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 accumulated in public databases in recent decades; however, only a small number of them have been examined for viral infections. Here, we screened for infections of 33 RNA viral families in publicly available mammalian and avian RNA-seq data and found over 900 hidden viral infections. We also discovered viral sequences 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, pegivirus in tree shrews, and seadornavirus in rats. 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.

46 **Importance**

47 Monitoring the spread of viral infections and identifying novel viruses capable of
48 infecting humans through animal reservoirs are necessary to control emerging viral
49 diseases. Massive sequencing data collected from various animals are publicly available,
50 but almost all these data have not been investigated regarding viral infections. Here, we
51 analyzed more than 46,000 public sequencing data and identified over 900 hidden RNA
52 viral infections in mammalian and avian samples. Some viruses discovered in this study
53 were genetically similar to pathogens that cause hepatitis, diarrhea, or encephalitis in
54 humans, suggesting the presence of new threats to public health. Our study demonstrates
55 the effectiveness of reusing public sequencing data to identify known and unknown viral
56 infections, indicating that future continuous monitoring of public sequencing data by
57 metagenomic analyses would help prepare and mitigate future viral pandemics.

58

Introduction

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

Metagenomic analysis can identify viruses with low bias and has substantially contributed to elucidating virus diversity for more than a decade (4). With the increase in publications using viral metagenomic analysis, new virus species, genera, and families have been successively established by the International Committee on Taxonomy of Viruses (ICTV) (5). However, a previous study estimated the existence of at least 40,000 mammalian viral species (6), which far exceeds the number of viral species classified by the ICTV to date (5, 7). Therefore, further research is needed to understand viral diversity and prepare for future viral pandemics. The quantity of RNA-seq data in public databases is growing exponentially (8); however, only limited dataset have been analyzed for viral infections (9, 10). The public sequencing data are derived from samples with various research backgrounds and may contain a wide variety of viruses. Therefore, analyzing publicly available RNA-seq data can be an effective way to assess the spread of viral infections and identify novel viruses.

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

83 We also discovered seven nearly complete viral genomes in livestock, wild, and
84 laboratory animals. Phylogenetic analyses showed some viruses were closely related to
85 human pathogenic viruses, suggesting the potential risk of causing disease in humans.
86 Furthermore, the viral infections were identified in several individuals collected by
87 independent studies, indicating that their infections may have already spread in the natural
88 host population. Our findings demonstrate the reusability of public sequencing data for
89 surveying viral infections that may present a threat to public health.
90

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 We used more than 46,000 mammalian and avian RNA-seq data to investigate
111 infections of 33 RNA virus families reported to infect vertebrates. Consequently, we
112 identified 907 infections of 22 RNA virus families in 709 sequencing data from 56 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 (11), such
118 as influenza A virus (12, 13) and West Nile virus (14). 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 3.28. Furthermore, among the investigated
121 species, we found relatively high viral detection rates in Gallus and Anas species at 20.1%
122 and 8.7%, 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 Database (15) (**Fig. 3A**).
130 This database lists virus-host relationships based on the identification of viral sequences
131 from a host animal. Using this database for comparison, we found 50 newly identified
132 virus-host relationships, and 17 of them were identified with more than 70% alignment
133 coverage. Notably, we identified nearly complete genomic sequences classified into the
134 family *Hepeviridae* in Spalax and Galago species for the first time. These discoveries
135 expanded our understanding of hepeviral host ranges (details of the viral characteristics
136 are 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 virus-
141 host relationships were also found in the families *Dicistroviridae*, *Iflaviridae*,
142 *Marnaviridae*, and *Nodaviridae*, suggesting that these viral host ranges are much broader
143 than previously expected. It should be noted that these relationships may be due to
144 contamination from environmental viruses, because few species in these virus families
145 have been reported to infect mammals or birds (16-19) (**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 genomic sequences (**Figs. 3B-C**). We found some viral sequences showing
151 low 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. 4-6 and S2-**
155 **3**). Consequently, we discovered seven viruses: hepatovirus in a goat, hepeviruses in blind
156 mole-rats and a galago, astrovirus in macaque monkeys, parechovirus in a cow, pegivirus
157 in tree shrews, and seadornavirus in rats.
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 (20). 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 goat samples (**Fig. 4A**). To our knowledge, this is the
166 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 goat samples, we found the viral
170 infection in nine samples from four independent studies with > 1.0 read per million reads
171 (RPM) (**Fig. 5A and Dataset S8**). These hepatoviral infections were detected in goat liver
172 and lung samples, suggesting that the goat hepatovirus can infect tissues other than the
173 liver. Although the lungs are not considered preferential tissues for hepatoviral replication,
174 a previous report also detected seal hepatoviral RNAs in the lungs (21). 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 (22) and members of Orthohepevirus A that infect
188 humans, pigs, rabbits, and camels (23) (**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 host livers, which coincided with
192 the tissue tropism of HEV (24). Additionally, we found that the 3'-portion of the blind
193 mole-rat hepevirus genome was highly transcribed (**Fig. S3B**), suggesting the
194 transcription of subgenomic RNAs (25). 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 viruses in a natural population using
199 RNA-seq data from blind mole-rats and galagos. Among 91 RNA-seq data from blind
200 mole-rats, we detected the hepeviral infections in six samples (**Fig. 5B**). The infected
201 individuals were from the same experiment, which were captured and kept as laboratory
202 animals in Israel (**Dataset S9**). There were two possibilities about when the hepeviruses
203 have infected blind mole-rats: the hepeviruses had already infected these blind mole-rats
204 when they were captured, or the viral infections had spread during the maintenance of
205 these individuals in the laboratory. To explore these possibilities, we investigated the
206 inter-individual diversity of the hepevirus sequences. We found that these individuals
207 were infected with relatively diverse hepeviruses representing nucleotide sequence
208 identities ranging from 83.6% to 99.5% (**Fig. 5C**). These results suggest that several
209 individuals had already been infected with distinct hepeviruses in the wild before being
210 captured. The galago hepeviral infections were detected in only two samples originating

211 from a study in which we first identified the virus (**Dataset S10**). This may be simply
212 because only four galago RNA-seq data obtained from the same study were available.
213 Taken together, we suggest that these hepeviruses can become a new threat to public
214 health, similar to HEV.
215
216 **MLB-like astrovirus detected in macaque monkeys with chronic diarrhea**
217 We found an astrovirus genetically similar to human astrovirus MLB (HAstV-MLB) in
218 fecal samples of macaque monkeys (**Fig. 4C**). Although HAstV-MLB infections are
219 typically asymptomatic (26, 27), several studies have reported the viral detection in cases
220 with diarrhea (28), encephalitis (29), or meningitis (30). Interestingly, the macaque MLB-
221 like astrovirus was found in macaque monkeys with chronic diarrhea. We analyzed the
222 viral read amounts in the patient (n = 12) and control (n = 12) monkeys to assess the
223 association between MLB-like astroviral infections and symptom prevalence (**Fig. 5D**
224 **and Dataset S11**). We detected abundant MLB-like astroviral reads in two patients,
225 suggesting that the viral infections are associated with host symptoms. However, we did
226 not observe the viral infection in other patients; further, we found the infection in a control
227 individual, although the viral read amount was approximately 100 times less than those
228 of the patients. Additionally, a previous study reported that monkeys, in which partial
229 sequences of MLB-like astroviruses were detected, had no obvious clinical signs,
230 including diarrhea (31). Thus, further experiments are needed to clarify the pathogenesis
231 of MLB-like astrovirus. Considering that there is no current experimental system for
232 examining HAstV-MLB infections (27), our findings suggest that macaque monkeys can
233 be used as animal model systems for researching MLB-like astroviruses.
234

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

236 Human parechovirus infection is especially problematic in infants and young children.
237 Although most parechovirus infections are considered asymptomatic, their infections
238 have been reported in patients with respiratory, digestive, and central nervous system
239 disorders (32). In this study, we identified a parechovirus, classified into the family
240 *Picornaviridae*, in the lower digestive tract of a cow (**Fig. 4D**). Despite the broad host
241 range of parechovirus, including mammals, birds, and reptiles (33), to our knowledge,
242 this is the first report on parechovirus infections in livestock animals.

243 Phylogenetic analysis indicated that this parechovirus was closely related to the
244 falcon parechovirus, a member of Parechovirus E. Next, we compared the bovine
245 parechovirus with the ICTV species demarcation criteria (33) to investigate whether this
246 virus is a novel species (**Fig. S2B**). Consequently, we found that the bovine parechovirus
247 was distant enough from other known parechovirus species and could be considered a
248 separate species based on the following criteria: divergence of amino acid sequences in
249 polyprotein (37.8%), P1 protein (37.8%), and 2C+3CD (29.9%) protein. Therefore, we
250 propose that this virus belongs to a new species in the genus *Parechovirus*.

251 We also investigated the prevalence of this parechovirus infection in a natural
252 host population using public RNA-seq data (**Fig. 5E and Dataset S12**). Among 8,284
253 cow samples, we detected the parechovirus infections in 944 samples from eight
254 independent studies with > 1.0 RPM. The viral infections were detected in various tissues,
255 such as the digestive, lymphatic, and central nervous system. These results suggest a
256 broad tissue tropism of the bovine parechovirus. To assess the parechovirus pathogenicity,
257 we analyzed the viral prevalence among 36 or 44 samples with a diagnosis for a
258 gastrointestinal disorder or respiratory lesion, respectively. We did not observe a

259 significant association between the viral infections and the presence/absence of
260 symptoms in these two studies (**Fig. 5F**). These results indicate that bovine parechovirus
261 infections may be asymptomatic, similar to the typical outcome of human parechoviral
262 infections. Furthermore, this also suggests that infected cows can spread parechoviral
263 infections as silent reservoirs.
264

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

267 We found a pegivirus belonging to the genus *Pegivirus* of the family *Flaviviridae* in tree
268 shrew liver samples. Phylogenetic analysis indicated that this pegivirus was closely
269 related to Pegivirus G identified in various bat species (**Fig. 4E**). According to the ICTV
270 species demarcation criteria (34), this virus appeared to be the same species as Pegivirus
271 G because the amino acid sequence identity in the NS5B gene was 70.9% (**Fig. S2C**).
272 These results indicate that Pegivirus G can infect distinct host lineages: tree shrews and
273 bats.

274 We also investigated the pegiviral infections in other tree shrew samples by read
275 mapping analysis. Among the 59 samples, the pegiviral infections were detected in four
276 samples collected from a research colony in the United Kingdom (**Dataset S13**). A recent
277 report partially identified a pegiviral sequence (MT085214) in tree shrews collected in
278 Southeast Asia (35), which showed 84.9% nucleotide sequence identity to the pegivirus
279 identified in this study (**Fig. 4E**). These results indicate that tree shrew pegivirus
280 infections were found in both Asia and Europe, suggesting an expanding geographic
281 distribution of Pegivirus G along with host animal transportation as experimental

282 resources. Thus, the global trade of host animals may lead to spreading pegiviral
283 infections hidden in tree shrews.
284
285 **Kadipiro virus in rats: a possible arbovirus that infects mosquitoes and mammals**
286 We identified Kadipiro virus (KDV), a member of the genus *Seadornavirus* of the family
287 *Reoviridae*, in rat spinal cord samples. Mosquitoes have been considered the hosts of
288 KDV (36); however, a previous report identified several KDV segments in plasma
289 samples from febrile humans (37). Phylogenetic analysis using VP1 amino acid
290 sequences indicated that the KDVs identified in humans, rats, and mosquitoes formed a
291 single cluster (**Fig. 4F**). Additionally, Banna virus, classified into the same genus as KDV,
292 is an arbovirus that transmits between mosquitoes and mammals, including humans, cows,
293 and pigs (38). Taken together with previous reports on seadornaviruses, KDV is also
294 expected to be an arbovirus.
295 Next, we calculated the sequence similarity among all segments between rat
296 KDV and known seadornaviruses to characterize the entire rat KDV genome (**Fig. 6**). We
297 found that several segments of rat KDV, especially segments 4-8, 10, and 11, showed
298 relatively low nucleotide sequence identities to those of mosquito KDV (**Fig. 6A**), even
299 though the amino acid sequences of rat KDV showed approximately 80% identity to
300 mosquito KDV throughout (**Fig. 6B**). These results suggest that rat KDV segments were
301 diversified among KDVs at the nucleotide sequence level due to virus-host coevolution
302 of codon usage and segment reassortment.
303 Various viral families, including coronaviruses and togaviruses, have been
304 reported to hijack the host macrodomain, leading to changes in virulence or immune
305 responses during viral infections (39). Interestingly, segment 8 in rat KDV may encode

306 chimeric VP8 containing a seadornaviral double-stranded RNA-binding domain (36) and
307 a macrodomain (**Fig. 6**). However, the mosquito KDV VP8 lacks a macrodomain. We
308 could not confirm whether human KDV encodes chimeric protein because human KDV
309 segment 8 was not identified in the previous study (37). Nonetheless, the presence of this
310 domain may be related to the determination of KDV host ranges. However, further
311 experiments are needed to confirm chimeric VP8 expression and function.
312
313

314 **Discussion**

315 Metagenomic analysis is a powerful approach for surveying viral infections (4, 5).
316 Although extensive deep sequencing data have accumulated in public databases, few data
317 have been investigated regarding viral infections. In this study, we analyzed the publicly
318 available RNA-seq data to search for hidden RNA viral infections in mammals and birds
319 and subsequently identified over 900 infections by 22 RNA virus families (**Figs. 1 and**
320 **2**). These results indicate that reusing public sequencing data is a cost-effective approach
321 for identifying viral infections. Furthermore, we discovered seven viruses in livestock,
322 wild, and experimental animals (**Fig. 4**). Some of these viruses were detected in different
323 individuals, suggesting that the viral infections may have already spread in the natural
324 host population (**Fig. 5**). Overall, our work demonstrates the reusability of public
325 sequencing data for surveying infections by both known and unknown viruses.

326 In this study, we determined viral infections by a combination of sequence
327 assembly and the alignment coverage-based method to solve several issues in viral
328 metagenomic analysis (**Fig. 1A**). One of the problems is detecting infections in data with
329 a small number of viral reads because almost all public sequencing data were collected
330 without using virus enrichment strategies. The result that most virus contigs were shorter
331 than the reference viral genomes reflects this difficulty (**Figs. 1B-C**). To resolve this issue,
332 we determined viral infections by the alignment coverage-based method, which uses
333 relatively short viral sequences as clues (**Figs. 1A and S1**). Consequently, we succeeded
334 in detecting over 900 RNA viral infections in public deep sequencing data (**Fig. 2A**).
335 Another problem in viral metagenomic analysis is that the viral detectability depends on
336 sequence similarity with known viruses. In this study, we discovered seven viral genomes
337 by sequence assembly (**Fig. 4**). Notably, these viral infections were undetectable in

338 almost all samples, even at the virus family and genus levels, by the NCBI SRA
339 Taxonomy Analysis Tool, which determines the taxonomic composition of reads in the
340 RNA-seq data without sequence assembly (**Dataset S8-S13**). These results indicate that
341 identifying viral sequences based on sequence assembly would effectively elucidate virus
342 diversity. Taken together, our strategy using sequence assembly and the alignment
343 coverage-based method can efficiently detect known and novel viral infections in publicly
344 available sequencing data.

345 However, there are still several challenges for identifying viral infections in
346 public sequencing data. First, we could not determine complete viral sequences mostly
347 (**Figs. 3B and 3C**). Further improvement in sequence assembly efficiency (40) or
348 integrative analysis using short- and long-read sequence datasets (41) can solve this
349 problem. Second, there may be a bias in virus detection using public sequencing data
350 depending on their genomic types. Among the 907 viral infections identified in this study,
351 75.2% were positive-sense single-stranded RNA (ssRNA(+)) viral infections, whereas
352 11.9% and 12.9% were double-stranded RNA and negative-sense single-stranded RNA
353 viral infections (**Fig. 2A**). The RNA-seq step, such as enrichment of polyadenylated
354 (poly-A) transcripts, can be relevant to this bias because many ssRNA(+) viruses have a
355 poly-A tract at the 3'-end of their genome (42). Alternatively, this bias may result from a
356 repertoire of reference viral genomes used for the viral search (**Fig. 1C**), which can be
357 solved in the future by database expansion.

358 Another challenge in viral metagenomic analysis using public data is
359 distinguishing true viral infections from contamination. To address this issue, we
360 performed integrative analyses using sample metadata and sequence information,
361 including sequence similarity and alignment coverage with known viruses (**details in**

362 **Materials and Methods**). Consequently, we found several possible contamination cases:
363 influenza A virus in Myotis bat, vesicular stomatitis Indiana virus (VSV) in chicken
364 cultured cells, and mammalian rubulavirus 5 (PIV5) in cultured cells and quail egg
365 samples (**Fig. 3A and Dataset S3**). For example, influenza A viral nucleotide sequence
366 identified in a bat sample showed 100% similarity to a laboratory strain of influenza A
367 virus (A/WSN/1933(H1N1)). Considering that the bat sample was collected in 2012, it is
368 difficult to expect that such a highly similar influenza A virus was maintained for
369 approximately 80 years. Likewise, the infections of VSVs and PIVs were also identified
370 with approximately 100% sequence similarity to the reference viral sequences (**Dataset**
371 **S3**). VSV is frequently used as an experimental tools; for example, as a pseudotype virus
372 (43). Additionally, previous studies have reported possible contamination of PIV5 in
373 cultured cells (44, 45). Therefore, we excluded these viral infections to avoid counting
374 false positives. These cases emphasize the importance of multilayered validations for
375 viral infections that were found only by viral metagenomic analysis.

376 Further research efforts to elucidate viral diversity are necessary to prepare for a
377 possible future viral pandemic (1, 5). A strategic approach, such as determining the host
378 samples used for virus search based on the expectation of viral infection frequency or
379 viral diversity, would be necessary. It has been discussed that birds may be high-risk viral
380 hosts of zoonoses because of their high species diversity and wide habitat range (11). In
381 this study, we found that viral infections were more frequently detected in birds,
382 especially Gallus and Anas species (**Figs. 2B-D**). Furthermore, among 223 viral
383 infections identified in Gallus and Anas samples, 78 infections (35.0%) showed less than
384 95% amino acid sequence similarity with known viruses, suggesting that these sequences

385 may be derived from unknown viruses. Therefore, further viral metagenomic analyses
386 targeting bird samples may effectively detect viral infections, including unknown ones.
387 In conclusion, we demonstrated the reusability of public sequencing data for
388 monitoring viral infections and discovering novel viral sequences, and elucidated diverse
389 RNA viruses hidden in animal samples. Our findings also emphasize the necessity of
390 continuous surveillance for viral infections using public sequencing data to prepare for
391 future viral pandemics, as well as the importance of developing a fundamental
392 bioinformatics platform for surveillance (46, 47).
393

394 **Materials and Methods**

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

396 RNA-seq data of 41,332 mammals (169 genera and 228 species) and 5,027 birds (70
397 genera and 83 species) were obtained from the NCBI Sequence Read Archive (SRA)
398 database (8) by pfastq-dump (https://github.com/inutano/pfastq-dump) and were then
399 preprocessed using fastp (version 0.20.0) (48) with options "-l 35", "-y -3", "-W 3", "-M
400 15", and "-x".

401 Sequence assembly was conducted by 1) mapping reads to the host or sister
402 species genome and 2) *de novo* assembly of sequences using unmapped reads. First, we
403 performed a mapping analysis to exclude the reads originating from host transcripts. We
404 mapped the reads in each RNA-seq data to the host genome by HISAT2 (version 2.1.0)
405 (49) with the default parameters or used the sister species genomes of the host in the same
406 genus when the host genome data were not available. Unmapped reads were extracted by
407 Samtools (version 1.9) (50) and Picard (version 2.20.4)
408 (http://broadinstitute.github.io/picard). When the relevant genome data were unavailable,
409 the preprocessed reads were directly used for sequence assembly. Sequence assembly was
410 conducted by SPAdes (version 3.13.0) (51) and/or metaSPAdes (version 3.13.0) (52) with
411 *k*-mers of 21, 33, 55, 77, and 99. Finally, we excluded contigs with lengths shorter than
412 500 bp by Seqkit (version 0.9.0) (53) and then clustered the contigs showing 95.0%
413 nucleotide sequence similarity by cd-hit-est (version 4.8.1) (54). Consequently, we
414 obtained 422,615,819 contigs and used them for subsequent analyses. We listed the SRA
415 Run accession numbers, genome files used for mapping analysis, and sequence assembly
416 tools in **Dataset S1**.

417

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

419 To determine the origins of the contigs, we analyzed the sequence similarity between the
420 contigs and known sequences in BLASTX screening (version 2.9.0) (55). First, we
421 performed BLASTX searches with the options "-word_size 2", "-evalue 1E-3", and
422 "max_target_seqs 1" using a custom database consisting of RNA viral proteins. We
423 constructed the custom database by downloading the viral protein sequences of the realm
424 *Riboviria* from the NCBI GenBank (version: 20190102) (56) and clustering the sequences
425 showing 98.0% similarity by cd-hit (version 4.8.1). Second, to confirm that the contigs
426 are not derived from organisms other than viruses, we further performed BLASTX
427 searches with the options "-word_size 2", "-evalue 1E-4", and "-max_target_seqs 10"
428 using the NCBI nr database (versions: 20190825-20190909 were used for screening
429 contigs in mammalian data and versions: 20190330-20190403 were used for screening
430 contigs in avian data). We determined the contig origins by comparing the bitscores in
431 the first and second BLASTX screening. Consequently, we obtained 17,060 contigs that
432 were deduced to encode RNA viral proteins.

433

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

435 Since most viral contigs were shorter than the reference viral genomic sizes (**Figs. 1B-C**),
436 we sought to determine viral infections based on the alignment coverage-based method
437 (**Fig. 1A**). First, we performed sequence alignment by TBLASTX (version 2.9.0) using
438 viral contigs from the same RNA-seq data and complete viral genomes in the NCBI
439 RefSeq genomic viral database (version 20200824). Next, we calculated the alignment
440 coverage with the genome of each viral species: the proportion of aligned sites in the
441 entire reference viral genome. In this study, we considered that an infection of the viral

442 family is present if the alignment coverage was greater than 20%. Validation of this
443 totalization method and evaluation of the criteria are described in the next section (**Fig.**
444 **S1**). Furthermore, we manually checked sequences with more than 70% alignment
445 coverage and more than 95% identity with known viruses in the TBLASTX alignment to
446 examine possible contamination with laboratory viral strains, as well as experimentally
447 inoculated viruses. We excluded experimental viral infections (**Dataset S2**) and possible
448 contamination (**Dataset S3**) from the final totalization (**Fig. 2A**). Overall, we investigated
449 the infections of 33 RNA viral families reported to infect vertebrates in 311 host species.
450

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

452 Using samples obtained from viral infection experiments, we first compared the
453 alignment coverage-based method with that based on viral read amounts in order to
454 validate the detection rate of viral infections of our method (**Fig. S1 and Dataset S2**). We
455 obtained the read amounts derived from experimentally infected viruses from the NCBI
456 SRA Taxonomy Analysis Tool results (https://github.com/ncbi/ngs-
457 tools/tree/tax/tools/tax). The calculation procedure for alignment coverage between viral
458 contigs in each RNA-seq data and viral reference genomes is described in the previous
459 section. We observed a positive correlation between the alignment coverage and viral
460 read amounts (Pearson's correlation coefficient: 0.19, p-value: 1.87E-6) (**Fig. S1A**).
461 Among the samples collected from experiments of viral infections, the true-positive rate
462 (the detection rate of experimentally inoculated viruses) was 88.3%, and the false-positive
463 rate (the rate that mock samples were determined to be infected samples) was 62.5% when
464 we used 20% alignment coverage as the criterion for determining viral infections (**Fig.**
465 **S1B**). The relatively high false-positive rate may be due to similar amounts of viral reads

466 in some mock samples as those in infected samples (**Fig. S1A**). Next, we analyzed the
467 association between alignment coverages and viral genome size (**Fig. S1C**) because the
468 detectability of viral infections in our method may depend on the reference viral genome
469 size. As expected, we observed a tendency for viruses with small genomes to be detected
470 with relatively high alignment coverage. However, more than 80% of experimentally
471 infected viral infections were detected with more than 20% alignment coverage,
472 regardless of the viral genome size. Based on these results, we established the alignment
473 coverage of 20 % to totalize the viral infections. Consequently, we identified a total of
474 1,410 RNA viral infections, including 503 infections in samples from viral infectious
475 experiments (**Fig. S1D**).
476
477 **Collection of information on experimentally infected viruses**
478 To exclude experimentally infected viruses from the final totalization, we analyzed the
479 experimental background of RNA-seq data. We first collected the experimental
480 descriptions of RNA-seq data: title and abstract from the NCBI BioProject database (57).
481 Then, we manually checked the terms relevant to viral infections in the descriptions,
482 focusing on viral name abbreviations and viral vector usage. We listed the obtained
483 information about viral infection experiments in **Dataset S2**.
484
485 **Summarization of virus-host relationships**
486 To identify novel reservoir hosts at the viral family levels, we compared the virus-host
487 relationships identified in this study with the dataset provided by the Virus-Host DB
488 (version: 20200629) (15). We define a "novel virus-host relationship" as one in which
489 the viral sequence has not been reported in the host. The virus-host relationships at the

490 viral family level were categorized as 1) a novel relationship detected with > 70%
491 alignment coverage, 2) a novel relationship detected with ≤ 70% alignment coverage, 3)
492 a known relationship that was also detected in this study, 4) a known relationship that
493 was not identified in this study, 5) a relationship unreported so far, and 6) a novel
494 relationship, which was possibly derived from contamination (**see Discussion**). To avoid
495 misclassification of the relationships, we analyzed reports manually by searching the
496 NCBI PubMed and Nucleotide databases using the combination of the host genus and
497 viral family names: for example, ["Pan" AND "Picobirnaviridae"]. The results of the
498 manual curation are listed in **Dataset S4**.
499
500 **Characterization of viral genomic architectures**
501 Open reading frames (ORFs) and polyadenylation signals in the viral genomes were
502 predicted by SnapGene software (snapgene.com). The positions of mature proteins,
503 frameshift signal sequences, and subgenomic RNA promoter sequences were predicted
504 based on sequence alignment using novel and reference viral sequences. The sequence
505 alignments were constructed by MAFFT (version 7.407) (58) with the option "--auto".
506 The reference viral sequences used for the genome annotations are listed in **Dataset S5**.
507 The macrodomain in rat KDV segment 8 was identified by CD-search (59) using the CDD
508 v3.18 database (60). The viral sequences identified in this study are registered under the
509 following accession numbers: BR001715-BR001732 and BR001751.
510
511 **Phylogenetic analyses**
512 Multiple sequence alignments (MSAs) of picornaviral P1 nucleotide sequences for **Fig.**
513 **4A**, hepeviral ORF1 amino acid sequences for **Fig. 4B**, picornaviral 3D nucleotide

514 sequences for **Fig. 4D**, and flaviviral NS5 nucleotide sequences for **Fig. 4E** were obtained
515 from the ICTV resources (the family of *Picornaviridae*: https://talk.ictvonline.org/ictv-
516 reports/ictv_online_report/positive-sense-rna-
517 viruses/picornavirales/w/picornaviridae/714/resources-picornaviridae, the family of
518 *Hepeviridae*: https://talk.ictvonline.org/ictv-reports/ictv_online_report/positive-sense-
519 rna-viruses/w/hepeviridae/731/resources-hepeviridae, and the family of *Flaviviridae*:
520 https://talk.ictvonline.org/ictv-reports/ictv_online_report/positive-sense-rna-
521 viruses/w/flaviviridae/371/resources-flaviviridae). For astroviruses (**Fig. 4C**) and
522 seadornaviruses (**Fig. 4F**), we collected reference sequences from the RefSeq protein
523 viral database (version 20210204) and extracted their amino acid sequences as follows:
524 ORF2 protein for viruses classified in the family *Astroviridae* and VP1 protein for viruses
525 classified in the genera *Seadornaviruses* and *Cardoreoviruses*. The MSAs of reference
526 and novel viral sequences were constructed by MAFFT with options "--add" and "--
527 keeplength". MSAs using astroviruses and seadornaviruses were trimmed by excluding
528 sites where > 20% of the sequences were gaps and subsequently removing sequences with
529 less than 80% of the total alignment sites. Phylogenetic trees were constructed by the
530 Maximum likelihood method using IQTREE (version 1.6.12) (61). The substitution
531 models were selected based on the Bayesian information criterion provided by
532 ModelFinder (62): GTR+R8 for **Fig. 4A**, LG+F+R4 for **Fig. 4B**, LG+F+R5 for **Fig. 4C**,
533 TVM+R9 for **Fig.4D**, GTR+R7 for **Fig. 4E**, and Blosum62 for **Fig. 4F**. The branch
534 supportive values were measured as the ultrafast bootstrap by UFBoot2 (63) with 1,000
535 replicates. Tree visualization was performed by the ggtree package (version 2.2.1) (64).
536 Sequence accession numbers used for the phylogenetic analyses are listed in **Dataset S5**.
537

538 **Comparison with the ICTV species demarcation criteria**

539 To assess whether the viruses identified in this study could be assigned to a novel species,
540 we compared their genetic distance with known viruses according to the ICTV species
541 demarcation criteria (33, 34) (**Fig. S2**). Amino acid sequences of the P1 and 3CD genes
542 in hepatoviruses and parechoviruses were extracted by referring to Hepatovirus A
543 (M14707) and Parechovirus A (S45208), respectively. Amino acid sequences of the NS3
544 and NS5B genes in pegiviruses were extracted by referring to Pegivirus A (U22303). We
545 constructed MSAs using these reference and novel viral sequences by MAFFT with the
546 option "--auto". We did not analyze other viruses identified in this study because the
547 ICTV did not provide criteria based on the genetic distance. The sequence accession
548 numbers used for these analyses are listed in **Dataset S5**.

549

550 **Calculation of genetic distances among the entire sequence of seadornaviral**
551 **segments**

552 To characterize the entire sequence of rat KDV segments, we visualized the sequence
553 identities between rat KDV and other seadornaviruses (**Fig. 6**). We first concatenated the
554 nucleotide and amino acid sequences of all the segments, and then constructed MSAs by
555 MAFFT with the option "--auto". The sequence identities were calculated by the recan
556 package (version 0.1.2) (65). The sequence accession numbers used for concatenation of
557 seadornaviral segments are listed in **Dataset S6**.

558

559 **Mapping analyses using viral genomes identified in this study**

560 To verify the quality of sequence assembly, we mapped the reads in the RNA-seq data,
561 in which a novel viral sequence was identified, to the viral genomes by STAR (version

562 2.7.6a) (66) (**Fig. S3**). The genome indexes were generated with the option "--
563 genomeSAindexNbases" according to each viral genomic size, and mapping analysis was
564 conducted with the options "--chimSegmentMin 20". The number of mapped reads in
565 each position was counted by Bedtools genomecov (version 2.27.1) (67) with the options
566 "-d" and "-split".

567 To identify novel viral infections in other individuals, we analyzed the publicly
568 available RNA-seq data of the host animals by quantifying viral reads (**Figs. 5A, B, and**
569 **5E**). We investigated 1,593 goat, 91 blind mole-rat, four galago, 8,282 cow, and 59 tree
570 shrew data for infections of goat hepatovirus, blind mole hepevirus, galago hepevirus,
571 bovine parechovirus, and tree shrew pegivirus, respectively. Mapping analyses were
572 performed using STAR (version 2.7.6a) as described above. The number of total and
573 mapped reads was extracted by Samtools (version 1.5). We considered that there was a
574 viral infection in the sample if the RPM was > 1.0.

575 We compared the viral read amounts between the patient and control monkeys
576 to investigate the association between chronic diarrhea and MLB-like astrovirus infection
577 (**Fig. 5D**). Viral read amounts were quantified as described above. The average RPM for
578 each individual is plotted in **Fig. 5D** because six samples were collected from each
579 individual. **Dataset S7** shows the SRA Run accession number used to investigate novel
580 viral infections. **Datasets S8-S13** list sample metadata in which the novel viral infections
581 were detected.

582

583 **Comparison of hepeviral sequences identified in different blind mole-rats**

584 We compared nucleotide sequence identities among the hepeviral sequences found in five
585 different individuals to predict when these viruses infected the blind mole-rats. The

586 sequence comparison was performed by BLASTN (version 2.11.0) with default
587 parameters. Because most hepeviral sequences were detected as short contigs, sequence
588 identities were represented by the percentage of identical matches in the longest aligned
589 region between the hepeviral sequences (**Fig. 5C**). We also analyzed the total of aligned
590 length between contigs identified in each individual and the hepeviral genome identified
591 in ERR1742977 and confirmed that these contigs covered 86.0-99.9% of the blind mole-
592 rat hepevirus genome.
593
594 **Data Availability**
595 Bioinformatics tools and their versions are listed in **Dataset S14**.
596

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 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.

620 **References**

621 1. Carroll D, Daszak P, Wolfe ND, Gao GF, Morel CM, Morzaria S, Pablos-Méndez
622 A, Tomori O, Mazet JAK. 2018. The Global Virome Project. Science 359:872-874.
623 2. Karesh WB, Dobson A, Lloyd-Smith JO, Lubroth J, Dixon MA, Bennett M, Aldrich
624 S, Harrington T, Formenty P, Loh EH, Machalaba CC, Thomas MJ, Heymann DL.
625 2012. Ecology of zoonoses: natural and unnatural histories. The Lancet 380:1936-
626 1945.
627 3. Otte M, Nugent R, McLeod A. 2004. Transboundary animal diseases: Assessment
628 of socio-economic impacts and institutional responses. Rome, Italy: Food and
629 Agriculture Organization (FAO):119-126.
630 4. Zhang Y-Z, Chen Y-M, Wang W, Qin X-C, Holmes EC. 2019. Expanding the RNA
631 Virosphere by Unbiased Metagenomics. Annual Review of Virology 6:119-139.
632 5. Greninger AL. 2018. A decade of RNA virus metagenomics is (not) enough. Virus
633 Research 244:218-229.
634 6. Carlson CJ, Zipfel CM, Garnier R, Bansal S. 2019. Global estimates of mammalian
635 viral diversity accounting for host sharing. Nature Ecology & Evolution 3:1070-
636 1075.
637 7. Gorbalenya AE, Krupovic M, Mushegian A, Kropinski AM, Siddell SG, Varsani
638 A, Adams MJ, Davison AJ, Dutilh BE, Harrach B, Harrison RL, Junglen S, King
639 AMQ, Knowles NJ, Lefkowitz EJ, Nibert ML, Rubino L, Sabanadzovic S,
640 Sanfaçon H, Simmonds P, Walker PJ, Zerbini FM, Kuhn JH, International
641 Committee on Taxonomy of Viruses Executive C. 2020. The new scope of virus
642 taxonomy: partitioning the virosphere into 15 hierarchical ranks. Nature
643 Microbiology 5:668-674.

644 8. Leinonen R, Sugawara H, Shumway M. 2011. The Sequence Read Archive.
645 Nucleic Acids Research 39:D19-D21.
646 9. Iwamoto M, Shibata Y, Kawasaki J, Kojima S, Li Y-T, Iwami S, Muramatsu M,
647 Wu H-L, Wada K, Tomonaga K, Watashi K, Horie M. 2021. Identification of novel
648 avian and mammalian deltaviruses provides new insights into deltavirus evolution.
649 Virus Evolution 7.
650 10. Horie M, Akashi H, Kawata M, Tomonaga K. 2020. Identification of a reptile
651 lyssavirus in Anolis allogus provided novel insights into lyssavirus evolution. Virus
652 Genes doi:10.1007/s11262-020-01803-y.
653 11. Nabi G, Wang Y, Lü L, Jiang C, Ahmad S, Wu Y, Li D. 2021. Bats and birds as
654 viral reservoirs: A physiological and ecological perspective. Science of The Total
655 Environment 754:142372.
656 12. Olsen B, Munster VJ, Wallensten A, Waldenstrom J, Osterhaus ADME, Fouchier
657 RAM. 2006. Global Patterns of Influenza A Virus in Wild Birds. Science 312:384-
658 388.
659 13. Lycett SJ, Duchatel F, Digard P. 2019. A brief history of bird flu. Philosophical
660 Transactions of the Royal Society B: Biological Sciences 374:20180257.
661 14. Habarugira G, Suen WW, Hobson-Peters J, Hall RA, Bielefeldt-Ohmann H. 2020.
662 West Nile Virus: An Update on Pathobiology, Epidemiology, Diagnostics, Control
663 and "One Health" Implications. Pathogens 9:589.
664 15. Mihara T, Nishimura Y, Shimizu Y, Nishiyama H, Yoshikawa G, Uehara H,
665 Hingamp P, Goto S, Ogata H. 2016. Linking Virus Genomes with Host Taxonomy.
666 Viruses 8:66.

667 16. Scherer WF, Verna JE, Richter GW. 1968. Nodamura Virus, an Ether- and
668 Chloroform-Resistant Arbovirus from Japan *. The American Journal of Tropical
669 Medicine and Hygiene 17:120-128.
670 17. Reuter G, Pankovics P, Gyöngyi Z, Delwart E, Boros Á. 2014. Novel dicistrovirus
671 from bat guano. Archives of Virology 159:3453-3456.
672 18. Greninger AL, Jerome KR. 2016. Draft Genome Sequence of Goose Dicistrovirus.
673 Genome Announcements 4:e00068-16.
674 19. Yinda CK, Zeller M, Conceição-Neto N, Maes P, Deboutte W, Beller L, Heylen E,
675 Ghogomu SM, Van Ranst M, Matthijnssens J. 2016. Novel highly divergent
676 reassortant bat rotaviruses in Cameroon, without evidence of zoonosis. Scientific
677 Reports 6:34209.
678 20. Lemon SM, Walker CM. 2019. Hepatitis A Virus and Hepatitis E Virus: Emerging
679 and Re-Emerging Enterically Transmitted Hepatitis Viruses. Cold Spring Harbor
680 Perspectives in Medicine 9:a031823.
681 21. Anthony SJ, St. Leger JA, Liang E, Hicks AL, Sanchez-Leon MD, Jain K,
682 Lefkowitch JH, Navarrete-Macias I, Knowles N, Goldstein T, Pugliares K, Ip HS,
683 Rowles T, Lipkin WI. 2015. Discovery of a Novel Hepatovirus (Phopivirus of
684 Seals) Related to Human Hepatitis A Virus. mBio 6:e01180-15.
685 22. Lin J, Norder H, Uhlhorn H, Belák S, Widén F. 2014. Novel hepatitis E like virus
686 found in Swedish moose. Journal of General Virology 95:557-570.
687 23. Purdy MA, Harrison TJ, Jameel S, Meng XJ, Okamoto H, Van Der Poel WHM,
688 Smith DB. 2017. ICTV Virus Taxonomy Profile: Hepeviridae. Journal of General
689 Virology 98:2645-2646.

690 24. Wang B, Meng X-J. 2021. Hepatitis E virus: host tropism and zoonotic infection.
691 Current Opinion in Microbiology 59:8-15.
692 25. Graff J, Torian U, Nguyen H, Emerson SU. 2006. A Bicistronic Subgenomic
693 mRNA Encodes both the ORF2 and ORF3 Proteins of Hepatitis E Virus. Journal of
694 Virology 80:5919-5926.
695 26. Cortez V, Meliopoulos VA, Karlsson EA, Hargest V, Johnson C, Schultz-Cherry S.
696 2017. Astrovirus Biology and Pathogenesis. Annual Review of Virology 4:327-348.
697 27. Johnson C, Hargest V, Cortez V, Meliopoulos V, Schultz-Cherry S. 2017.
698 Astrovirus Pathogenesis. Viruses 9:22.
699 28. Finkbeiner SR, Kirkwood CD, Wang D. 2008. Complete genome sequence of a
700 highly divergent astrovirus isolated from a child with acute diarrhea. Virology
701 Journal 5:117.
702 29. Sato M, Kuroda M, Kasai M, Matsui H, Fukuyama T, Katano H, Tanaka-Taya K.
703 2016. Acute encephalopathy in an immunocompromised boy with astrovirus-
704 MLB1 infection detected by next generation sequencing. Journal of Clinical
705 Virology 78:66-70.
706 30. Cordey S, Vu D-L, Schibler M, L'Huillier AG, Brito F, Docquier M, Posfay-Barbe
707 KM, Petty TJ, Turin L, Zdobnov EM, Kaiser L. 2016. Astrovirus MLB2, a New
708 Gastroenteric Virus Associated with Meningitis and Disseminated Infection.
709 Emerging Infectious Diseases 22:846-853.
710 31. Karlsson EA, Small CT, Freiden P, Feeroz M, Matsen FA, San S, Hasan MK, Wang
711 D, Jones-Engel L, Schultz-Cherry S. 2015. Non-Human Primates Harbor Diverse
712 Mammalian and Avian Astroviruses Including Those Associated with Human
713 Infections. PLOS Pathogens 11:e1005225.

714 32. Britton PN, Jones CA, Macartney K, Cheng AC. 2018. Parechovirus: an important
715 emerging infection in young infants. Medical Journal of Australia 208:365-369.
716 33. Zell R, Delwart E, Gorbalenya AE, Hovi T, King AMQ, Knowles NJ, Lindberg
717 AM, Pallansch MA, Palmenberg AC, Reuter G, Simmonds P, Skern T, Stanway G,
718 Yamashita T. 2017. ICTV Virus Taxonomy Profile: Picornaviridae. Journal of
719 General Virology 98:2421-2422.
720 34. Simmonds P, Becher P, Bukh J, Gould EA, Meyers G, Monath T, Muerhoff S,
721 Pletnev A, Rico-Hesse R, Smith DB, Stapleton JT. 2017. ICTV Virus Taxonomy
722 Profile: Flaviviridae. Journal of General Virology 98:2-3.
723 35. Wu Z, Han Y, Liu B, Li H, Zhu G, Latinne A, Dong J, Sun L, Su H, Liu L, Du J,
724 Zhou S, Chen M, Kritiyakan A, Jittapalapong S, Chaisiri K, Buchy P, Duong V,
725 Yang J, Jiang J, Xu X, Zhou H, Yang F, Irwin DM, Morand S, Daszak P, Wang J,
726 Jin Q. 2021. Decoding the RNA viromes in rodent lungs provides new insight into
727 the origin and evolutionary patterns of rodent-borne pathogens in Mainland
728 Southeast Asia. Microbiome 9.
729 36. Attoui H, De Micco P, De Lamballerie X, Billoir F, Biagini P. 2000. Complete
730 sequence determination and genetic analysis of Banna virus and Kadipiro virus:
731 proposal for assignment to a new genus (Seadornavirus) within the family
732 Reoviridae. Journal of General Virology 81:1507-1515.
733 37. Ngoi CN, Siqueira J, Li L, Deng X, Mugo P, Graham SM, Price MA, Sanders EJ,
734 Delwart E. 2016. The plasma virome of febrile adult Kenyans shows frequent
735 parvovirus B19 infections and a novel arbovirus (Kadipiro virus). Journal of
736 General Virology 97:3359-3367.

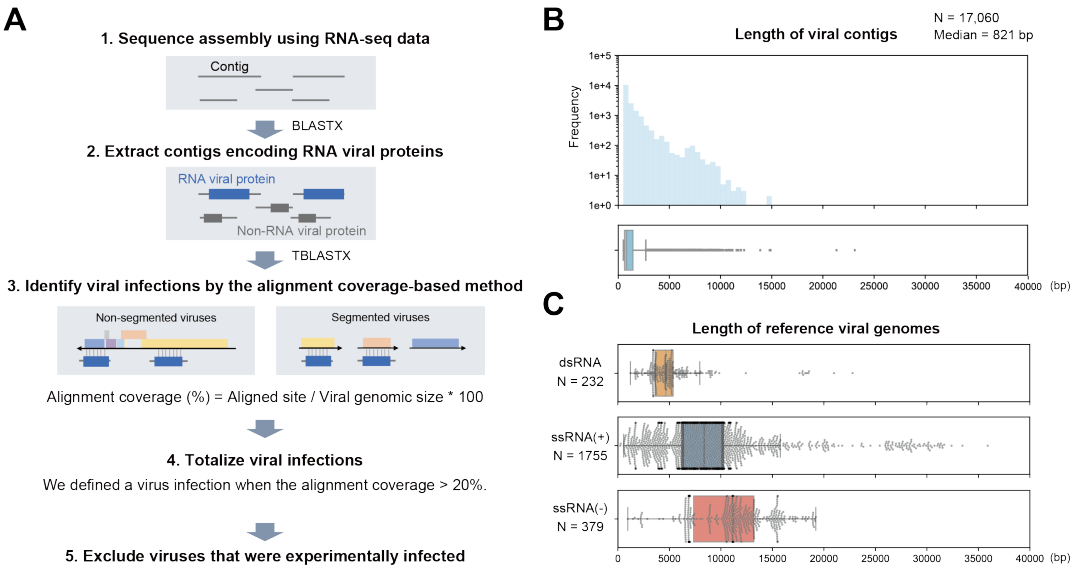
737 38. Liu H, Li M-H, Zhai Y-G, Meng W-S, Sun X-H, Cao Y-X, Fu S-H, Wang H-Y, Xu
738 L-H, Tang Q, Liang G-D. 2010. Banna Virus, China, 1987–2007. Emerging
739 Infectious Diseases 16:514-517.
740 39. Rack JGM, Perina D, Ahel I. 2016. Macrodomains: Structure, Function, Evolution,
741 and Catalytic Activities. Annual Review of Biochemistry 85:431-454.
742 40. Antipov D, Raiko M, Lapidus A, Pevzner PA. 2020. MetaviralSPAdes: assembly
743 of viruses from metagenomic data. Bioinformatics 36:4126-4129.
744 41. Yahara K, Suzuki M, Hirabayashi A, Suda W, Hattori M, Suzuki Y, Okazaki Y.
745 2021. Long-read metagenomics using PromethION uncovers oral bacteriophages
746 and their interaction with host bacteria. Nature Communications 12.
747 42. Dreher TW. 1999. FUNCTIONS OF THE 3′-UNTRANSLATED REGIONS OF
748 POSITIVE STRAND RNA VIRAL GENOMES. Annual Review of
749 Phytopathology 37:151-174.
750 43. Munis AM, Bentley EM, Takeuchi Y. 2020. A tool with many applications:
751 vesicular stomatitis virus in research and medicine. Expert Opinion on Biological
752 Therapy 20:1187-1201.
753 44. Feehan BJ, Penin AA, Mukhin AN, Kumar D, Moskvina AS, Khametova KM,
754 Yuzhakov AG, Musienko MI, Zaberezhny AD, Aliper TI, Marthaler D, Alekseev
755 KP. 2019. Novel Mammalian orthorubulavirus 5 Discovered as Accidental Cell
756 Culture Contaminant. Viruses 11:777.
757 45. Wignall-Fleming E, Young DF, Goodbourn S, Davison AJ, Randall RE. 2016.
758 Genome Sequence of the Parainfluenza Virus 5 Strain That Persistently Infects
759 AGS Cells. Genome Announcements 4:e00653-16.

760 46. Edgar RC, Taylor J, Altman T, Barbera P, Meleshko D, Lin V, Lohr D, Novakovsky
761 G, Al-Shayeb B, Banfield JF, Korobeynikov A, Chikhi R, Babaian A. 2020.
762 Petabase-scale sequence alignment catalyses viral discovery. bioRxiv
763 doi:10.1101/2020.08.07.241729:2020.08.07.241729.
764 47. Gibb R, Albery GF, Becker DJ, Brierley L, Connor R, Dallas TA, Eskew EA,
765 Farrell MJ, Rasmussen AL, Ryan SJ, Sweeny A, Carlson CJ, Poisot T. 2021. Data
766 proliferation, reconciliation, and synthesis in viral ecology. bioRxiv
767 doi:10.1101/2021.01.14.426572:2021.01.14.426572.
768 48. Chen S, Zhou Y, Chen Y, Gu J. 2018. fastp: an ultra-fast all-in-one FASTQ
769 preprocessor. Bioinformatics 34:i884-i890.
770 49. Kim D, Langmead B, Salzberg SL. 2015. HISAT: a fast spliced aligner with low
771 memory requirements. Nature Methods 12:357-360.
772 50. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis
773 G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools.
774 Bioinformatics 25:2078-2079.
775 51. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin
776 VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N,
777 Tesler G, Alekseyev MA, Pevzner PA. 2012. SPAdes: A New Genome Assembly
778 Algorithm and Its Applications to Single-Cell Sequencing. Journal of
779 Computational Biology 19:455-477.
780 52. Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. 2017. metaSPAdes: a new
781 versatile metagenomic assembler. Genome Research 27:824-834.
782 53. Shen W, Le S, Li Y, Hu F. 2016. SeqKit: A Cross-Platform and Ultrafast Toolkit
783 for FASTA/Q File Manipulation. PLOS ONE 11:e0163962.

784 54. Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large
785 sets of protein or nucleotide sequences. Bioinformatics 22:1658-1659.
786 55. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden
787 TL. 2009. BLAST+: architecture and applications. BMC Bioinformatics 10:421.
788 56. Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. 2016. GenBank.
789 Nucleic Acids Research 44:D67-D72.
790 57. Barrett T, Clark K, Gevorgyan R, Gorelenkov V, Gribov E, Karsch-Mizrachi I,
791 Kimelman M, Pruitt KD, Resenchuk S, Tatusova T, Yaschenko E, Ostell J. 2012.
792 BioProject and BioSample databases at NCBI: facilitating capture and organization
793 of metadata. Nucleic Acids Research 40:D57-D63.
794 58. Katoh K, Standley DM. 2013. MAFFT Multiple Sequence Alignment Software
795 Version 7: Improvements in Performance and Usability. Molecular Biology and
796 Evolution 30:772-780.
797 59. Marchler-Bauer A, Bryant SH. 2004. CD-Search: protein domain annotations on
798 the fly. Nucleic Acids Research 32:W327-W331.
799 60. Lu S, Wang J, Chitsaz F, Derbyshire MK, Geer RC, Gonzales NR, Gwadz M,
800 Hurwitz DI, Marchler GH, Song JS, Thanki N, Yamashita RA, Yang M, Zhang D,
801 Zheng C, Lanczycki CJ, Marchler-Bauer A. 2020. CDD/SPARCLE: the conserved
802 domain database in 2020. Nucleic Acids Research 48:D265-D268.
803 61. Nguyen L-T, Schmidt HA, Von Haeseler A, Minh BQ. 2015. IQ-TREE: A Fast and
804 Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies.
805 Molecular Biology and Evolution 32:268-274.

806 62. Kalyaanamoorthy S, Minh BQ, Wong TKF, Von Haeseler A, Jermiin LS. 2017.
807 ModelFinder: fast model selection for accurate phylogenetic estimates. Nature
808 Methods 14:587-589.
809 63. Hoang DT, Chernomor O, Von Haeseler A, Minh BQ, Vinh LS. 2018. UFBoot2:
810 Improving the Ultrafast Bootstrap Approximation. Molecular Biology and
811 Evolution 35:518-522.
812 64. Yu G, Smith DK, Zhu H, Guan Y, Lam TTY. 2017. ggtree : an r package for
813 visualization and annotation of phylogenetic trees with their covariates and other
814 associated data. Methods in Ecology and Evolution 8:28-36.
815 65. Babin Y. 2020. Recan: Python tool for analysis of recombination events in viral
816 genomes. Journal of Open Source Software 5:2014.
817 66. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson
818 M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. Bioinformatics
819 29:15-21.
820 67. Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing
821 genomic features. Bioinformatics 26:841-842.
822
823

824 **Figure Legend**

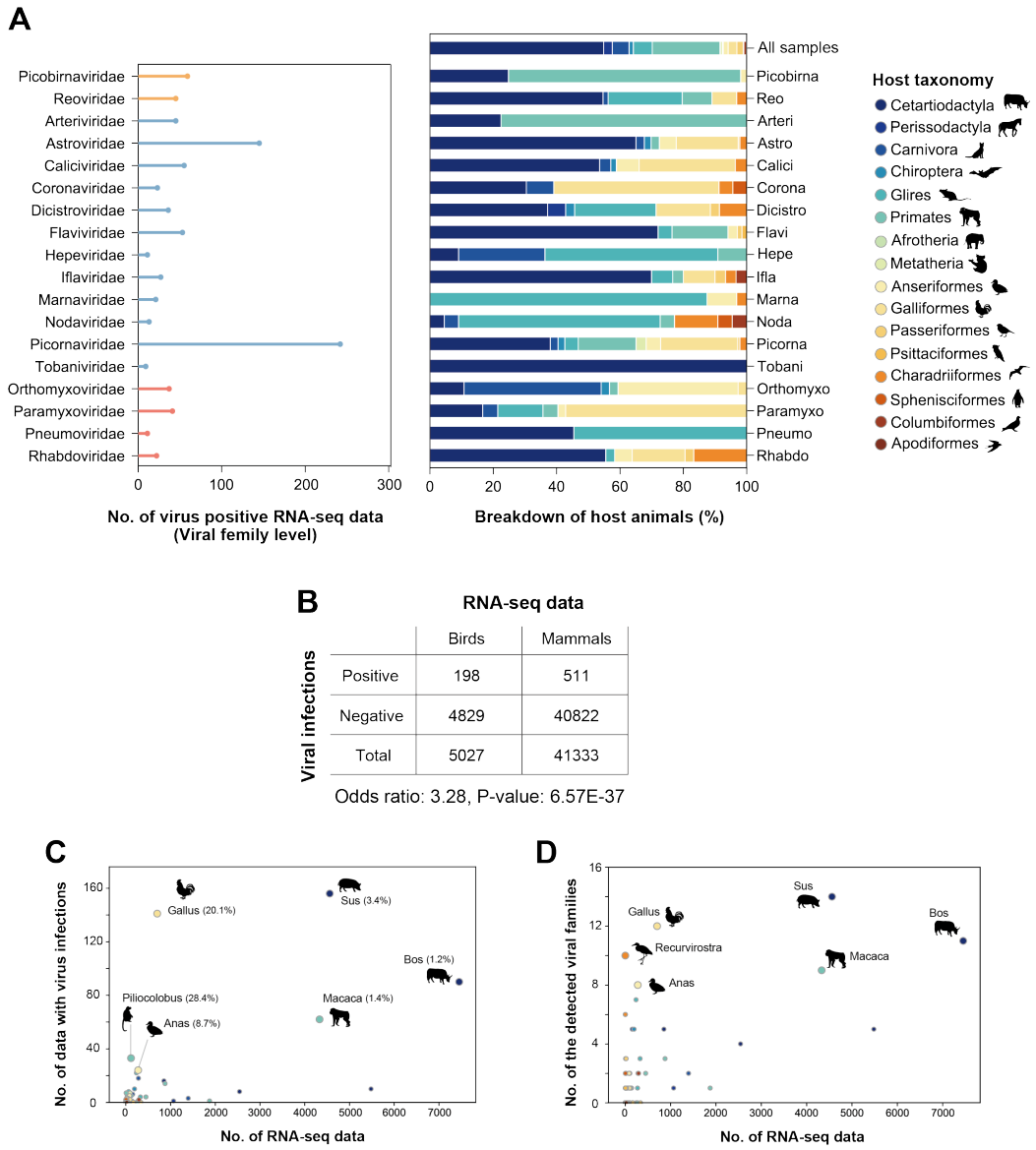


825

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

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

837 (B) Distributions of viral contig length: histogram (upper panel) and box plot (lower
838 panel). The x-axis indicates the viral contig length. Among 17,060 viral contigs, the
839 median length was 821 bp.
840 (C) Length of reference viral genomes. Each panel corresponds to the Baltimore
841 classification: the upper, middle, and lower panels show double-stranded RNA (dsRNA)
842 viruses, positive-sense single-stranded RNA (ssRNA(+)) viruses, and negative-sense
843 single-stranded RNA (ssRNA(-)) viruses, respectively. The x-axis indicates the viral
844 genome size. These viral genomes were obtained from the RefSeq genomic viral database.
845 The genomic size of segmented viruses is the sum length of all segments in a virus species.
846



847

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

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

viral family infections were detected. The filled colors correspond to the host taxonomy shown in the legend. The top row indicates the animal-wide breakdown of all RNA-seq data used in this study.

(B) Comparison of viral detection rate between avian and mammalian samples. The table shows the number of RNA-seq data with and without viral infections. The odds ratio and p-value were obtained by Fisher's exact test.

(C) Scatter plot between the numbers of RNA-seq data investigated in this study (x-axis) and those with viral infections (y-axis). Each dot indicates the animal genus. Dot colors correspond to the host taxonomy shown in (A). The animal genera, in which viral infections were detected in ≥ 24 samples, are annotated with the representative animal species silhouettes. The percentages in parentheses indicate the ratio of virus-positive RNA-seq data to the investigated data.

(D) Scatter plot between the number of RNA-seq data investigated in this study (x-axis) and those of detected viral families (y-axis). Each dot indicates the animal genus. Dot colors correspond to the host taxonomy shown in (A). The animal genera, in which ≥ eight viral families were detected, are annotated with the representative animal species silhouettes.

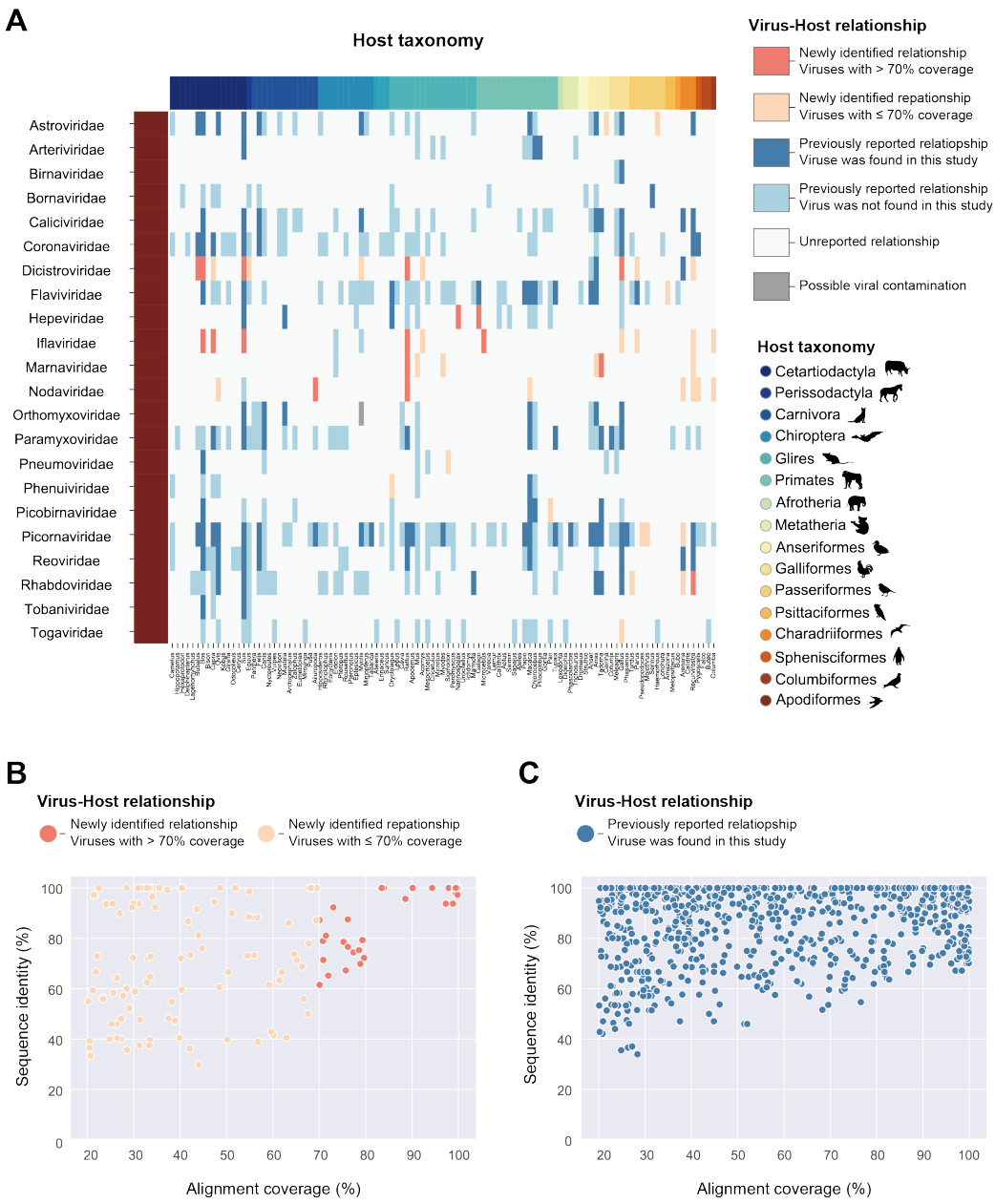


Figure 3. Search for unknown reservoir hosts and novel virus sequences.

(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:

879 a relationship was newly identified in this study, and the viral infection was detected with
880 > 70% alignment coverage (coral), a relationship was newly identified in this study, but
881 the viral infection was detected with ≤ 70% alignment coverage (salmon), a relationship
882 was previously reported, and the viral infection was also detected in this study (blue), a
883 relationship was previously reported, but the viral infection was not detected in this study
884 (light blue), a relationship was unreported so far (white), and a relationship was newly
885 identified in this study, but it may be attributed to contamination (gray) (**see Discussion**).
886 (B-C) Scatter plot between alignment coverages (x-axis) and sequence similarities with
887 known viruses (y-axis). Each dot represents the viral infections identified in this study.
888 Viral infections related to novel virus-host relationships are shown in (B), and those
889 related to known relationships are shown in (C). The dot colors correspond to virus-host
890 relationships shown in (A). Sequence identity represents the maximum value of the
891 percentage of identical matches obtained by TBLASTX alignment.
892

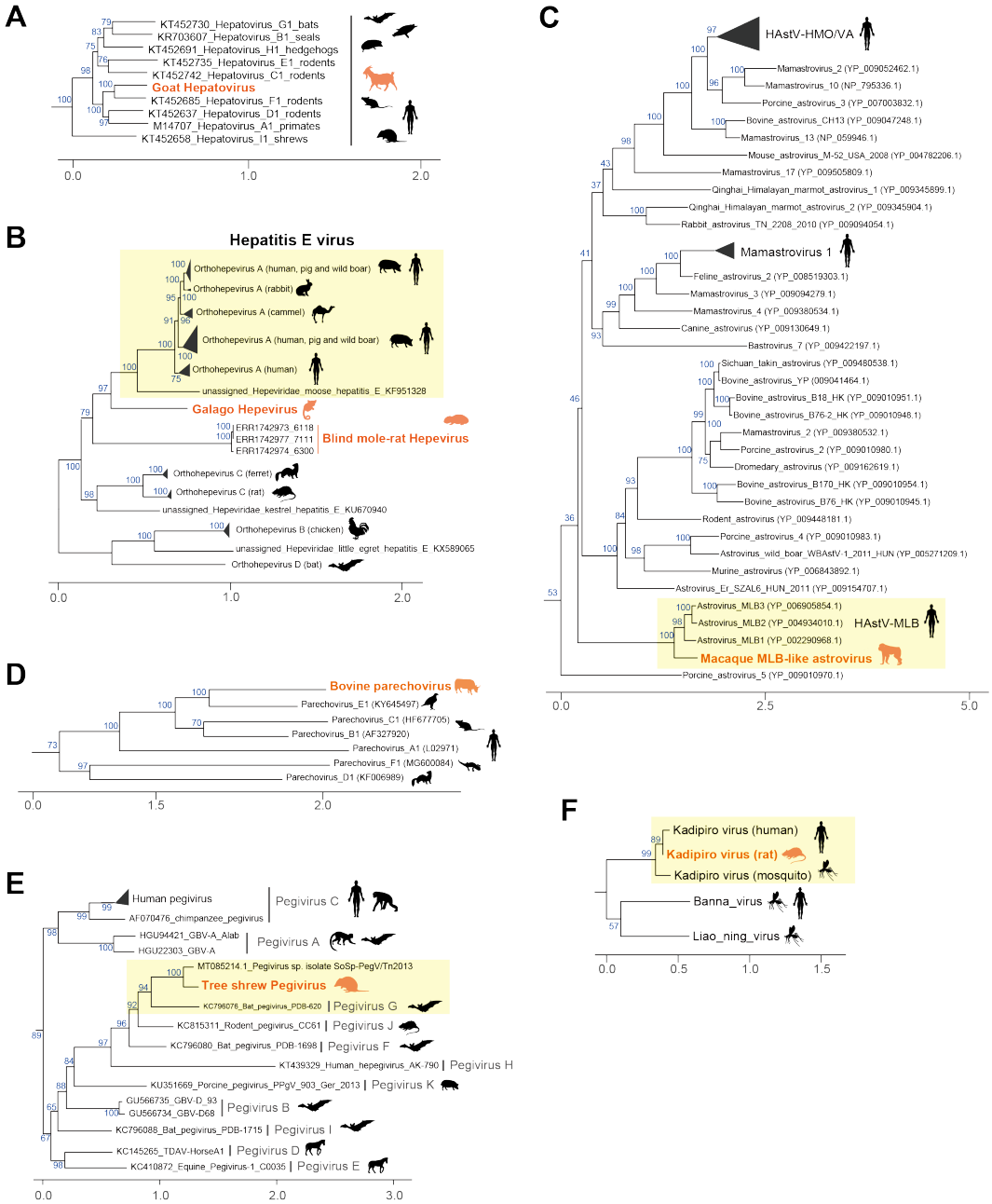


Figure 4. Characterization of virus sequences identified in this study.

(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), the genus *Pegivirus* of the family *Flaviviridae* (E), and the genus *Seadornavirus* of the family *Reoviridae* (F). These

899 phylogenetic trees were constructed based on the maximum likelihood method (**details**
900 **in Materials and Methods**). The orange labels indicate viruses identified in this study,
901 and the colored animal silhouette indicates the viral host species. The black label and
902 animal silhouette indicate known viruses and their representative hosts, respectively.
903 Scale bars indicate the genetic distance (substitutions per site). The blue labels on
904 branches indicate the bootstrap supporting values (%) with 1,000 replicates. Yellow
905 boxes highlight viruses genetically similar to the virus identified in this study.
906

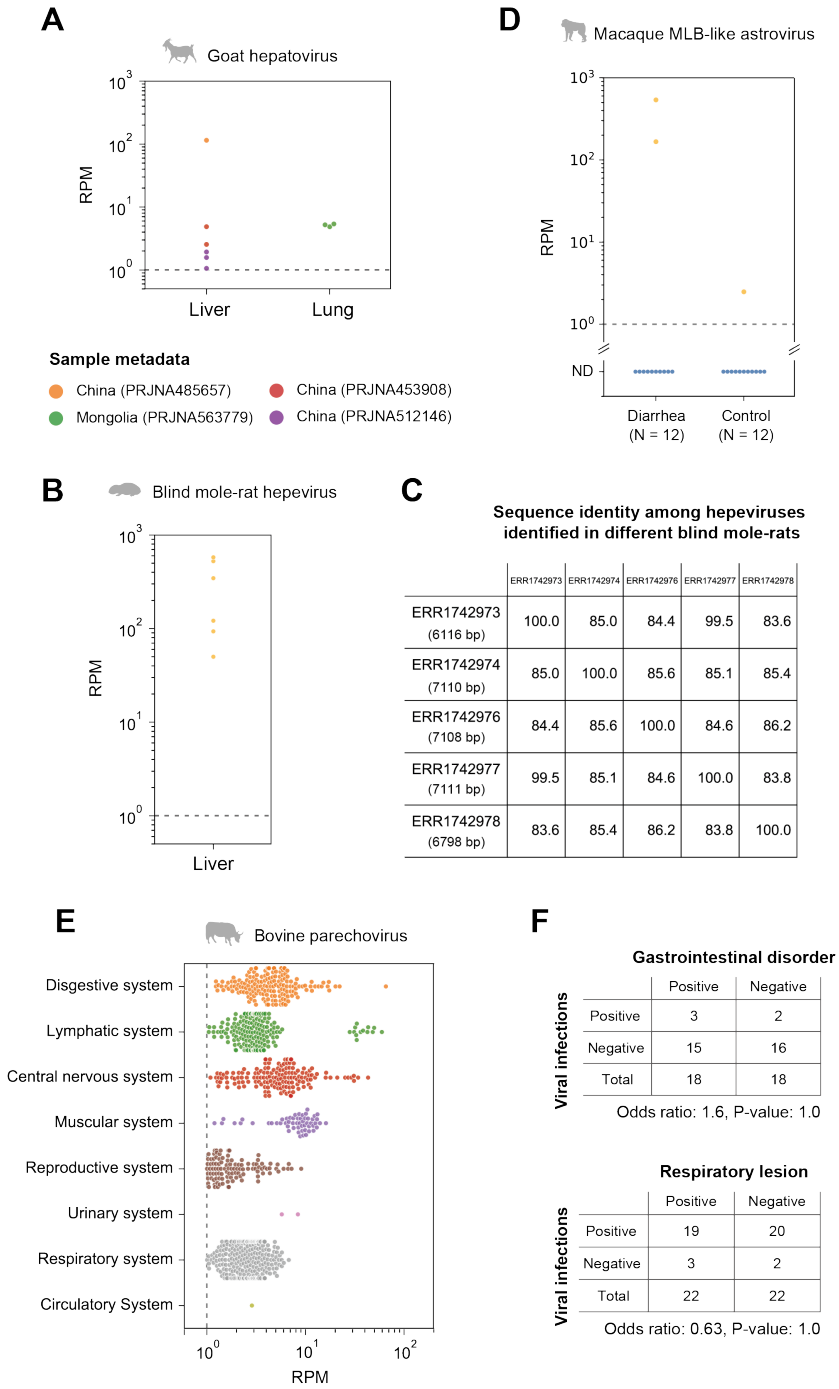


Figure 5. Detection of viral infections in the natural host population.

(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

912 or organ system. The gray dotted line indicates the criterion used to determine viral
913 infections (RPM: 1.0). The lower panel in (A) represents the sample metadata.
914 (C) Comparison of nucleotide sequence identity among the hepeviral sequences identified
915 in five different blind mole-rats. The numbers in parentheses in the row indicate the total
916 number of aligned sites between the viral contigs identified in each individual and the
917 blind mole-rat hepevirus identified in ERR1742977.
918 (D) Quantification of the macaque MLB-like viral infection levels in the patient with
919 diarrhea and control macaque monkeys. The x-axis indicates the diagnosis for the 24
920 monkeys, and the y-axis indicates the RPM. The average RPM for each individual is
921 plotted because six samples were collected from each individual. The dotted line indicates
922 the criterion used for detecting viral infections (RPM: 1.0). We considered samples with
923 RPMs below the criterion as non-detectable (ND).
924 (F) Association between the parechovirus infections and symptoms. The tables show the
925 number of RNA-seq data with and without the parechovirus infections in two independent
926 studies, which provide diagnostic information: gastrointestinal disorder (upper panel) and
927 respiratory lesion (lower panel).
928

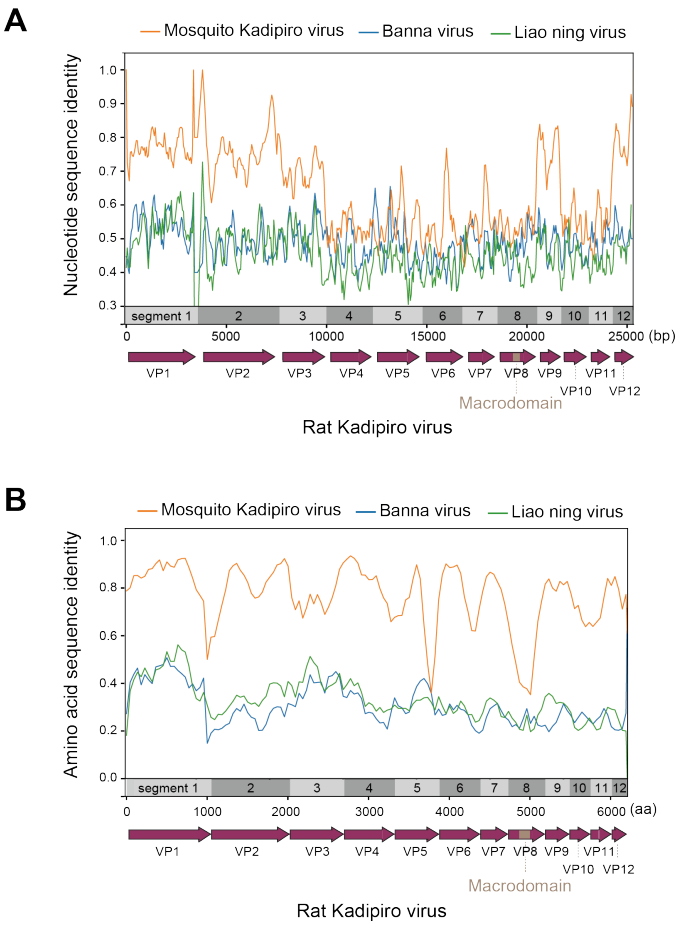
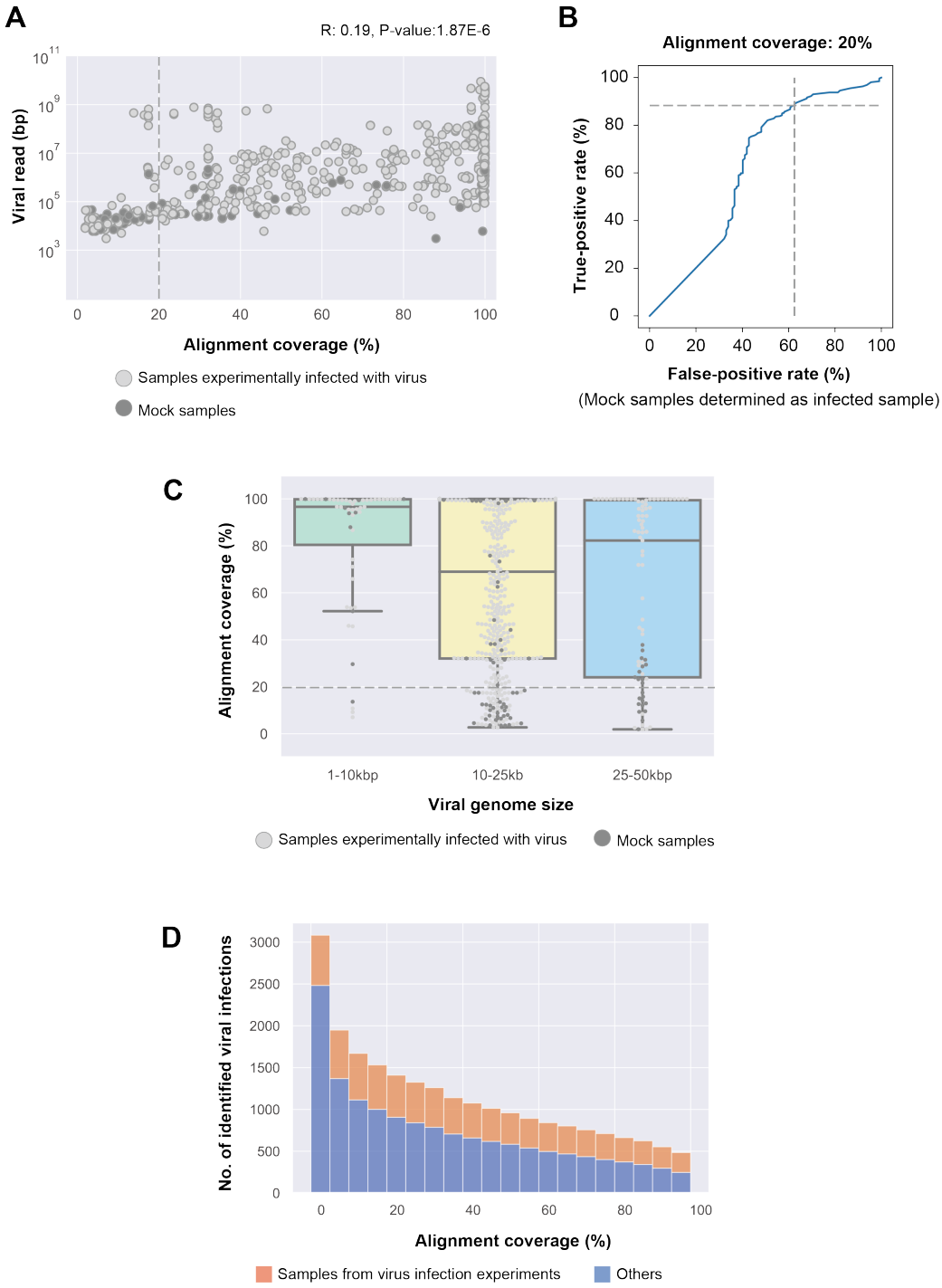


Figure 6. Sequence identity plots between rat Kadipiro virus and other known seadornaviruses.

Sequence identity plots using nucleotide sequences (A) and amino acid sequences (B). Line colors correspond to the viruses shown in the upper legend. The x-axis indicates the alignment positions, and the y-axis indicates sequence identity between rat Kadipiro virus and each virus. Light gray and dark gray boxes indicate the segments of rat Kadipiro virus. Dark purple arrows indicate open reading frames in the viral genome. Segment 8 of rat Kadipiro virus was expected to encode chimeric VP8, containing a macrodomain, shown as a light brown box.

940 **Supplemental Materials**

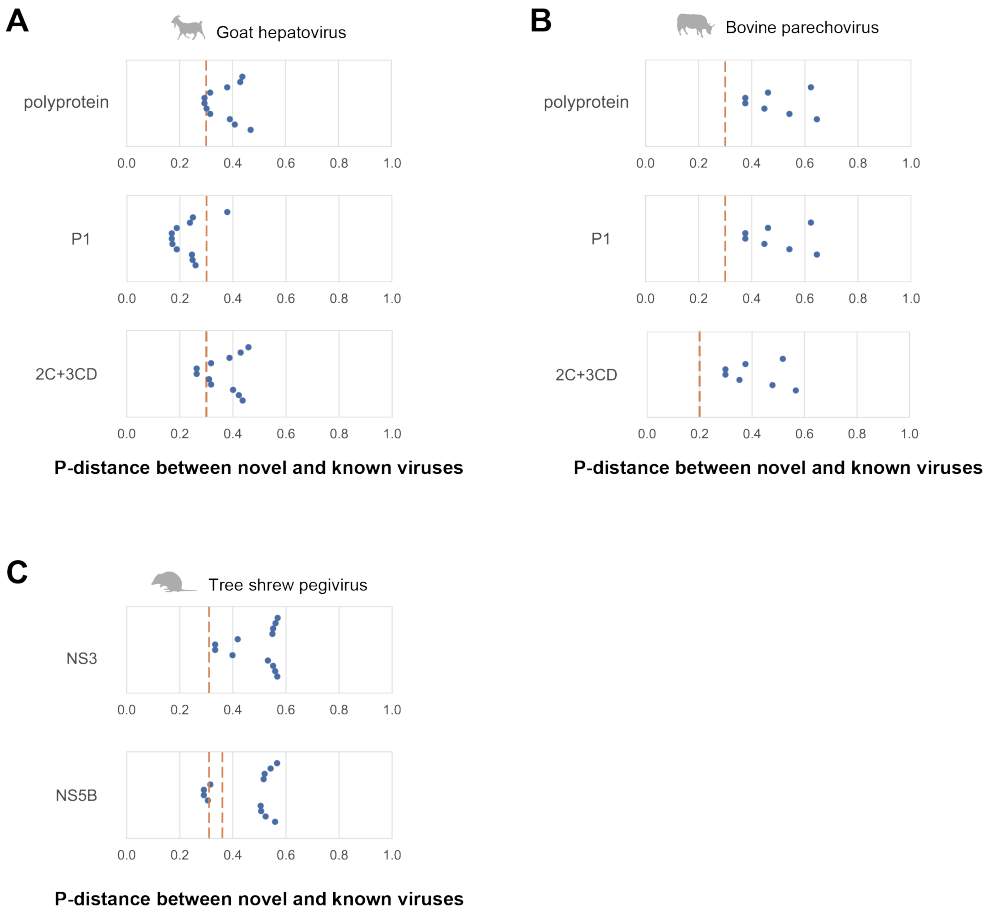


Supplementary Figure 1

941

942 **Supplemental Figure 1. Validation of the alignment coverage-based method for**
943 **detecting viral infections using samples obtained from viral infection experiments.**
944 (A) Comparison between the alignment coverage-based method and the viral read-based
945 method using samples obtained from viral infection experiments. The x-axis indicates
946 alignment coverage between viral contigs in each RNA-seq data and the reference viral
947 genome used for the experiments. The y-axis indicates the total read length of the virus
948 family used for the experiment, which was obtained from the NCBI SRA Taxonomy
949 Analysis Tool. Light gray dots indicate samples experimentally infected with viruses, and
950 dark gray dots indicate mock samples. R: Pearson's correlation coefficient. Dotted line
951 indicates 20% alignment coverage.
952 (B) Changes in the true-positive and the false-positive rates depending on the criteria to
953 determine viral infections. The true-positive rate (y-axis) indicates the number of samples
954 experimentally infected with viruses correctly determined as the infected sample, and the
955 false-positive rate (x-axis) indicates the number of mock samples determined as the
956 infected sample. Dotted line indicates the true-positive rate (88.3%) and the false-positive
957 rate (62.5%) when 20% alignment coverage was used as the criterion (**details in**
958 **Materials and Methods**).
959 (C) Detection rate of viral infections depending on the viral genome size. Box plots show
960 the distributions of alignment coverage of the viral genome with 1-10kbp (green), 10-
961 25kbp (yellow), and 25-50kbp (blue). Light gray dots indicate samples infected with
962 viruses experimentally, and dark gray dots indicate mock samples. Dotted line indicates
963 20% alignment coverage.
964 (D) The number of detected viral infections depending on the alignment coverage criteria.
965 The x-axis indicates alignment coverage used as a criterion for defining viral infections.

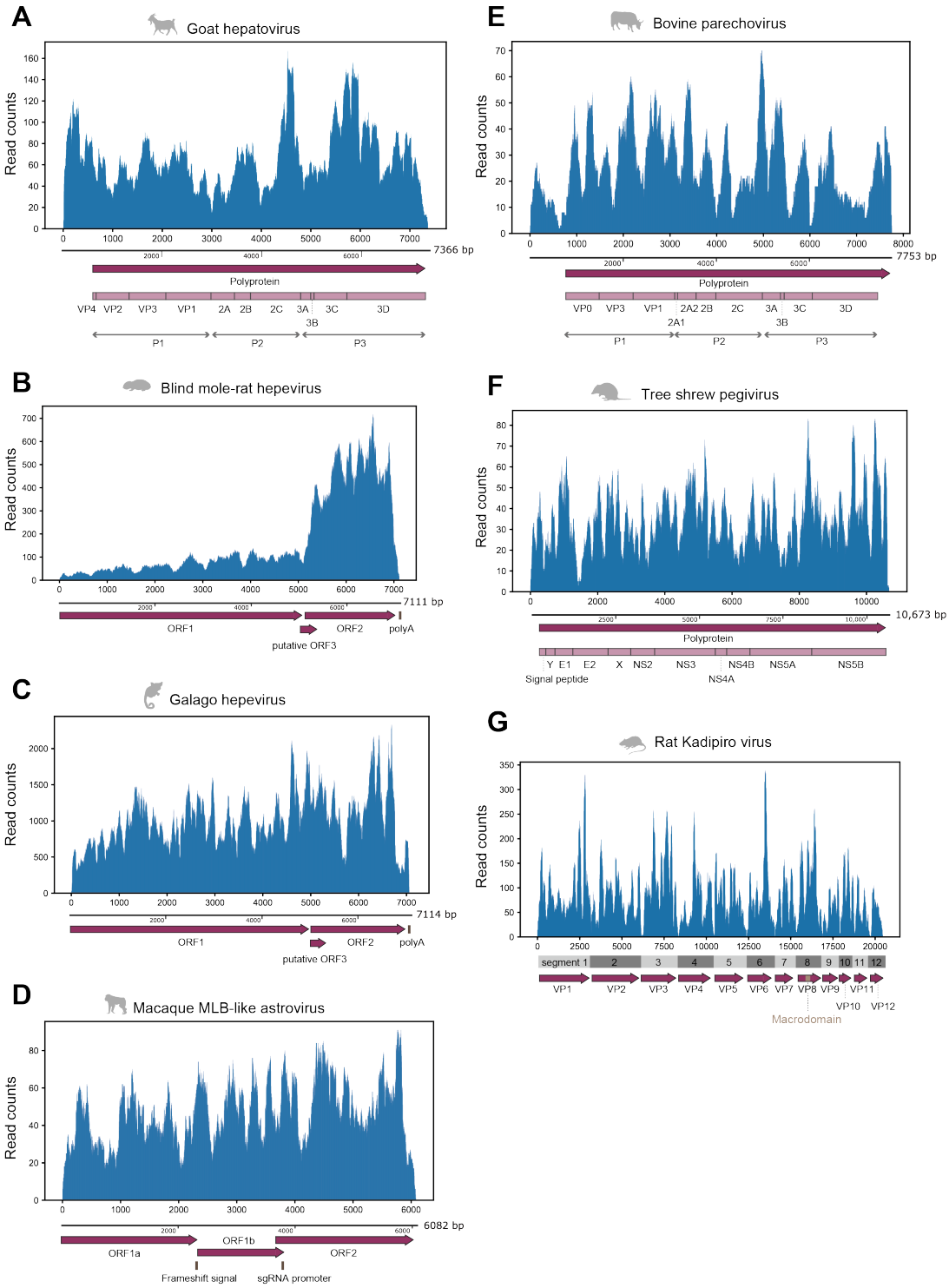
Bar graphs show the number of detected viral infections using the criterion shown on the x-axis. Filled colors indicate infections in samples from viral infection experiments (orange) or those in others (blue). When we used 20% alignment coverage as the criterion, a total of 1,410 viral infections were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



Supplementary Figure 2

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

974 (A-C) Genetic distance among the amino acid sequences of novel and known viruses in
975 the genera *Hepatovirus* (A), *Parechovirus* (B), and *Pegivirus* (C). The x-axis indicates
976 the proportion of different sites: p-distance. Each dot shows the amino acid sequence p-
977 distance between the novel and known virus species. The International Committee on
978 Taxonomy of Viruses species demarcation criteria are shown as orange dotted lines:
979 greater than 0.3 in polyprotein, P1, and 2C+3CD regions for hepatoviruses (A), greater
980 than 0.3 in polyprotein, P1 regions and 0.2 in 2C+3CD region for parechoviruses (B), and
981 greater than 0.31 in the NS3 region and 0.31-0.36 in the NS5B region for pegiviruses (C).
982



Supplementary Figure 3

983

984 **Supplemental Figure 3. Read mapping analysis using RNA-seq data in which the**
985 **viral sequence was identified.**
986 (A-G) Read distributions mapped to the viral sequence: goat hepatovirus (A), blind mole-
987 rat hepevirus (B), galago hepevirus (C), macaque MLB-like astrovirus (D), bovine
988 parechovirus (E), tree shrew pegivirus (F), and rat Kadipiro virus (G). The upper panel
989 shows the virus genomic positions (x-axis) and read counts at each position (y-axis). The
990 lower panel shows genomic annotations, such as protein-coding regions or signal
991 sequences. Dark purple arrows indicate open reading frames (ORFs) in the viral genome.
992 Light purple boxes show mature proteins predicted based on aligned positions with
993 reference viruses (**details in Materials and Methods**). Brown vertical lines indicate
994 nucleotide sequence features, such as polyadenylation signal (poly-A), ribosomal
995 frameshift signal (frameshift signal), and promoter sequence for subgenomic RNA
996 synthesis (sgRNA promoter). Light and dark gray boxes indicate the segments of rat
997 Kadipiro virus. Segment 8 of rat Kadipiro virus was expected to encode chimeric VP8,
998 containing a macrodomain, shown as a brown box in the dark purple arrow.
999
1000
1001 **Supplemental Dataset 1. List of Sequence Read Archive run accession numbers,**
1002 **genome file, and sequence assembly method.**
1003 **Supplemental Dataset 2. Information on RNA-seq data from experimental infection**
1004 **with viruses.**
1005 **Supplemental Dataset 3. Information on possible viral contamination excluded from**
1006 **the totalization.**

1007 **Supplemental Dataset 4. Information on manual curation for virus-host**
1008 **relationships.**
1009 **Supplemental Dataset 5. Accession numbers of viral sequences used for phylogenetic**
1010 **analyses, viral genomic annotations, and comparing the International Committee on**
1011 **Taxonomy of Viruses species demarcation criteria.**
1012 **Supplemental Dataset 6. Information on concatenated seadornaviral sequences.**
1013 **Supplemental Dataset 7. Sequence Read Archive run accessions used for mapping**
1014 **analyses.**
1015 **Supplemental Dataset 8. Sample metadata in which the goat hepatoviral infections**
1016 **were detected.**
1017 **Supplemental Dataset 9. Sample metadata in which the blind mole-rat hepeviral**
1018 **infections were detected.**
1019 **Supplemental Dataset 10. Sample metadata in which the galago hepeviral infections**
1020 **were detected.**
1021 **Supplemental Dataset 11. Sample metadata in which the macaque MLB-like**
1022 **astrovirus infections were detected.**
1023 **Supplemental Dataset 12. Sample metadata in which the bovine parechovirus**
1024 **infections were detected.**
1025 **Supplemental Dataset 13. Sample metadata in which the tree shrew pegiviral**
1026 **infections were detected.**
1027 **Supplemental Dataset 14. Bioinformatics tools and their versions used in this study.**
1028