral family level were categorized as 1) a novel relationship detected with > 70 %
479 alignment coverage, 2) a novel relationship detected with ≤ 70 % alignment coverage, 3)
480 a known relationship that was also detected in this study, 4) a known relationship that
481 was not identified in this study, 5) a relationship unreported so far, and 6) a novel
482 relationship, which was possibly derived from contamination (**see Discussion**). To avoid
483 misclassification of the relationships, we analyzed reports manually by searching the
484 NCBI PubMed and Nucleotide databases using the combination of the host genus and
485 viral family names: for example, ["Pan" AND "Picobirnaviridae"]. The results of the
486 manual curation are listed in **Dataset S4**.

487

488 **Characterization of viral genomic architectures**

489 Open reading frames (ORFs) and polyadenylation signals in the viral genomes were
490 predicted by the SnapGene software (snapgene.com). The positions of mature proteins,
491 frameshift signal sequences, and subgenomic RNA promoter sequences were predicted
492 based on sequence alignment using novel and reference viral sequences. The sequence
493 alignments were constructed by MAFFT (version 7.407) (55) with the option "--auto".
494 The reference viral sequences used for the genome annotations are listed in **Dataset S5**.

The viral sequences identified in this study are registered under the following accession numbers: BR001715-BR001732 and BR001751.

Phylogenetic analyses

Multiple sequence alignments (MSAs) of picornaviral P1 nucleotide sequences for **Fig. 4A**, hepeviral ORF1 amino acid sequences for **Fig. 4B**, picornaviral 3D nucleotide sequences for **Fig. 4D**, and flaviviral NS5 nucleotide sequences for **Fig. 4E** were obtained from the ICTV resources (the family of *Picornaviridae*: https://talk.ictvonline.org/ictv-reports/ictv_online_report/positive-sense-rna-viruses/picornavirales/w/picornaviridae/714/resources-picornaviridae, the family of *Hepeviridae*: https://talk.ictvonline.org/ictv-reports/ictv_online_report/positive-sense-rna-viruses/w/hepeviridae/731/resources-hepeviridae, and the family of *Flaviviridae*: https://talk.ictvonline.org/ictv-reports/ictv_online_report/positive-sense-rna-viruses/w/flaviviridae/371/resources-flaviviridae). Because an astroviral MSA was not available in the ICTV resources, we extracted astroviral ORF2 amino acid sequences from RefSeq protein viral database (version 20201007). The MSAs of reference and novel viral sequences were constructed by MAFFT with options "--add" and "--keeplength". The astroviral MSA was trimmed by excluding sites where > 20 % of the sequences were gaps and subsequently removing sequences with less than 80 % of the total alignment sites. Phylogenetic trees were constructed by the Maximum likelihood method using IQTREE (version 1.6.12) (56). The substitution models were selected based on the Bayesian information criterion provided by ModelFinder (57): GTR+R8 for **Fig. 4A**, LG+F+R4 for **Fig. 4B**, LG+F+R5 for **Fig. 4C**, TVM+R9 for **Fig.4D**, and GTR+R7 for **Fig. 4E**. The branch supportive values were measured as the ultrafast bootstrap by

519 UFBoot2 (58) with 1,000 replicates. Tree visualization was performed by the ggtree
520 package (version 2.2.1) (59). Sequence accession numbers used for the phylogenetic
521 analyses are listed in **Dataset S5**.
522
523 **Comparison with the ICTV species demarcation criteria**
524 To assess whether the viruses identified in this study could be assigned to a novel species,
525 we compared their genetic distance with known viruses according to the ICTV species
526 demarcation criteria (34, 35) (**Fig. S2**). Amino acid sequences of the P1 and 3CD genes
527 in hepatoviruses and parechoviruses were extracted by referring to Hepatovirus A
528 (GenBank: M14707.1) and Parechovirus A (GenBank: S45208.1), respectively. Amino
529 acid sequences of the NS3 and NS5B genes in pegiviruses were extracted by referring to
530 Pegivirus A (GenBank: U22303.1). We constructed MSAs using these reference and
531 novel viral sequences by MAFFT with the option "--auto". We did not analyze other
532 viruses identified in this study because the ICTV did not provide criteria based on the
533 genetic distance. The sequence accession numbers used for these analyses are listed in
534 **Dataset S5**.
535
536 **Mapping analyses using viral genomes identified in this study**
537 To verify the quality of sequence assembly, we mapped the reads in the RNA-seq data,
538 in which a novel viral sequence was identified, to the viral genomes by STAR (version
539 2.7.6a) (60) (**Fig. S3**). The genome indexes were generated with the option "--
540 genomeSAindexNbases" according to each viral genomic size, and mapping analysis was
541 conducted with the options "--chimSegmentMin 20". The number of mapped reads in

542 each position was counted by Bedtools genomecov (version 2.27.1) (61) with the options
543 "-d" and "-split".
544 To identify novel viral infections in other individuals, we analyzed the publicly
545 available RNA-seq data of the host animals by quantifying viral reads (**Figs. 5A, B, and**
546 **5E**). We investigated 1,593 goat, 91 blind mole-rat, four galago, 8,282 cow, and 59 tree
547 shrew data for infections of goat hepatovirus, blind mole hepevirus, galago hepevirus,
548 bovine parechovirus, and tree shrew pegivirus, respectively. Mapping analyses were
549 performed using STAR (version 2.7.6a) as described above. The number of total and
550 mapped reads was extracted by Samtools (version 1.5). We considered that there was a
551 viral infection in the sample if the RPM was > 1.0.
552 We compared the viral read amounts between the patient and control monkeys
553 to investigate the association between chronic diarrhea and MLB-like astrovirus infection
554 (**Fig. 5D**). Viral read amounts were quantified as described above. The average RPM for
555 each individual is plotted in **Fig. 5D** because six samples were collected from each
556 individual. **Dataset S6** shows the SRA Run accession number used to investigate novel
557 viral infections. **Datasets S7-S12** list sample metadata in which the novel viral infections
558 were detected.
559
560 **Comparison of hepeviral sequences identified in different blind mole-rats**
561 We compared nucleotide sequence identities among the hepeviral sequences found in five
562 different individuals to predict when these viruses infected the blind mole-rats. The
563 sequence comparison was performed by BLASTN (version 2.11.0) with default
564 parameters. Because most hepeviral sequences were detected as short contigs, sequence
565 identities were represented by the percentage of identical matches in the longest aligned

566 region between the hepeviral sequences (**Fig. 5C**). We also analyzed the total of aligned
567 length between contigs identified in each individual and the hepeviral genome identified
568 in ERR1742977 and confirmed that these contigs covered 86.0-99.9% of the blind mole-
569 rat hepevirus genome.
570
571 **Data Availability**
572 Bioinformatics tools and their versions are listed in **Dataset S13**.
573

Acknowledgments

We thank Jumpei Ito (Institute of Medical Science, The University of Tokyo, Japan) and Dr. Keiko Takemoto (Institute for Virus Research, Kyoto University, Japan) for their technical support. We are grateful to Dr. Nicholas F Parrish (RIKEN Hakubi Research Team, RIKEN Cluster for Pioneering Research, Yokohama, Japan), Bea Clarise Garcia, Yahiro Mukai, Hsien Hen Lin, and Koichi Kitao (Institute for Frontier Life and Medical Sciences, Kyoto University) for helpful discussions. We thank Editage (http://www.editage.com) for editing and reviewing this manuscript for English language.

This study was supported by JSPS KAKENHI JP19J22241 (JK), JP18K19443 (MH); MEXT KAKENHI JP17H05821 (MH) and JP19H04833 (MH); Hakubi project at Kyoto University (MH). Computations were partially performed on the supercomputing systems: SHIROKANE (Human Genome Center, the Institute of Medical Science, The University of Tokyo) and the NIG supercomputer (ROIS National Institute of Genetics).

Competing interests

The authors declare that they have no competing interests.

Author contributions

MH and JK conceived the study; JK and MH mainly performed bioinformatics analyses; SK supported bioinformatics analyses; JK and MH prepared the figures and wrote the initial draft of the manuscript; all authors designed the study, interpreted data, revised the paper, and approved the final manuscript.

598 **References**

599 1. Carroll D, Daszak P, Wolfe ND, Gao GF, Morel CM, Morzaria S, Pablos-Méndez
600 A, Tomori O, Mazet JAK. 2018. The Global Virome Project. Science 359:872-874.
601 2. Karesh WB, Dobson A, Lloyd-Smith JO, Lubroth J, Dixon MA, Bennett M, Aldrich
602 S, Harrington T, Formenty P, Loh EH, Machalaba CC, Thomas MJ, Heymann DL.
603 2012. Ecology of zoonoses: natural and unnatural histories. The Lancet 380:1936-
604 1945.
605 3. Otte M, Nugent R, McLeod A. 2004. Transboundary animal diseases: Assessment
606 of socio-economic impacts and institutional responses. Rome, Italy: Food and
607 Agriculture Organization (FAO):119-126.
608 4. Zhang Y-Z, Chen Y-M, Wang W, Qin X-C, Holmes EC. 2019. Expanding the RNA
609 Virosphere by Unbiased Metagenomics. Annual Review of Virology 6:119-139.
610 5. Greninger AL. 2018. A decade of RNA virus metagenomics is (not) enough. Virus
611 Research 244:218-229.
612 6. Carlson CJ, Zipfel CM, Garnier R, Bansal S. 2019. Global estimates of mammalian
613 viral diversity accounting for host sharing. Nature Ecology & Evolution 3:1070-
614 1075.
615 7. Gorbalenya AE, Krupovic M, Mushegian A, Kropinski AM, Siddell SG, Varsani
616 A, Adams MJ, Davison AJ, Dutilh BE, Harrach B, Harrison RL, Junglen S, King
617 AMQ, Knowles NJ, Lefkowitz EJ, Nibert ML, Rubino L, Sabanadzovic S,
618 Sanfaçon H, Simmonds P, Walker PJ, Zerbini FM, Kuhn JH, International
619 Committee on Taxonomy of Viruses Executive C. 2020. The new scope of virus
620 taxonomy: partitioning the virosphere into 15 hierarchical ranks. Nature
621 Microbiology 5:668-674.

622 8. Leinonen R, Sugawara H, Shumway M. 2011. The Sequence Read Archive.
623 Nucleic Acids Research 39:D19-D21.
624 9. Iwamoto M, Shibata Y, Kawasaki J, Kojima S, Li Y-T, Iwami S, Muramatsu M,
625 Wu H-L, Wada K, Tomonaga K, Watashi K, Horie M. 2021. Identification of novel
626 avian and mammalian deltaviruses provides new insights into deltavirus evolution.
627 Virus Evolution 7.
628 10. Horie M, Akashi H, Kawata M, Tomonaga K. 2020. Identification of a reptile
629 lyssavirus in Anolis allogus provided novel insights into lyssavirus evolution. Virus
630 Genes doi:10.1007/s11262-020-01803-y.
631 11. Edgar RC, Taylor J, Altman T, Barbera P, Meleshko D, Lin V, Lohr D, Novakovsky
632 G, Al-Shayeb B, Banfield JF, Korobeynikov A, Chikhi R, Babaian A. 2020.
633 Petabase-scale sequence alignment catalyses viral discovery. bioRxiv
634 doi:10.1101/2020.08.07.241729:2020.08.07.241729.
635 12. Nabi G, Wang Y, Lü L, Jiang C, Ahmad S, Wu Y, Li D. 2021. Bats and birds as
636 viral reservoirs: A physiological and ecological perspective. Science of The Total
637 Environment 754:142372.
638 13. Olsen B, Munster VJ, Wallensten A, Waldenstrom J, Osterhaus ADME, Fouchier
639 RAM. 2006. Global Patterns of Influenza A Virus in Wild Birds. Science 312:384-
640 388.
641 14. Lycett SJ, Duchatel F, Digard P. 2019. A brief history of bird flu. Philosophical
642 Transactions of the Royal Society B: Biological Sciences 374:20180257.
643 15. Habarugira G, Suen WW, Hobson-Peters J, Hall RA, Bielefeldt-Ohmann H. 2020.
644 West Nile Virus: An Update on Pathobiology, Epidemiology, Diagnostics, Control
645 and “One Health” Implications. Pathogens 9:589.

646 16. Mihara T, Nishimura Y, Shimizu Y, Nishiyama H, Yoshikawa G, Uehara H,
647 Hingamp P, Goto S, Ogata H. 2016. Linking Virus Genomes with Host Taxonomy.
648 Viruses 8:66.
649 17. Scherer WF, Verna JE, Richter GW. 1968. Nodamura Virus, an Ether- and
650 Chloroform-Resistant Arbovirus from Japan *. The American Journal of Tropical
651 Medicine and Hygiene 17:120-128.
652 18. Reuter G, Pankovics P, Gyöngyi Z, Delwart E, Boros Á. 2014. Novel dicistrovirus
653 from bat guano. Archives of Virology 159:3453-3456.
654 19. Greninger AL, Jerome KR. 2016. Draft Genome Sequence of Goose Dicistrovirus.
655 Genome Announcements 4:e00068-16.
656 20. Yinda CK, Zeller M, Conceição-Neto N, Maes P, Deboutte W, Beller L, Heylen E,
657 Ghogomu SM, Van Ranst M, Matthijnssens J. 2016. Novel highly divergent
658 reassortant bat rotaviruses in Cameroon, without evidence of zoonosis. Scientific
659 Reports 6:34209.
660 21. Lemon SM, Walker CM. 2019. Hepatitis A Virus and Hepatitis E Virus: Emerging
661 and Re-Emerging Enterically Transmitted Hepatitis Viruses. Cold Spring Harbor
662 Perspectives in Medicine 9:a031823.
663 22. Anthony SJ, St. Leger JA, Liang E, Hicks AL, Sanchez-Leon MD, Jain K,
664 Lefkowitch JH, Navarrete-Macias I, Knowles N, Goldstein T, Pugliares K, Ip HS,
665 Rowles T, Lipkin WI. 2015. Discovery of a Novel Hepatovirus (Phopivirus of
666 Seals) Related to Human Hepatitis A Virus. mBio 6:e01180-15.
667 23. Lin J, Norder H, Uhlhorn H, Belák S, Widén F. 2014. Novel hepatitis E like virus
668 found in Swedish moose. Journal of General Virology 95:557-570.

669 24. Purdy MA, Harrison TJ, Jameel S, Meng XJ, Okamoto H, Van Der Poel WHM,
670 Smith DB. 2017. ICTV Virus Taxonomy Profile: Hepeviridae. Journal of General
671 Virology 98:2645-2646.
672 25. Wang B, Meng X-J. 2021. Hepatitis E virus: host tropism and zoonotic infection.
673 Current Opinion in Microbiology 59:8-15.
674 26. Graff J, Torian U, Nguyen H, Emerson SU. 2006. A Bicistronic Subgenomic
675 mRNA Encodes both the ORF2 and ORF3 Proteins of Hepatitis E Virus. Journal of
676 Virology 80:5919-5926.
677 27. Cortez V, Meliopoulos VA, Karlsson EA, Hargest V, Johnson C, Schultz-Cherry S.
678 2017. Astrovirus Biology and Pathogenesis. Annual Review of Virology 4:327-348.
679 28. Johnson C, Hargest V, Cortez V, Meliopoulos V, Schultz-Cherry S. 2017.
680 Astrovirus Pathogenesis. Viruses 9:22.
681 29. Finkbeiner SR, Kirkwood CD, Wang D. 2008. Complete genome sequence of a
682 highly divergent astrovirus isolated from a child with acute diarrhea. Virology
683 Journal 5:117.
684 30. Sato M, Kuroda M, Kasai M, Matsui H, Fukuyama T, Katano H, Tanaka-Taya K.
685 2016. Acute encephalopathy in an immunocompromised boy with astrovirus-
686 MLB1 infection detected by next generation sequencing. Journal of Clinical
687 Virology 78:66-70.
688 31. Cordey S, Vu D-L, Schibler M, L'Huillier AG, Brito F, Docquier M, Posfay-Barbe
689 KM, Petty TJ, Turin L, Zdobnov EM, Kaiser L. 2016. Astrovirus MLB2, a New
690 Gastroenteric Virus Associated with Meningitis and Disseminated Infection.
691 Emerging Infectious Diseases 22:846-853.

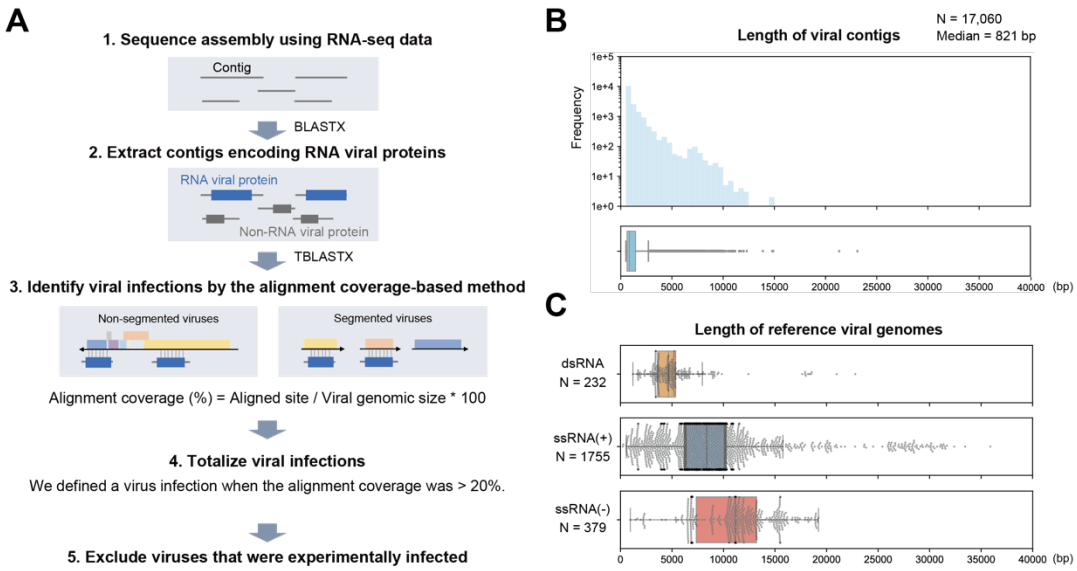
692 32. Karlsson EA, Small CT, Freiden P, Feeroz M, Matsen FA, San S, Hasan MK, Wang
693 D, Jones-Engel L, Schultz-Cherry S. 2015. Non-Human Primates Harbor Diverse
694 Mammalian and Avian Astroviruses Including Those Associated with Human
695 Infections. PLOS Pathogens 11:e1005225.
696 33. Britton PN, Jones CA, Macartney K, Cheng AC. 2018. Parechovirus: an important
697 emerging infection in young infants. Medical Journal of Australia 208:365-369.
698 34. Zell R, Delwart E, Gorbalenya AE, Hovi T, King AMQ, Knowles NJ, Lindberg
699 AM, Pallansch MA, Palmenberg AC, Reuter G, Simmonds P, Skern T, Stanway G,
700 Yamashita T. 2017. ICTV Virus Taxonomy Profile: Picornaviridae. Journal of
701 General Virology 98:2421-2422.
702 35. Simmonds P, Becher P, Bukh J, Gould EA, Meyers G, Monath T, Muerhoff S,
703 Pletnev A, Rico-Hesse R, Smith DB, Stapleton JT. 2017. ICTV Virus Taxonomy
704 Profile: Flaviviridae. Journal of General Virology 98:2-3.
705 36. Wu Z, Han Y, Liu B, Li H, Zhu G, Latinne A, Dong J, Sun L, Su H, Liu L, Du J,
706 Zhou S, Chen M, Kritiyakan A, Jittapalapong S, Chaisiri K, Buchy P, Duong V,
707 Yang J, Jiang J, Xu X, Zhou H, Yang F, Irwin DM, Morand S, Daszak P, Wang J,
708 Jin Q. 2021. Decoding the RNA viromes in rodent lungs provides new insight into
709 the origin and evolutionary patterns of rodent-borne pathogens in Mainland
710 Southeast Asia. Microbiome 9.
711 37. Antipov D, Raiko M, Lapidus A, Pevzner PA. 2020. MetaviralSPAdes: assembly
712 of viruses from metagenomic data. Bioinformatics 36:4126-4129.
713 38. Yahara K, Suzuki M, Hirabayashi A, Suda W, Hattori M, Suzuki Y, Okazaki Y.
714 2021. Long-read metagenomics using PromethION uncovers oral bacteriophages
715 and their interaction with host bacteria. Nature Communications 12.

716 39. Dreher TW. 1999. FUNCTIONS OF THE 3′-UNTRANSLATED REGIONS OF
717 POSITIVE STRAND RNA VIRAL GENOMES. Annual Review of
718 Phytopathology 37:151-174.
719 40. Munis AM, Bentley EM, Takeuchi Y. 2020. A tool with many applications:
720 vesicular stomatitis virus in research and medicine. Expert Opinion on Biological
721 Therapy 20:1187-1201.
722 41. Feehan BJ, Penin AA, Mukhin AN, Kumar D, Moskvina AS, Khametova KM,
723 Yuzhakov AG, Musienko MI, Zaberezhny AD, Aliper TI, Marthaler D, Alekseev
724 KP. 2019. Novel Mammalian orthorubulavirus 5 Discovered as Accidental Cell
725 Culture Contaminant. Viruses 11:777.
726 42. Wignall-Fleming E, Young DF, Goodbourn S, Davison AJ, Randall RE. 2016.
727 Genome Sequence of the Parainfluenza Virus 5 Strain That Persistently Infects
728 AGS Cells. Genome Announcements 4:e00653-16.
729 43. Ngoi CN, Siqueira J, Li L, Deng X, Mugo P, Graham SM, Price MA, Sanders EJ,
730 Delwart E. 2017. Corrigendum: The plasma virome of febrile adult Kenyans shows
731 frequent parvovirus B19 infections and a novel arbovirus (Kadipiro virus). Journal
732 of General Virology 98:517-517.
733 44. Gibb R, Albery GF, Becker DJ, Brierley L, Connor R, Dallas TA, Eskew EA,
734 Farrell MJ, Rasmussen AL, Ryan SJ, Sweeny A, Carlson CJ, Poisot T. 2021. Data
735 proliferation, reconciliation, and synthesis in viral ecology. bioRxiv
736 doi:10.1101/2021.01.14.426572:2021.01.14.426572.
737 45. Chen S, Zhou Y, Chen Y, Gu J. 2018. fastp: an ultra-fast all-in-one FASTQ
738 preprocessor. Bioinformatics 34:i884-i890.

739 46. Kim D, Langmead B, Salzberg SL. 2015. HISAT: a fast spliced aligner with low
740 memory requirements. Nature Methods 12:357-360.
741 47. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis
742 G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools.
743 Bioinformatics 25:2078-2079.
744 48. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin
745 VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N,
746 Tesler G, Alekseyev MA, Pevzner PA. 2012. SPAdes: A New Genome Assembly
747 Algorithm and Its Applications to Single-Cell Sequencing. Journal of
748 Computational Biology 19:455-477.
749 49. Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. 2017. metaSPAdes: a new
750 versatile metagenomic assembler. Genome Research 27:824-834.
751 50. Shen W, Le S, Li Y, Hu F. 2016. SeqKit: A Cross-Platform and Ultrafast Toolkit
752 for FASTA/Q File Manipulation. PLOS ONE 11:e0163962.
753 51. Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large
754 sets of protein or nucleotide sequences. Bioinformatics 22:1658-1659.
755 52. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden
756 TL. 2009. BLAST+: architecture and applications. BMC Bioinformatics 10:421.
757 53. Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. 2016. GenBank.
758 Nucleic Acids Research 44:D67-D72.
759 54. Barrett T, Clark K, Gevorgyan R, Gorelenkov V, Gribov E, Karsch-Mizrachi I,
760 Kimelman M, Pruitt KD, Resenchuk S, Tatusova T, Yaschenko E, Ostell J. 2012.
761 BioProject and BioSample databases at NCBI: facilitating capture and organization
762 of metadata. Nucleic Acids Research 40:D57-D63.

763 55. Katoh K, Standley DM. 2013. MAFFT Multiple Sequence Alignment Software
764 Version 7: Improvements in Performance and Usability. Molecular Biology and
765 Evolution 30:772-780.
766 56. Nguyen L-T, Schmidt HA, Von Haeseler A, Minh BQ. 2015. IQ-TREE: A Fast and
767 Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies.
768 Molecular Biology and Evolution 32:268-274.
769 57. Kalyaanamoorthy S, Minh BQ, Wong TKF, Von Haeseler A, Jermiin LS. 2017.
770 ModelFinder: fast model selection for accurate phylogenetic estimates. Nature
771 Methods 14:587-589.
772 58. Hoang DT, Chernomor O, Von Haeseler A, Minh BQ, Vinh LS. 2018. UFBoot2:
773 Improving the Ultrafast Bootstrap Approximation. Molecular Biology and
774 Evolution 35:518-522.
775 59. Yu G, Smith DK, Zhu H, Guan Y, Lam TTY. 2017. ggtree : an r package for
776 visualization and annotation of phylogenetic trees with their covariates and other
777 associated data. Methods in Ecology and Evolution 8:28-36.
778 60. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson
779 M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. Bioinformatics
780 29:15-21.
781 61. Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing
782 genomic features. Bioinformatics 26:841-842.
783
784

785 **Figure Legend**



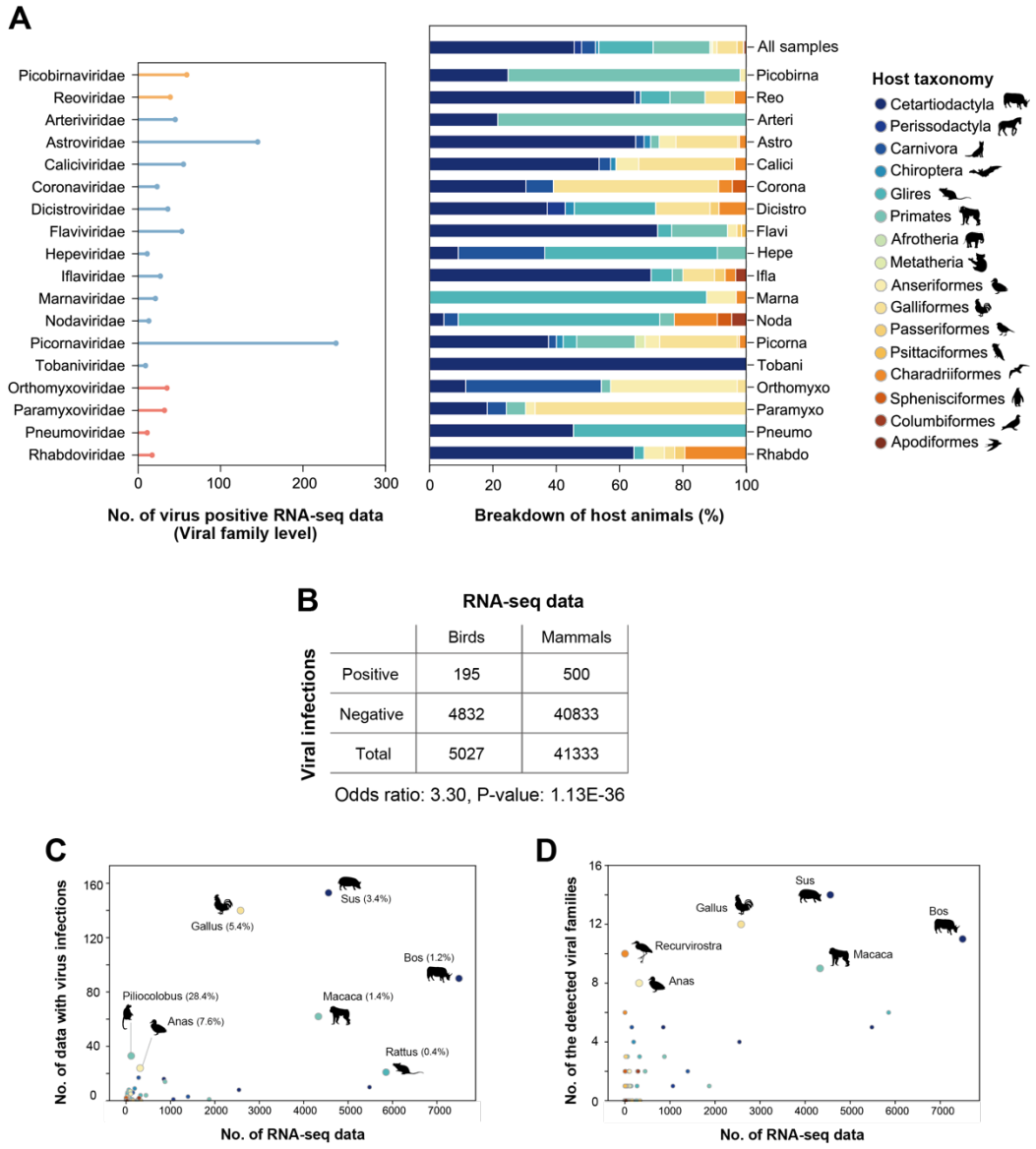
786

787 **Figure 1. Strategy for detecting viral infections in public RNA-seq data.**

788 (A) Schematic diagram of the procedure for detecting viral infections. First, we performed
789 *de novo* sequence assembly using publicly available mammalian and avian RNA-seq data.
790 Next, we extracted contigs encoding RNA viral proteins by BLASTX. Third, we
791 constructed sequence alignments by TBLASTX using the viral contigs in each RNA-seq
792 data and reference viral genomes because most viral contigs were shorter than complete
793 viral genomes, as shown in (B-C). The alignment coverage is defined as the proportion
794 of aligned sites in the entire reference viral genome. Fourth, we determined a viral
795 infection when the alignment coverage was > 20 %. Finally, we totalized the infections
796 at the virus family level after excluding experimentally infected viruses (**details in**
797 **Materials and Methods**).

798 (B) Distributions of viral contig length: histogram (upper panel) and box plot (lower
799 panel). The x-axis indicates the viral contig length. Among 17,060 viral contigs, the
800 median length was 821 bp.

801 (C) Length of reference viral genomes. Each panel corresponds to the Baltimore
802 classification: the upper, middle, and lower panels show double-stranded RNA (dsRNA)
803 viruses, positive-sense single-stranded RNA (ssRNA(+)) viruses, and negative-sense
804 single-stranded RNA (ssRNA(-)) viruses, respectively. The x-axis indicates the viral
805 genome size. These viral genomes were obtained from the RefSeq genomic viral database.
806 The genomic size of segmented viruses is the sum length of all segments in a virus species.
807



808

809 **Figure 2. RNA viral infections in the public sequencing data.**

810 (A) RNA viral infections detected in public sequencing data. Left panel: the x-axis
811 indicates the number of virus-positive RNA-seq data, and the y-axis indicates viral
812 families. Although infections by 22 RNA viral families were identified in this study, 18
813 families that were detected in more than five RNA-seq data are shown here. Bar colors
814 correspond to the Baltimore classification, dsRNA viruses (orange), ssRNA(+) viruses
815 (blue), and ssRNA(-) viruses (red). Right panel: breakdown by host animals in which

816 viral family infections were detected. The filled colors correspond to the host taxonomy
817 shown in the legend. The top row indicates the animal-wide breakdown of all RNA-seq
818 data used in this study.
819 (B) Comparison of viral detection rate between avian and mammalian samples. The table
820 shows the number of RNA-seq data with and without viral infections. The odds ratio and
821 p-value were obtained by Fisher's exact test.
822 (C) Scatter plot between the numbers of RNA-seq data investigated in this study (x-axis)
823 and those with viral infections (y-axis). Each dot indicates the animal genus. Dot colors
824 correspond to the host taxonomy shown in (A). The animal genera, in which viral
825 infections were detected in ≥ 20 samples, are annotated with the representative animal
826 species silhouettes. The percentages in parentheses indicate the ratio of virus-positive
827 RNA-seq data to the investigated data.
828 (D) Scatter plot between the number of RNA-seq data investigated in this study (x-axis)
829 and those of detected viral families (y-axis). Each dot indicates the animal genus. Dot
830 colors correspond to the host taxonomy shown in (A). The animal genera, in which ≥
831 eight viral families were detected, are annotated with the representative animal species
832 silhouettes.
833

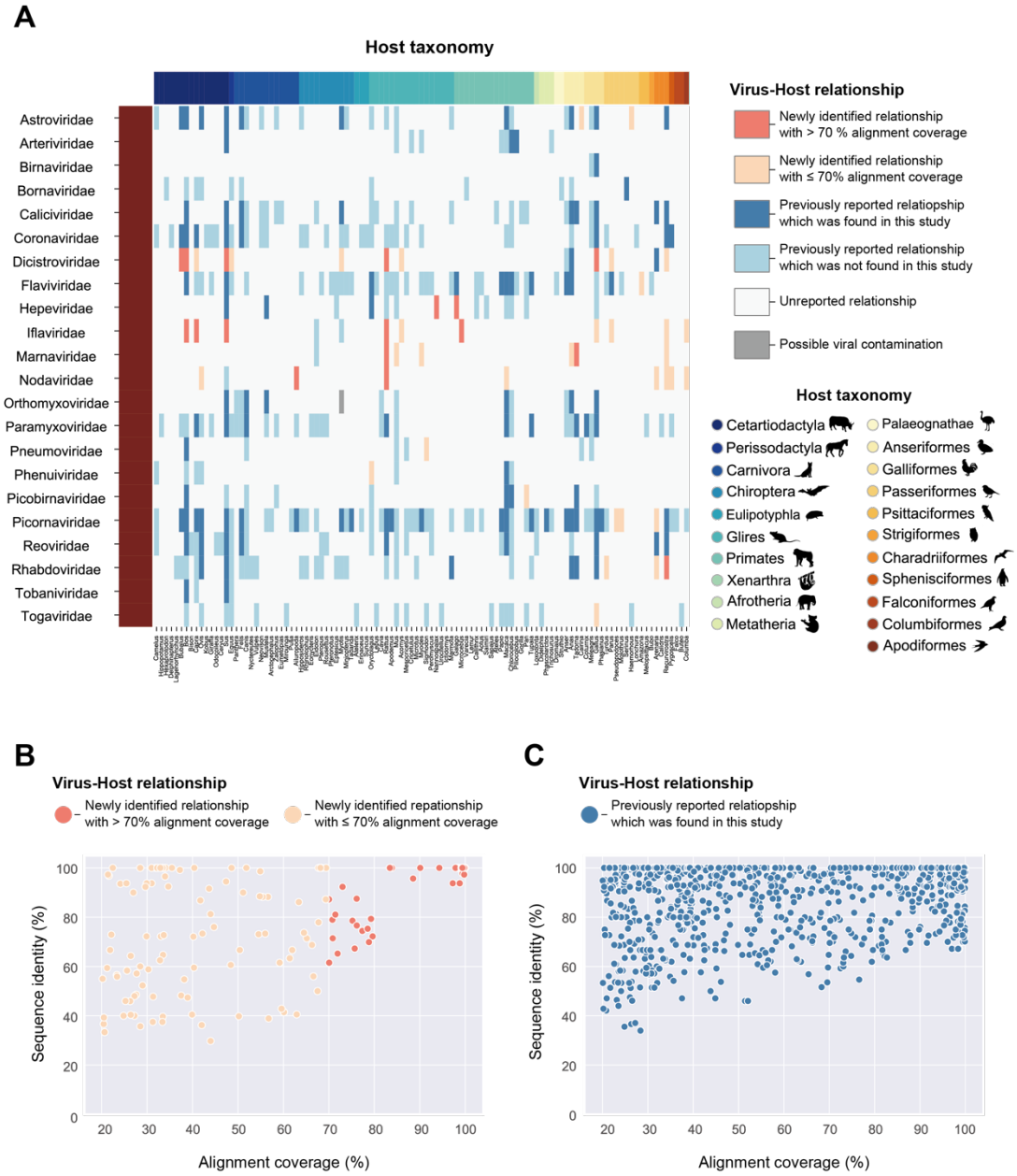


Figure 3. Search for unknown reservoir hosts and novel virus sequences.

(A) Heatmap showing the newness of virus-host relationships. Rows indicate viral families that reportedly infect vertebrate hosts. Columns indicate animal genus, and filled colors correspond to the host taxonomy shown in the lower right corner. Heatmap colors are according to six categories of virus-host relationships shown in the upper right corner: a relationship was newly identified in this study, and the viral infection was detected with

841 > 70 % alignment coverage (coral), a relationship was newly identified in this study, but
842 the viral infection was detected with ≤ 70 % alignment coverage (salmon), a relationship
843 was previously reported, and the viral infection was also detected in this study (blue), a
844 relationship was previously reported, but the viral infection was not detected in this study
845 (light blue), a relationship was unreported so far (white), and a relationship was newly
846 identified in this study, but it may be attributed to contamination (gray) (**see Discussion**).
847 (B-C) Scatter plot between alignment coverages (x-axis) and sequence similarities with
848 known viruses (y-axis). Each dot represents the viral infections identified in this study.
849 Viral infections related to novel virus-host relationships are shown in (B), and those
850 related to known relationships are shown in (C). The dot colors correspond to virus-host
851 relationships shown in (A). Sequence identity represents the maximum value of the
852 percentage of identical matches obtained by TBLASTX.
853

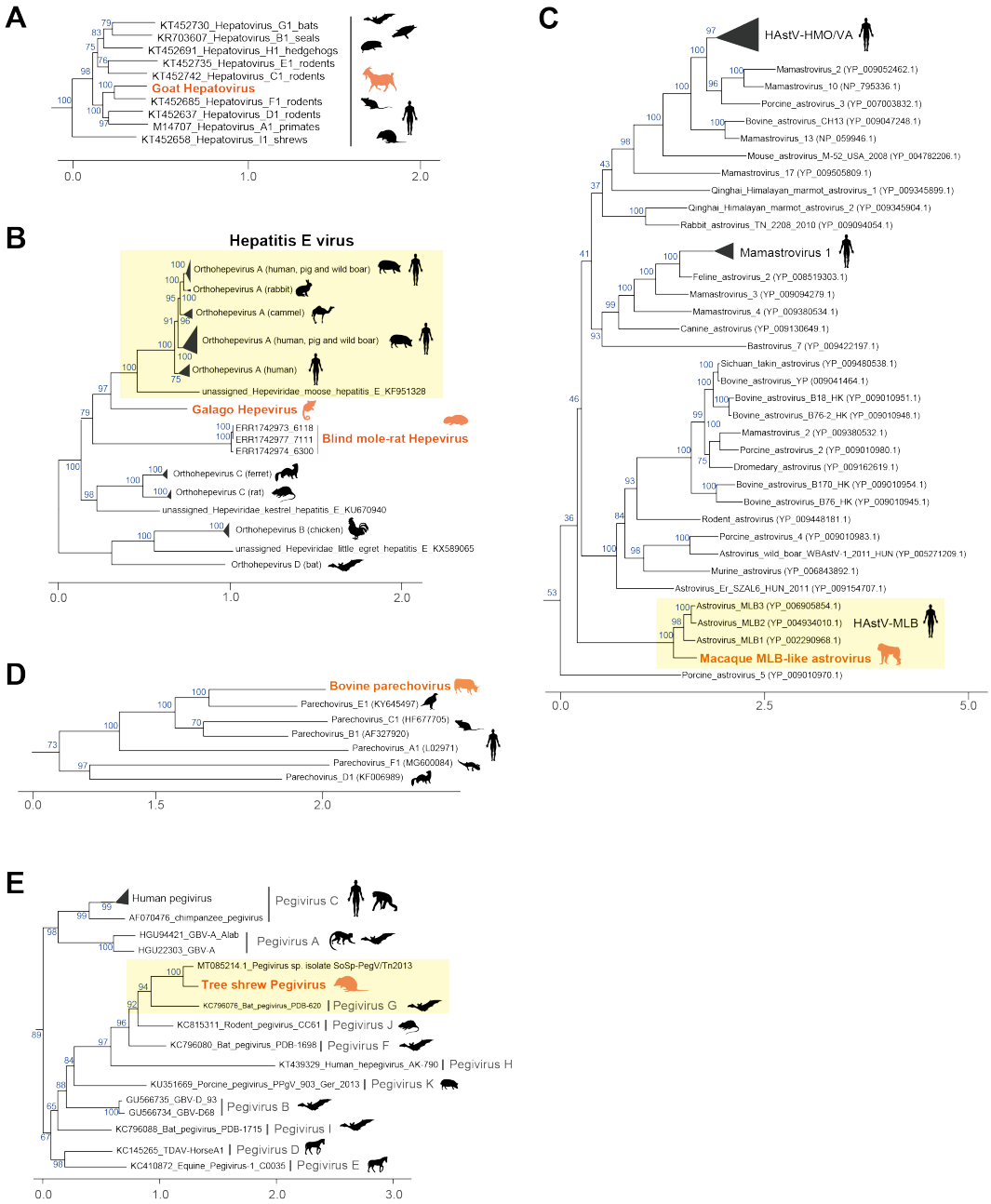


Figure 4. Characterization of virus sequences identified in this study.

(A-E) Phylogenetic analyses: the genus *Hepatovirus* of the family *Picornaviridae* (A), the family *Hepeviridae* (B), the genus *Mamastrovirus* of the family *Astroviridae* (C), the genus *Parechovirus* of the family *Picornaviridae* (D), and the genus *Pegivirus* of the family *Flaviviridae* (E). These phylogenetic trees were constructed based on the

860 maximum likelihood method (**details in Materials and Methods**). The orange labels
861 indicate viruses identified in this study, and the colored animal silhouette indicates the
862 viral host species. The black label and animal silhouette indicate known viruses and their
863 representative hosts, respectively. Scale bars indicate the genetic distance (substitutions
864 per site). The blue labels on branches indicate the bootstrap supporting values (%) with
865 1,000 replicates. Yellow boxes highlight viruses genetically similar to the novel virus
866 identified in this study.
867

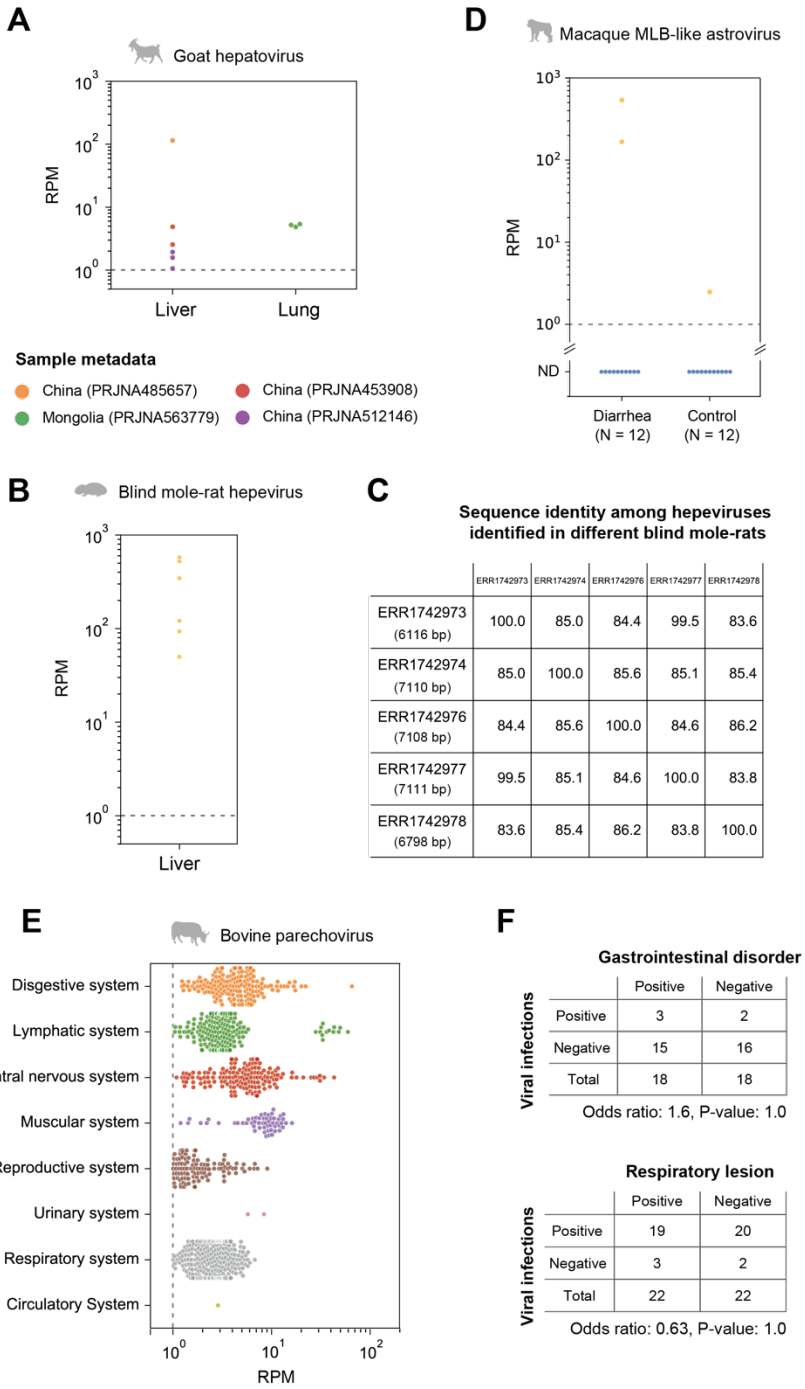


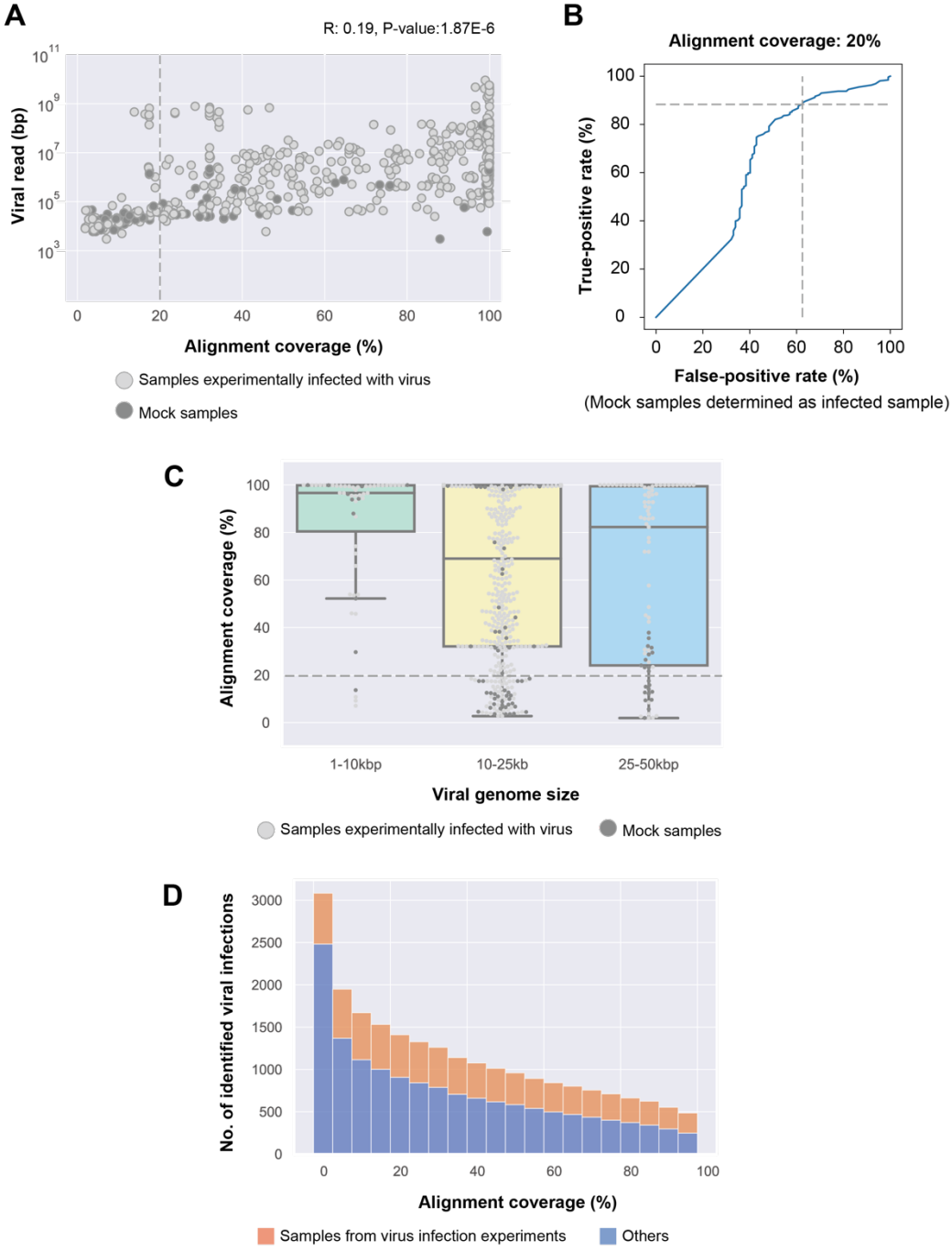
Figure 5. Detection of viral infections in the natural host population.

(A, B, and E) Investigation of viral infections in the natural host population by quantifying viral reads: goat hepatovirus (A), blind mole-rat hepevirus (B), and bovine parechovirus (E). Panel indicates the viral read amount (read per million reads [RPM]) in each tissue

873 or organ system. The gray dotted line indicates the criterion used to determine viral
874 infections (RPM: 1.0). The lower panel in (A) represents the sample metadata.
875 (C) Comparison of nucleotide sequence identity among the hepeviral sequences identified
876 in five different blind mole-rats. The numbers in parentheses in the row indicate the total
877 number of aligned sites between the viral contigs identified in each individual and the
878 blind mole-rat hepevirus identified in ERR1742977.
879 (D) Quantification of the macaque MLB-like viral infection levels in the patient with
880 diarrhea and control macaque monkeys. The x-axis indicates the diagnosis for the 24
881 monkeys, and the y-axis indicates the RPM. The average RPM for each individual is
882 plotted because six samples were collected from each individual. The dotted line indicates
883 the criterion used for detecting viral infections (RPM: 1.0). We considered samples with
884 RPMs below the criterion as non-detectable (ND).
885 (F) Association between the parechovirus infections and symptoms. The tables show the
886 number of RNA-seq data with and without the parechovirus infections in two independent
887 studies, which provide diagnostic information: gastrointestinal disorder (upper panel) and
888 respiratory lesion (lower panel). The odds ratio and p-value were obtained by Fisher's
889 exact test.
890

891 **Supplemental Materials**

Supplemental Figure 1

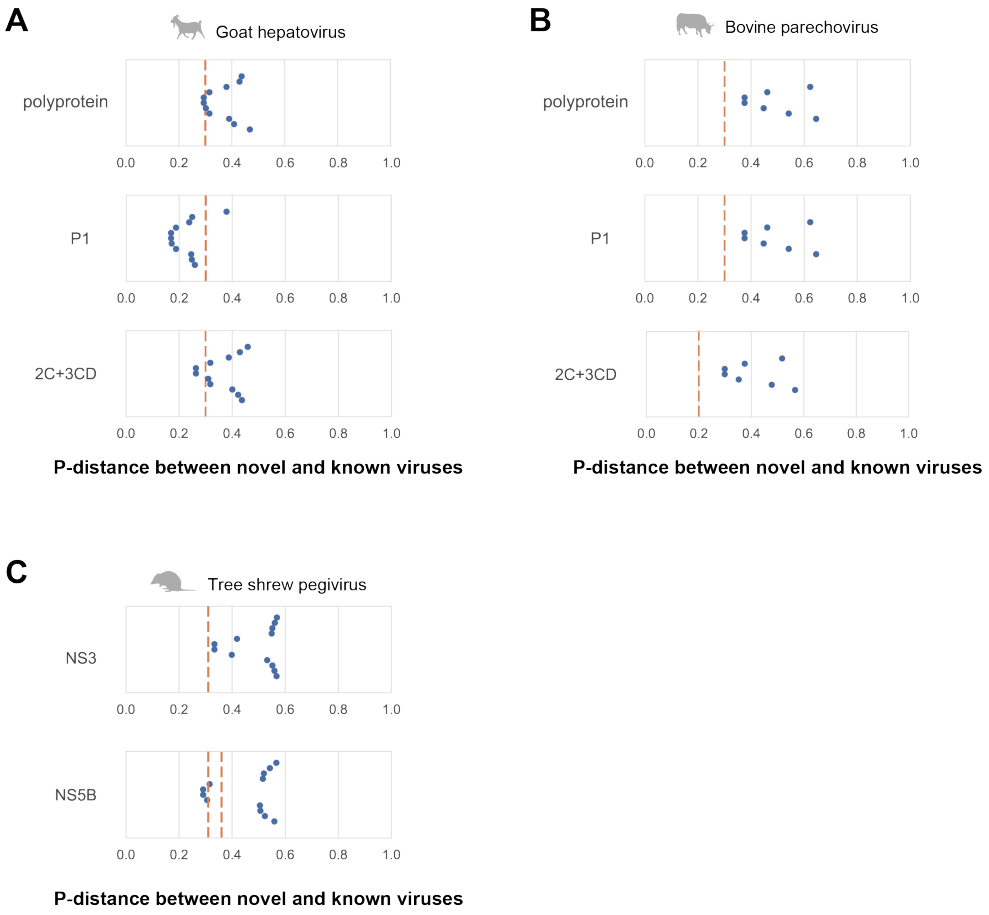


892

893 **Supplemental Figure 1. Validation of the alignment coverage-based method for**
894 **detecting viral infections.**
895 (A) Comparison between the alignment coverage-based method and the viral read-based
896 method using samples obtained from viral infection experiments. The x-axis indicates
897 alignment coverage between viral contigs in each RNA-seq data and the reference viral
898 genome used for the experiments. The y-axis indicates the total read length of the virus
899 family used for the experiment, which was obtained from the NCBI SRA Taxonomy
900 Analysis Tool. Light gray dots indicate samples experimentally infected with viruses, and
901 dark gray dots indicate mock samples. R: Pearson's correlation coefficient. Dotted line
902 indicates 20 % alignment coverage.
903 (B) Changes in the true-positive and the false-positive rates depending on the criteria to
904 determine viral infections. The true-positive rate (y-axis) indicates the number of samples
905 experimentally infected with viruses correctly determined as the infected sample, and the
906 false-positive rate (x-axis) indicates the number of mock samples determined as the
907 infected sample. Dotted line indicates the true-positive rate (88.3 %) and the false-positive
908 rate (62.5 %) when 20 % alignment coverage was used as the criterion (**details in**
909 **Materials and Methods**).
910 (C) Detection rate of viral infections depending on the viral genome size. Box plots show
911 the distributions of alignment coverage of the viral genome with 1-10kbp (green), 10-
912 25kbp (yellow), and 25-50kbp (blue). Light gray dots indicate samples infected with
913 viruses experimentally, and dark gray dots indicate mock samples. Dotted line indicates
914 20 % alignment coverage.
915 (D) The number of detected viral infections depending on the alignment coverage criteria.
916 The x-axis indicates alignment coverage used as a criterion for defining viral infections.

Bar graphs show the number of detected viral infections using the criterion shown on the x-axis. Filled colors indicate infections in samples from viral infection experiments (orange) or those in others (blue). When we used 20 % alignment coverage as the criterion, a total of 1,410 viral infections were identified, including 503 experimentally infected samples.

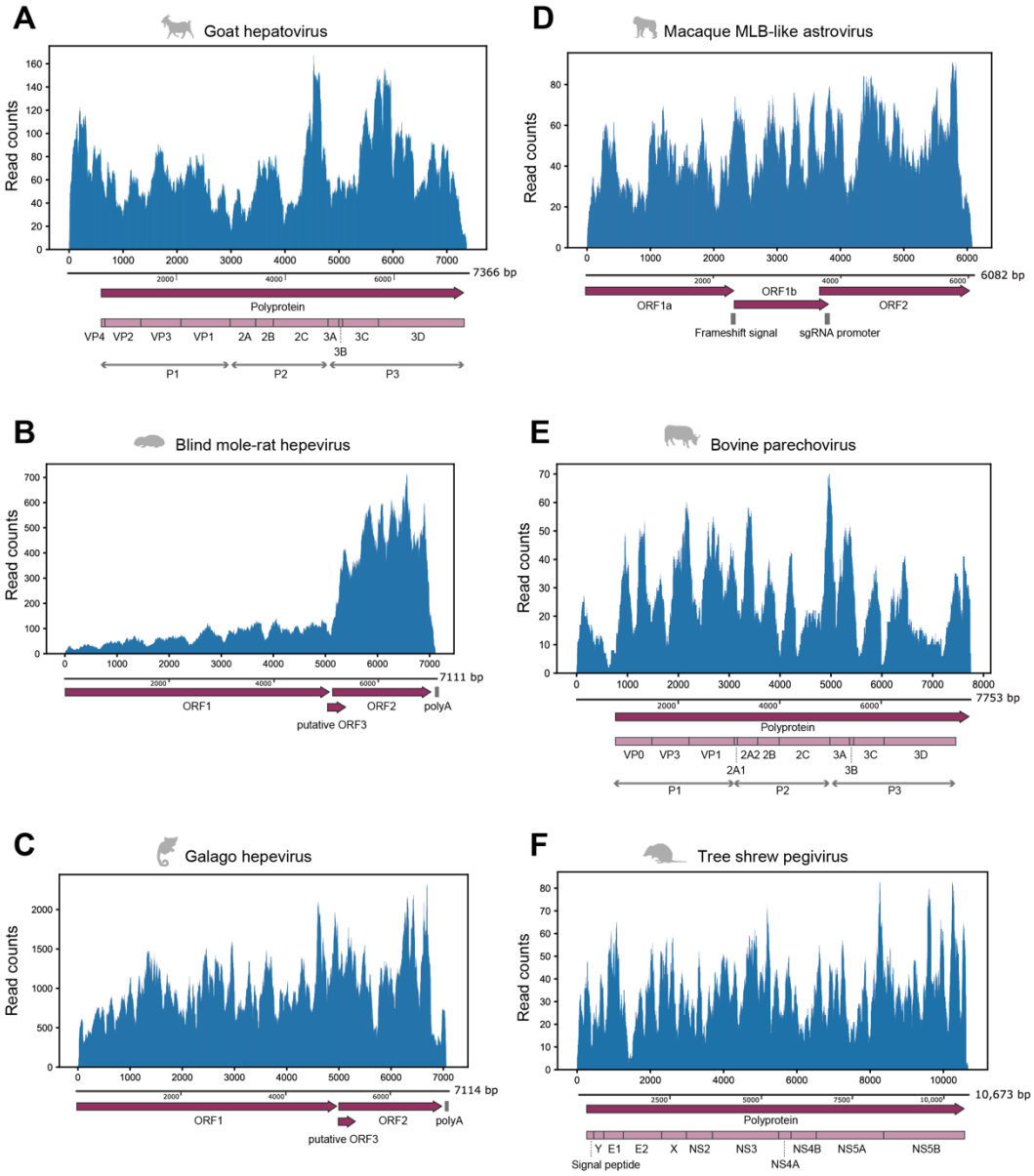
Supplemental Figure 2



Supplemental Figure 2. Comparison with the ICTV species demarcation criteria.

926 (A-C) Genetic distance among the amino acid sequences of novel and known viruses in
927 the genera *Hepatovirus* (A), *Parechovirus* (B), and *Pegivirus* (C). The x-axis indicates
928 the proportion of different sites: p-distance. Each dot shows the amino acid sequence p-
929 distance between the novel and known virus species. The International Committee on
930 Taxonomy of Viruses species demarcation criteria are shown as orange dotted lines:
931 greater than 0.3 in polyprotein, P1, and 2C+3CD regions for hepatoviruses (A), greater
932 than 0.3 in polyprotein, P1 regions and 0.2 in 2C+3CD region for parechoviruses (B), and
933 greater than 0.31 in the NS3 region and 0.31-0.36 in the NS5B region for pegiviruses (C).
934

Supplemental Figure 3



935

936 **Supplemental Figure 3. Mapping analysis using RNA-seq data in which the full-**
937 **length viral genome was identified.**
938 (A-F) Read distributions mapped to the viral sequence: goat hepatovirus (A), blind mole-
939 rat hepevirus (B), galago hepevirus (C), macaque MLB-like astrovirus (D), bovine
940 parechovirus (E), and tree shrew pegivirus (F). The upper panel shows the virus genomic
941 positions (x-axis) and read counts at each position (y-axis). The lower panel shows
942 genomic annotations, such as protein-coding regions or signal sequences. Dark purple
943 arrows indicate open reading frames (ORFs) in the viral genome. Light purple boxes show
944 mature proteins predicted based on aligned positions with reference viruses (**details in**
945 **Materials and Methods**). Gray vertical lines indicate nucleotide sequence features, such
946 as polyadenylation signal (poly-A), ribosomal frameshift signal (frameshift signal), and
947 promoter sequence for subgenomic RNA synthesis (sgRNA promoter).
948
949
950 **Supplemental Dataset 1. List of Sequence Read Archive run accession numbers,**
951 **genome file, and sequence assembly method.**
952 **Supplemental Dataset 2. Information on RNA-seq data from experimental infection**
953 **with viruses.**
954 **Supplemental Dataset 3. Information on possible viral contamination excluded from**
955 **the totalization.**
956 **Supplemental Dataset 4. Information on manual curation for virus-host**
957 **relationships.**

958 **Supplemental Dataset 5. Accession numbers of viral sequences used for phylogenetic**
959 **analyses, viral genomic annotations, and comparing the International Committee on**
960 **Taxonomy of Viruses species demarcation criteria.**
961 **Supplemental Dataset 6. Sequence Read Archive run accessions used for mapping**
962 **analyses.**
963 **Supplemental Dataset 7. Sample metadata in which the goat hepatoviral infections**
964 **were detected.**
965 **Supplemental Dataset 8. Sample metadata in which the blind mole-rat hepeviral**
966 **infections were detected.**
967 **Supplemental Dataset 9. Sample metadata in which the galago hepeviral infections**
968 **were detected.**
969 **Supplemental Dataset 10. Sample metadata in which the macaque MLB-like**
970 **astrovirus infections were detected.**
971 **Supplemental Dataset 11. Sample metadata in which the bovine parechovirus**
972 **infections were detected.**
973 **Supplemental Dataset 12. Sample metadata in which the tree shrew pegiviral**
974 **infections were detected.**
975 **Supplemental Dataset 13. Bioinformatics tools and their versions used in this study.**
976