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.

Importance

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

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.

We also discovered seven nearly complete viral genomes in livestock, wild, and laboratory animals. Phylogenetic analyses showed some 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.

Results

Detection of RNA viral infections hidden in public sequence data

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

We used more than 46,000 mammalian and avian RNA-seq data to investigate infections of 33 RNA virus families reported to infect vertebrates. Consequently, we identified 907 infections of 22 RNA virus families in 709 sequencing data from 56 host species (**Fig. 2A**). These results indicate that analyzing public sequencing data by metagenomic analysis is useful for identifying hidden viral infections.

Frequent detection of diverse virus families in bird samples

Many viral infectious diseases associated with birds have been reported so far (11), such as influenza A virus (12, 13) and West Nile virus (14). In this study, we frequently detected viral infections in bird samples (**Fig. 2B**). The odds ratio of RNA virus detection in birds compared with that in mammals was 3.28. Furthermore, among the investigated species, we found relatively high viral detection rates in Gallus and Anas species at 20.1% and 8.7%, respectively (**Fig. 2C**). We also found infections of 12 and 8 virus families in Gallus and Anas species, respectively (**Fig. 2D**). These results indicate that birds, especially Gallus and Anas species, are frequently infected with various virus families, suggesting that these species are reservoirs for a wide variety of viruses (**see Discussion**).

Identification of unknown reservoir hosts at virus family levels

To identify novel virus-host relationships at virus family levels, we compared our data with known virus-host relationships provided in the Virus-Host Database (15) (**Fig. 3A**). This database lists virus-host relationships based on the identification of viral sequences from a host animal. Using this database for comparison, we found 50 newly identified virus-host relationships, and 17 of them were identified with more than 70% alignment coverage. Notably, we identified nearly complete genomic sequences classified into the family *Hepeviridae* in Spalax and Galago species for the first time. These discoveries expanded our understanding of hepeviral host ranges (details of the viral characteristics are described in the section: "*Hepeviruses in blind mole-rats and a galago: expanding understanding of the hepatitis E virus host range*"). A novel relationship was also identified between the family *Rhabdoviridae* and Recurvirostra species. We did not perform further investigations because the complete rhabdovirus genome could not be obtained, although the alignment coverage was more than 70%. Additionally, novel virus-host relationships were also found in the families *Dicistroviridae*, *Iflaviridae*, *Marnaviridae*, and *Nodaviridae*, suggesting that these viral host ranges are much broader than previously expected. It should be noted that these relationships may be due to contamination from environmental viruses, because few species in these virus families have been reported to infect mammals or birds (16-19) (**see Discussion**).

Investigation of novel viruses with complete genomic sequences

To identify novel sequences comparable to a complete viral genome, we simultaneously analyzed sequence similarity with known viruses and the alignment coverages with reference viral genomic sequences (**Figs. 3B-C**). We found some viral sequences showing low sequence similarity with known viruses and high alignment coverage, which were expected to be novel viruses with a nearly complete genome. Therefore, we further characterized these viral sequences by phylogenetic analyses, annotations of viral genomic features, and quantification of viral reads in RNA-seq data (**Figs. 4-6 and S2-3**). Consequently, we discovered seven viruses: 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.

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

Hepatitis A virus (HAV), belonging to the genus *Hepatovirus* of the family *Picornaviridae*, can cause acute and fulminant hepatitis and is typically transmitted via fecal-oral routes, including contaminated water or foods (20). The World Health

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

We further analyzed the hepatovirus prevalence in a natural host population by quantifying the viral reads in other goat RNA-seq data because this virus was initially identified in only one goat sample. Among 1,593 goat samples, we found the viral infection in nine samples from four independent studies with > 1.0 read per million reads (RPM) (**Fig. 5A and Dataset S8**). These hepatoviral infections were detected in goat liver and lung samples, suggesting that the goat hepatovirus can infect tissues other than the liver. Although the lungs are not considered preferential tissues for hepatoviral replication, a previous report also detected seal hepatoviral RNAs in the lungs (21). The infected goat samples were collected in East Asia, including China and Mongolia. Therefore, goat hepatoviruses may be prevalent in the natural host population, suggesting this virus can be a new threat to public health through the contamination of water and foods by infected animals.

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

Several million infections of hepatitis E virus (HEV) are estimated to occur worldwide; the WHO reported approximately 44,000 deaths due to HEV infection in 2015 (https://www.who.int/news-room/fact-sheets/detail/hepatitis-e). Here, we found hepeviruses, classified into the same viral family as HEV, in blind mole-rats and a galago for the first time (**Fig. 3A**). Phylogenetic analysis indicated that these hepeviruses formed a single cluster with moose HEV (22) and members of Orthohepevirus A that infect humans, pigs, rabbits, and camels (23) (**Fig. 4B**). However, the hepeviruses identified in this study appeared to have an early divergence from the HEV common ancestor. These results suggest a high diversity and broader host range of HEV-like viruses.

The blind mole-rat hepevirus was identified in host livers, which coincided with the tissue tropism of HEV (24). Additionally, we found that the 3'-portion of the blind mole-rat hepevirus genome was highly transcribed (**Fig. S3B**), suggesting the transcription of subgenomic RNAs (25). In contrast, we could not determine the tissues infected by the galago hepevirus because the relevant metadata were not available. Further, we did not observe a clear read-mapping pattern that suggests any subgenomic RNA transcription in the galago sample (**Fig. S3C**).

We also investigated the spread of these viruses in a natural population using RNA-seq data from blind mole-rats and galagos. Among 91 RNA-seq data from blind mole-rats, we detected the hepeviral infections in six samples (**Fig. 5B**). The infected individuals were from the same experiment, which were captured and kept as laboratory animals in Israel (**Dataset S9**). There were two possibilities about when the hepeviruses have infected blind mole-rats: the hepeviruses had already infected these blind mole-rats when they were captured, or the viral infections had spread during the maintenance of these individuals in the laboratory. To explore these possibilities, we investigated the inter-individual diversity of the hepevirus sequences. We found that these individuals were infected with relatively diverse hepeviruses representing nucleotide sequence identities ranging from 83.6% to 99.5% (**Fig. 5C**). These results suggest that several individuals had already been infected with distinct hepeviruses in the wild before being captured. The galago hepeviral infections were detected in only two samples originating from a study in which we first identified the virus (**Dataset S10**). This may be simply because only four galago RNA-seq data obtained from the same study were available. Taken together, we suggest that these hepeviruses can become a new threat to public health, similar to HEV.

MLB-like astrovirus detected in macaque monkeys with chronic diarrhea

We found an astrovirus genetically similar to human astrovirus MLB (HAstV-MLB) in fecal samples of macaque monkeys (**Fig. 4C**). Although HAstV-MLB infections are typically asymptomatic (26, 27), several studies have reported the viral detection in cases with diarrhea (28), encephalitis (29), or meningitis (30). Interestingly, the macaque MLB-like astrovirus was found in macaque monkeys with chronic diarrhea. We analyzed the viral read amounts in the patient (n = 12) and control (n = 12) monkeys to assess the association between MLB-like astroviral infections and symptom prevalence (**Fig. 5D and Dataset S11**). We detected abundant MLB-like astroviral reads in two patients, suggesting that the viral infections are associated with host symptoms. However, we did not observe the viral infection in other patients; further, we found the infection in a control individual, although the viral read amount was approximately 100 times less than those of the patients. Additionally, a previous study reported that monkeys, in which partial sequences of MLB-like astroviruses were detected, had no obvious clinical signs, including diarrhea (31). Thus, further experiments are needed to clarify the pathogenesis of MLB-like astrovirus. Considering that there is no current experimental system for examining HAstV-MLB infections (27), our findings suggest that macaque monkeys can be used as animal model systems for researching MLB-like astroviruses.

Silent infections of bovine parechovirus having a broad tissue tropism

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

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

We also investigated the prevalence of this parechovirus infection in a natural host population using public RNA-seq data (**Fig. 5E and Dataset S12**). Among 8,284 cow samples, we detected the parechovirus infections in 944 samples from eight independent studies with > 1.0 RPM. The viral infections were detected in various tissues, such as the digestive, lymphatic, and central nervous system. These results suggest a broad tissue tropism of the bovine parechovirus. To assess the parechovirus pathogenicity, we analyzed the viral prevalence among 36 or 44 samples with a diagnosis for a gastrointestinal disorder or respiratory lesion, respectively. We did not observe a significant association between the viral infections and the presence/absence of symptoms in these two studies (**Fig. 5F**). These results indicate that bovine parechovirus infections may be asymptomatic, similar to the typical outcome of human parechoviral infections. Furthermore, this also suggests that infected cows can spread parechoviral infections as silent reservoirs.

Geographical expansion of tree shrew pegivirus infection associated with host migration

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

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

Kadipiro virus in rats: a possible arbovirus that infects mosquitoes and mammals

We identified Kadipiro virus (KDV), a member of the genus *Seadornavirus* of the family *Reoviridae*, in rat spinal cord samples. Mosquitoes have been considered the hosts of KDV (36); however, a previous report identified several KDV segments in plasma samples from febrile humans (37). Phylogenetic analysis using VP1 amino acid sequences indicated that the KDVs identified in humans, rats, and mosquitoes formed a single cluster (**Fig. 4F**). Additionally, Banna virus, classified into the same genus as KDV, is an arbovirus that transmits between mosquitoes and mammals, including humans, cows, and pigs (38). Taken together with previous reports on seadornaviruses, KDV is also expected to be an arbovirus.

Next, we calculated the sequence similarity among all segments between rat KDV and known seadornaviruses to characterize the entire rat KDV genome (**Fig. 6**). We found that several segments of rat KDV, especially segments 4-8, 10, and 11, showed relatively low nucleotide sequence identities to those of mosquito KDV (**Fig. 6A**), even though the amino acid sequences of rat KDV showed approximately 80% identity to mosquito KDV throughout (**Fig. 6B**). These results suggest that rat KDV segments were diversified among KDVs at the nucleotide sequence level due to virus-host coevolution of codon usage and segment reassortment.

Various viral families, including coronaviruses and togaviruses, have been reported to hijack the host macrodomain, leading to changes in virulence or immune responses during viral infections (39). Interestingly, segment 8 in rat KDV may encode chimeric VP8 containing a seadornaviral double-stranded RNA-binding domain (36) and a macrodomain (**Fig. 6**). However, the mosquito KDV VP8 lacks a macrodomain. We could not confirm whether human KDV encodes chimeric protein because human KDV segment 8 was not identified in the previous study (37). Nonetheless, the presence of this domain may be related to the determination of KDV host ranges. However, further experiments are needed to confirm chimeric VP8 expression and function.

Discussion

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

In this study, we determined viral infections by a combination of sequence assembly and the alignment coverage-based method to solve several issues in viral metagenomic analysis (**Fig. 1A**). One of the problems is detecting infections in data with a small number of viral reads because almost all public sequencing data were collected without using virus enrichment strategies. The result that most virus contigs were shorter than the reference viral genomes reflects this difficulty (**Figs. 1B-C**). To resolve this issue, we determined viral infections by the alignment coverage-based method, which uses relatively short viral sequences as clues (**Figs. 1A and S1**). Consequently, we succeeded in detecting over 900 RNA viral infections in public deep sequencing data (**Fig. 2A**). Another problem in viral metagenomic analysis is that the viral detectability depends on sequence similarity with known viruses. In this study, we discovered seven viral genomes by sequence assembly (**Fig. 4**). Notably, these viral infections were undetectable in almost all samples, even at the virus family and genus levels, by the NCBI SRA Taxonomy Analysis Tool, which determines the taxonomic composition of reads in the RNA-seq data without sequence assembly (**Dataset S8-S13**). These results indicate that identifying viral sequences based on sequence assembly would effectively elucidate virus diversity. Taken together, our strategy using sequence assembly and the alignment coverage-based method can efficiently detect known and novel viral infections in publicly available sequencing data.

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

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

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

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

In conclusion, we demonstrated the reusability of public sequencing data for monitoring viral infections and discovering novel viral sequences, and elucidated diverse RNA viruses hidden in animal samples. Our findings also emphasize the necessity of continuous surveillance for viral infections using public sequencing data to prepare for future viral pandemics, as well as the importance of developing a fundamental bioinformatics platform for surveillance (46, 47).

Materials and Methods

Sequence assembly using publicly available RNA-seq data

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

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

Identification of contigs originating from RNA viruses

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

Totalization of RNA viral infections in public RNA-seq data

Since most viral contigs were shorter than the reference viral genomic sizes (**Figs. 1B-C**), we sought to determine viral infections based on the alignment coverage-based method (**Fig. 1A**). First, we performed sequence alignment by TBLASTX (version 2.9.0) using viral contigs from the same RNA-seq data and complete viral genomes in the NCBI RefSeq genomic viral database (version 20200824). Next, we calculated the alignment coverage with the genome of each viral species: the proportion of aligned sites in the entire reference viral genome. In this study, we considered that an infection of the viral family is present if the alignment coverage was greater than 20%. Validation of this totalization method and evaluation of the criteria are described in the next section (**Fig. S1**). Furthermore, we manually checked sequences with more than 70% alignment coverage and more than 95% identity with known viruses in the TBLASTX alignment to examine possible contamination with laboratory viral strains, as well as experimentally inoculated viruses. We excluded experimental viral infections (**Dataset S2**) and possible contamination (**Dataset S3**) from the final totalization (**Fig. 2A**). Overall, we investigated the infections of 33 RNA viral families reported to infect vertebrates in 311 host species.

Validation of the procedure used to totalize viral infections

Using samples obtained from viral infection experiments, we first compared the alignment coverage-based method with that based on viral read amounts in order to validate the detection rate of viral infections of our method (**Fig. S1 and Dataset S2**). We obtained the read amounts derived from experimentally infected viruses from the NCBI SRA Taxonomy Analysis Tool results (https://github.com/ncbi/ngs-tools/tree/tax/tools/tax). The calculation procedure for alignment coverage between viral contigs in each RNA-seq data and viral reference genomes is described in the previous section. We observed a positive correlation between the alignment coverage and viral read amounts (Pearson's correlation coefficient: 0.19, p-value: 1.87E-6) (**Fig. S1A**). Among the samples collected from experiments of viral infections, the true-positive rate (the detection rate of experimentally inoculated viruses) was 88.3%, and the false-positive rate (the rate that mock samples were determined to be infected samples) was 62.5% when we used 20% alignment coverage as the criterion for determining viral infections (**Fig. S1B**). The relatively high false-positive rate may be due to similar amounts of viral reads in some mock samples as those in infected samples (**Fig. S1A**). Next, we analyzed the association between alignment coverages and viral genome size (**Fig. S1C**) because the detectability of viral infections in our method may depend on the reference viral genome size. As expected, we observed a tendency for viruses with small genomes to be detected with relatively high alignment coverage. However, more than 80% of experimentally infected viral infections were detected with more than 20% alignment coverage, regardless of the viral genome size. Based on these results, we established the alignment coverage of 20 % to totalize the viral infections. Consequently, we identified a total of 1,410 RNA viral infections, including 503 infections in samples from viral infectious experiments (**Fig. S1D**).

Collection of information on experimentally infected viruses

To exclude experimentally infected viruses from the final totalization, we analyzed the experimental background of RNA-seq data. We first collected the experimental descriptions of RNA-seq data: title and abstract from the NCBI BioProject database (57). Then, we manually checked the terms relevant to viral infections in the descriptions, focusing on viral name abbreviations and viral vector usage. We listed the obtained information about viral infection experiments in **Dataset S2**.

Summarization of virus-host relationships

To identify novel reservoir hosts at the viral family levels, we compared the virus-host relationships identified in this study with the dataset provided by the Virus-Host DB (version: 20200629) (15). We define a “novel virus-host relationship” as one in which the viral sequence has not been reported in the host. The virus-host relationships at the viral family level were categorized as 1) a novel relationship detected with > 70% alignment coverage, 2) a novel relationship detected with ≤ 70% alignment coverage, 3) a known relationship that was also detected in this study, 4) a known relationship that was not identified in this study, 5) a relationship unreported so far, and 6) a novel relationship, which was possibly derived from contamination (**see Discussion**). To avoid misclassification of the relationships, we analyzed reports manually by searching the NCBI PubMed and Nucleotide databases using the combination of the host genus and viral family names: for example, [“Pan” AND “Picobirnaviridae”]. The results of the manual curation are listed in **Dataset S4**.

Characterization of viral genomic architectures

Open reading frames (ORFs) and polyadenylation signals in the viral genomes were predicted by SnapGene software (snapgene.com). The positions of mature proteins, frameshift signal sequences, and subgenomic RNA promoter sequences were predicted based on sequence alignment using novel and reference viral sequences. The sequence alignments were constructed by MAFFT (version 7.407) (58) with the option “--auto”. The reference viral sequences used for the genome annotations are listed in **Dataset S5**. The macrodomain in rat KDV segment 8 was identified by CD-search (59) using the CDD v3.18 database (60). 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). For astroviruses (**Fig. 4C**) and seadornaviruses (**Fig. 4F**), we collected reference sequences from the RefSeq protein viral database (version 20210204) and extracted their amino acid sequences as follows: ORF2 protein for viruses classified in the family *Astroviridae* and VP1 protein for viruses classified in the genera *Seadornaviruses* and *Cardoreoviruses*. The MSAs of reference and novel viral sequences were constructed by MAFFT with options "--add" and "--keeplength". MSAs using astroviruses and seadornaviruses were 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) (61). The substitution models were selected based on the Bayesian information criterion provided by ModelFinder (62): GTR+R8 for **Fig. 4A**, LG+F+R4 for **Fig. 4B**, LG+F+R5 for **Fig. 4C**, TVM+R9 for **Fig.4D**, GTR+R7 for **Fig. 4E**, and Blosum62 for **Fig. 4F**. The branch supportive values were measured as the ultrafast bootstrap by UFBoot2 (63) with 1,000 replicates. Tree visualization was performed by the ggtree package (version 2.2.1) (64). Sequence accession numbers used for the phylogenetic analyses are listed in **Dataset S5**.

Comparison with the ICTV species demarcation criteria

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

Calculation of genetic distances among the entire sequence of seadornaviral segments

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

Mapping analyses using viral genomes identified in this study

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

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

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

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

Comparison of hepeviral sequences identified in different blind mole-rats

We compared nucleotide sequence identities among the hepeviral sequences found in five different individuals to predict when these viruses infected the blind mole-rats. The sequence comparison was performed by BLASTN (version 2.11.0) with default parameters. Because most hepeviral sequences were detected as short contigs, sequence identities were represented by the percentage of identical matches in the longest aligned region between the hepeviral sequences (**Fig. 5C**). We also analyzed the total of aligned length between contigs identified in each individual and the hepeviral genome identified in ERR1742977 and confirmed that these contigs covered 86.0-99.9% of the blind mole-rat hepevirus genome.

Data Availability

Bioinformatics tools and their versions are listed in **Dataset S14**.

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.

References

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

8. Leinonen R, Sugawara H, Shumway M. 2011. The Sequence Read Archive. Nucleic Acids Research 39:D19-D21.

9. Iwamoto M, Shibata Y, Kawasaki J, Kojima S, Li Y-T, Iwami S, Muramatsu M, Wu H-L, Wada K, Tomonaga K, Watashi K, Horie M. 2021. Identification of novel avian and mammalian deltaviruses provides new insights into deltavirus evolution. Virus Evolution 7.

10. Horie M, Akashi H, Kawata M, Tomonaga K. 2020. Identification of a reptile lyssavirus in Anolis allogus provided novel insights into lyssavirus evolution. Virus Genes doi:10.1007/s11262-020-01803-y.

11. Nabi G, Wang Y, Lü L, Jiang C, Ahmad S, Wu Y, Li D. 2021. Bats and birds as viral reservoirs: A physiological and ecological perspective. Science of The Total Environment 754:142372.

12. Olsen B, Munster VJ, Wallensten A, Waldenstrom J, Osterhaus ADME, Fouchier RAM. 2006. Global Patterns of Influenza A Virus in Wild Birds. Science 312:384-388.

13. Lycett SJ, Duchatel F, Digard P. 2019. A brief history of bird flu. Philosophical Transactions of the Royal Society B: Biological Sciences 374:20180257.

14. Habarugira G, Suen WW, Hobson-Peters J, Hall RA, Bielefeldt-Ohmann H. 2020. West Nile Virus: An Update on Pathobiology, Epidemiology, Diagnostics, Control and “One Health” Implications. Pathogens 9:589.

15. Mihara T, Nishimura Y, Shimizu Y, Nishiyama H, Yoshikawa G, Uehara H, Hingamp P, Goto S, Ogata H. 2016. Linking Virus Genomes with Host Taxonomy. Viruses 8:66.

16. Scherer WF, Verna JE, Richter GW. 1968. Nodamura Virus, an Ether- and Chloroform-Resistant Arbovirus from Japan *. The American Journal of Tropical Medicine and Hygiene 17:120-128.

17. Reuter G, Pankovics P, Gyöngyi Z, Delwart E, Boros Á. 2014. Novel dicistrovirus from bat guano. Archives of Virology 159:3453-3456.

18. Greninger AL, Jerome KR. 2016. Draft Genome Sequence of Goose Dicistrovirus. Genome Announcements 4:e00068-16.

19. Yinda CK, Zeller M, Conceição-Neto N, Maes P, Deboutte W, Beller L, Heylen E, Ghogomu SM, Van Ranst M, Matthijnssens J. 2016. Novel highly divergent reassortant bat rotaviruses in Cameroon, without evidence of zoonosis. Scientific Reports 6:34209.

20. Lemon SM, Walker CM. 2019. Hepatitis A Virus and Hepatitis E Virus: Emerging and Re-Emerging Enterically Transmitted Hepatitis Viruses. Cold Spring Harbor Perspectives in Medicine 9:a031823.

21. Anthony SJ, St. Leger JA, Liang E, Hicks AL, Sanchez-Leon MD, Jain K, Lefkowitch JH, Navarrete-Macias I, Knowles N, Goldstein T, Pugliares K, Ip HS, Rowles T, Lipkin WI. 2015. Discovery of a Novel Hepatovirus (Phopivirus of Seals) Related to Human Hepatitis A Virus. mBio 6:e01180-15.

22. Lin J, Norder H, Uhlhorn H, Belák S, Widén F. 2014. Novel hepatitis E like virus found in Swedish moose. Journal of General Virology 95:557-570.

23. Purdy MA, Harrison TJ, Jameel S, Meng XJ, Okamoto H, Van Der Poel WHM, Smith DB. 2017. ICTV Virus Taxonomy Profile: Hepeviridae. Journal of General Virology 98:2645-2646.

24. Wang B, Meng X-J. 2021. Hepatitis E virus: host tropism and zoonotic infection. Current Opinion in Microbiology 59:8-15.

25. Graff J, Torian U, Nguyen H, Emerson SU. 2006. A Bicistronic Subgenomic mRNA Encodes both the ORF2 and ORF3 Proteins of Hepatitis E Virus. Journal of Virology 80:5919-5926.

26. Cortez V, Meliopoulos VA, Karlsson EA, Hargest V, Johnson C, Schultz-Cherry S. 2017. Astrovirus Biology and Pathogenesis. Annual Review of Virology 4:327-348.

27. Johnson C, Hargest V, Cortez V, Meliopoulos V, Schultz-Cherry S. 2017. Astrovirus Pathogenesis. Viruses 9:22.

28. Finkbeiner SR, Kirkwood CD, Wang D. 2008. Complete genome sequence of a highly divergent astrovirus isolated from a child with acute diarrhea. Virology Journal 5:117.

29. Sato M, Kuroda M, Kasai M, Matsui H, Fukuyama T, Katano H, Tanaka-Taya K. 2016. Acute encephalopathy in an immunocompromised boy with astrovirus-MLB1 infection detected by next generation sequencing. Journal of Clinical Virology 78:66-70.

30. Cordey S, Vu D-L, Schibler M, L'Huillier AG, Brito F, Docquier M, Posfay-Barbe KM, Petty TJ, Turin L, Zdobnov EM, Kaiser L. 2016. Astrovirus MLB2, a New Gastroenteric Virus Associated with Meningitis and Disseminated Infection. Emerging Infectious Diseases 22:846-853.

31. Karlsson EA, Small CT, Freiden P, Feeroz M, Matsen FA, San S, Hasan MK, Wang D, Jones-Engel L, Schultz-Cherry S. 2015. Non-Human Primates Harbor Diverse Mammalian and Avian Astroviruses Including Those Associated with Human Infections. PLOS Pathogens 11:e1005225.

32. Britton PN, Jones CA, Macartney K, Cheng AC. 2018. Parechovirus: an important emerging infection in young infants. Medical Journal of Australia 208:365-369.

33. Zell R, Delwart E, Gorbalenya AE, Hovi T, King AMQ, Knowles NJ, Lindberg AM, Pallansch MA, Palmenberg AC, Reuter G, Simmonds P, Skern T, Stanway G, Yamashita T. 2017. ICTV Virus Taxonomy Profile: Picornaviridae. Journal of General Virology 98:2421-2422.

34. Simmonds P, Becher P, Bukh J, Gould EA, Meyers G, Monath T, Muerhoff S, Pletnev A, Rico-Hesse R, Smith DB, Stapleton JT. 2017. ICTV Virus Taxonomy Profile: Flaviviridae. Journal of General Virology 98:2-3.

35. Wu Z, Han Y, Liu B, Li H, Zhu G, Latinne A, Dong J, Sun L, Su H, Liu L, Du J, Zhou S, Chen M, Kritiyakan A, Jittapalapong S, Chaisiri K, Buchy P, Duong V, Yang J, Jiang J, Xu X, Zhou H, Yang F, Irwin DM, Morand S, Daszak P, Wang J, Jin Q. 2021. Decoding the RNA viromes in rodent lungs provides new insight into the origin and evolutionary patterns of rodent-borne pathogens in Mainland Southeast Asia. Microbiome 9.

36. Attoui H, De Micco P, De Lamballerie X, Billoir F, Biagini P. 2000. Complete sequence determination and genetic analysis of Banna virus and Kadipiro virus: proposal for assignment to a new genus (Seadornavirus) within the family Reoviridae. Journal of General Virology 81:1507-1515.

37. Ngoi CN, Siqueira J, Li L, Deng X, Mugo P, Graham SM, Price MA, Sanders EJ, Delwart E. 2016. The plasma virome of febrile adult Kenyans shows frequent parvovirus B19 infections and a novel arbovirus (Kadipiro virus). Journal of General Virology 97:3359-3367.

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 L-H, Tang Q, Liang G-D. 2010. Banna Virus, China, 1987–2007. Emerging Infectious Diseases 16:514-517.

39. Rack JGM, Perina D, Ahel I. 2016. Macrodomains: Structure, Function, Evolution, and Catalytic Activities. Annual Review of Biochemistry 85:431-454.

40. Antipov D, Raiko M, Lapidus A, Pevzner PA. 2020. MetaviralSPAdes: assembly of viruses from metagenomic data. Bioinformatics 36:4126-4129.

41. Yahara K, Suzuki M, Hirabayashi A, Suda W, Hattori M, Suzuki Y, Okazaki Y. 2021. Long-read metagenomics using PromethION uncovers oral bacteriophages and their interaction with host bacteria. Nature Communications 12.

42. Dreher TW. 1999. FUNCTIONS OF THE 3′-UNTRANSLATED REGIONS OF POSITIVE STRAND RNA VIRAL GENOMES. Annual Review of Phytopathology 37:151-174.

43. Munis AM, Bentley EM, Takeuchi Y. 2020. A tool with many applications: vesicular stomatitis virus in research and medicine. Expert Opinion on Biological Therapy 20:1187-1201.

44. Feehan BJ, Penin AA, Mukhin AN, Kumar D, Moskvina AS, Khametova KM, Yuzhakov AG, Musienko MI, Zaberezhny AD, Aliper TI, Marthaler D, Alekseev KP. 2019. Novel Mammalian orthorubulavirus 5 Discovered as Accidental Cell Culture Contaminant. Viruses 11:777.

45. Wignall-Fleming E, Young DF, Goodbourn S, Davison AJ, Randall RE. 2016. Genome Sequence of the Parainfluenza Virus 5 Strain That Persistently Infects AGS Cells. Genome Announcements 4:e00653-16.

46. Edgar RC, Taylor J, Altman T, Barbera P, Meleshko D, Lin V, Lohr D, Novakovsky G, Al-Shayeb B, Banfield JF, Korobeynikov A, Chikhi R, Babaian A. 2020. Petabase-scale sequence alignment catalyses viral discovery. bioRxiv doi:10.1101/2020.08.07.241729:2020.08.07.241729.

47. Gibb R, Albery GF, Becker DJ, Brierley L, Connor R, Dallas TA, Eskew EA, Farrell MJ, Rasmussen AL, Ryan SJ, Sweeny A, Carlson CJ, Poisot T. 2021. Data proliferation, reconciliation, and synthesis in viral ecology. bioRxiv doi:10.1101/2021.01.14.426572:2021.01.14.426572.

48. Chen S, Zhou Y, Chen Y, Gu J. 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 34:i884-i890.

49. Kim D, Langmead B, Salzberg SL. 2015. HISAT: a fast spliced aligner with low memory requirements. Nature Methods 12:357-360.

50. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078-2079.

51. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. 2012. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. Journal of Computational Biology 19:455-477.

52. Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. 2017. metaSPAdes: a new versatile metagenomic assembler. Genome Research 27:824-834.

53. Shen W, Le S, Li Y, Hu F. 2016. SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. PLOS ONE 11:e0163962.

54. Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22:1658-1659.

55. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. BMC Bioinformatics 10:421.

56. Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. 2016. GenBank. Nucleic Acids Research 44:D67-D72.

57. Barrett T, Clark K, Gevorgyan R, Gorelenkov V, Gribov E, Karsch-Mizrachi I, Kimelman M, Pruitt KD, Resenchuk S, Tatusova T, Yaschenko E, Ostell J. 2012. BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata. Nucleic Acids Research 40:D57-D63.

58. Katoh K, Standley DM. 2013. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Molecular Biology and Evolution 30:772-780.

59. Marchler-Bauer A, Bryant SH. 2004. CD-Search: protein domain annotations on the fly. Nucleic Acids Research 32:W327-W331.

60. Lu S, Wang J, Chitsaz F, Derbyshire MK, Geer RC, Gonzales NR, Gwadz M, Hurwitz DI, Marchler GH, Song JS, Thanki N, Yamashita RA, Yang M, Zhang D, Zheng C, Lanczycki CJ, Marchler-Bauer A. 2020. CDD/SPARCLE: the conserved domain database in 2020. Nucleic Acids Research 48:D265-D268.

61. Nguyen L-T, Schmidt HA, Von Haeseler A, Minh BQ. 2015. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. Molecular Biology and Evolution 32:268-274.

62. Kalyaanamoorthy S, Minh BQ, Wong TKF, Von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. Nature Methods 14:587-589.

63. Hoang DT, Chernomor O, Von Haeseler A, Minh BQ, Vinh LS. 2018. UFBoot2: Improving the Ultrafast Bootstrap Approximation. Molecular Biology and Evolution 35:518-522.

64. Yu G, Smith DK, Zhu H, Guan Y, Lam TTY. 2017. ggtree : an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods in Ecology and Evolution 8:28-36.

65. Babin Y. 2020. Recan: Python tool for analysis of recombination events in viral genomes. Journal of Open Source Software 5:2014.

66. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29:15-21.

67. Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26:841-842.

Figure Legend

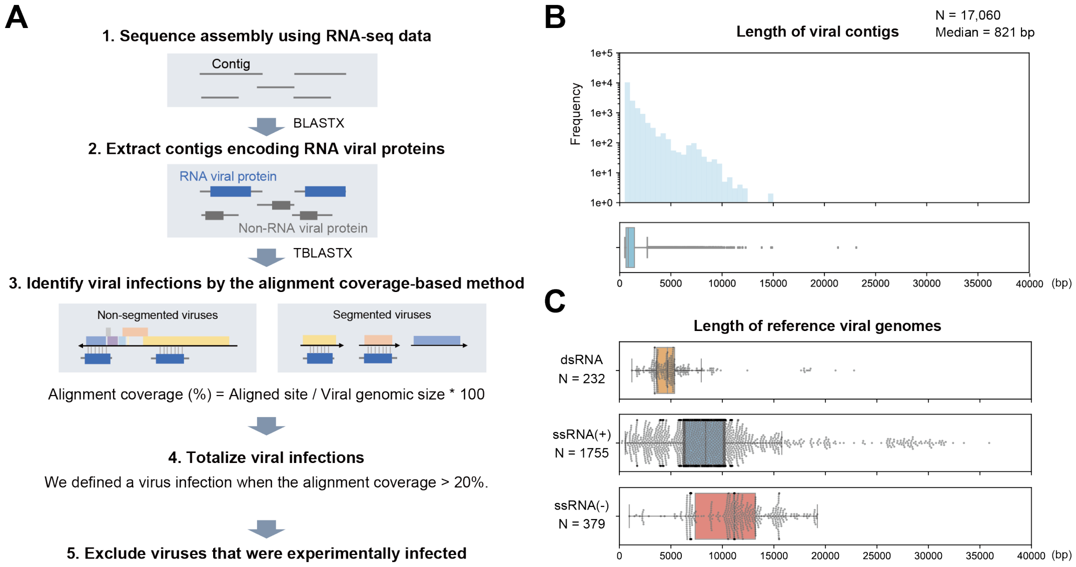
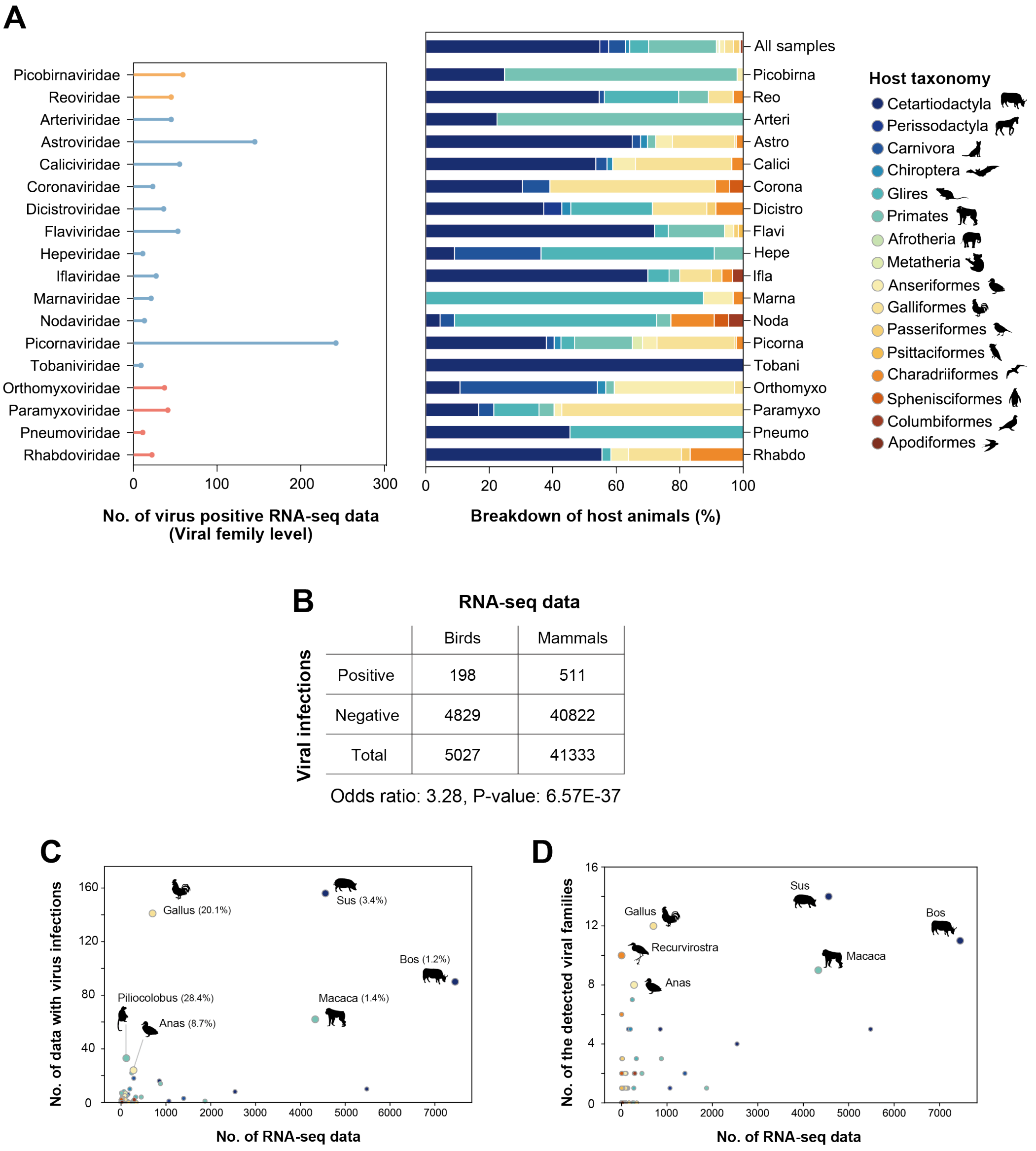

A
1. Sequence assembly using RNA-seq data
Contig
BLASTX
2. Extract contigs encoding RNA viral proteins
RNA viral protein
Non-RNA viral protein
TBLASTX
3. Identify viral infections by the alignment coverage-based method
Non-segmented viruses
Segmented viruses
Alignment coverage (%) = Aligned site / Viral genomic size * 100
4. Totalize viral infections
We defined a virus infection when the alignment coverage > 20%.
5. Exclude viruses that were experimentally infected
B
Length of viral contigs
N = 17,060
Median = 821 bp
Frequency
1e+5
1e+4
1e+3
1e+2
1e+1
1e+0
0
5000
10000
15000
20000
25000
30000
35000
40000
(bp)
C
Length of reference viral genomes
dsRNA
N = 232
ssRNA(+)
N = 1755
ssRNA(-)
N = 379
0
5000
10000
15000
20000
25000
30000
35000
40000
(bp)

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

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

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

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



	Birds	Mammals
Positive	198	511
Negative	4829	40822
Total	5027	41333

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

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

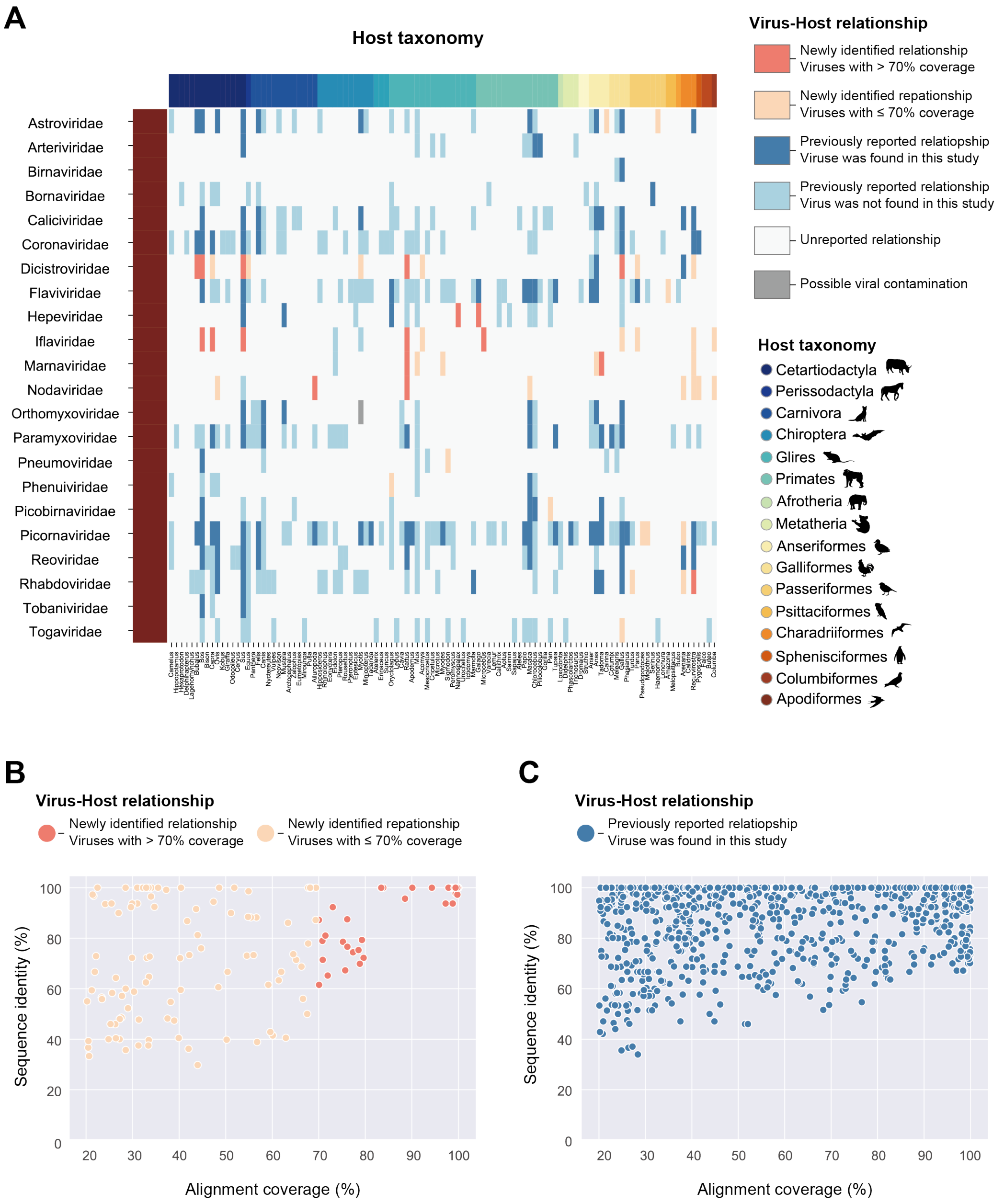


Figure 3. Search for unknown reservoir hosts and novel virus sequences.

(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:

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

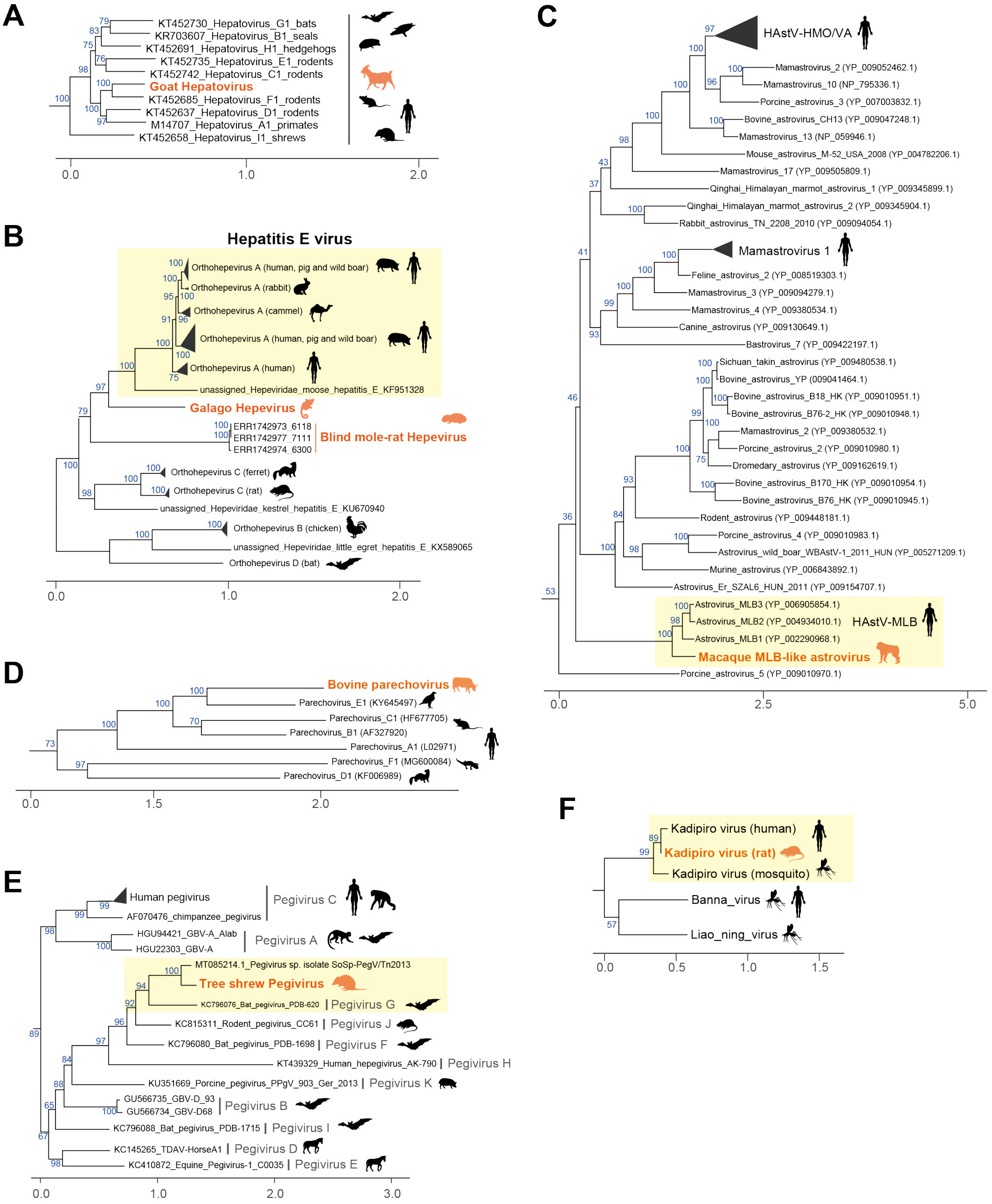
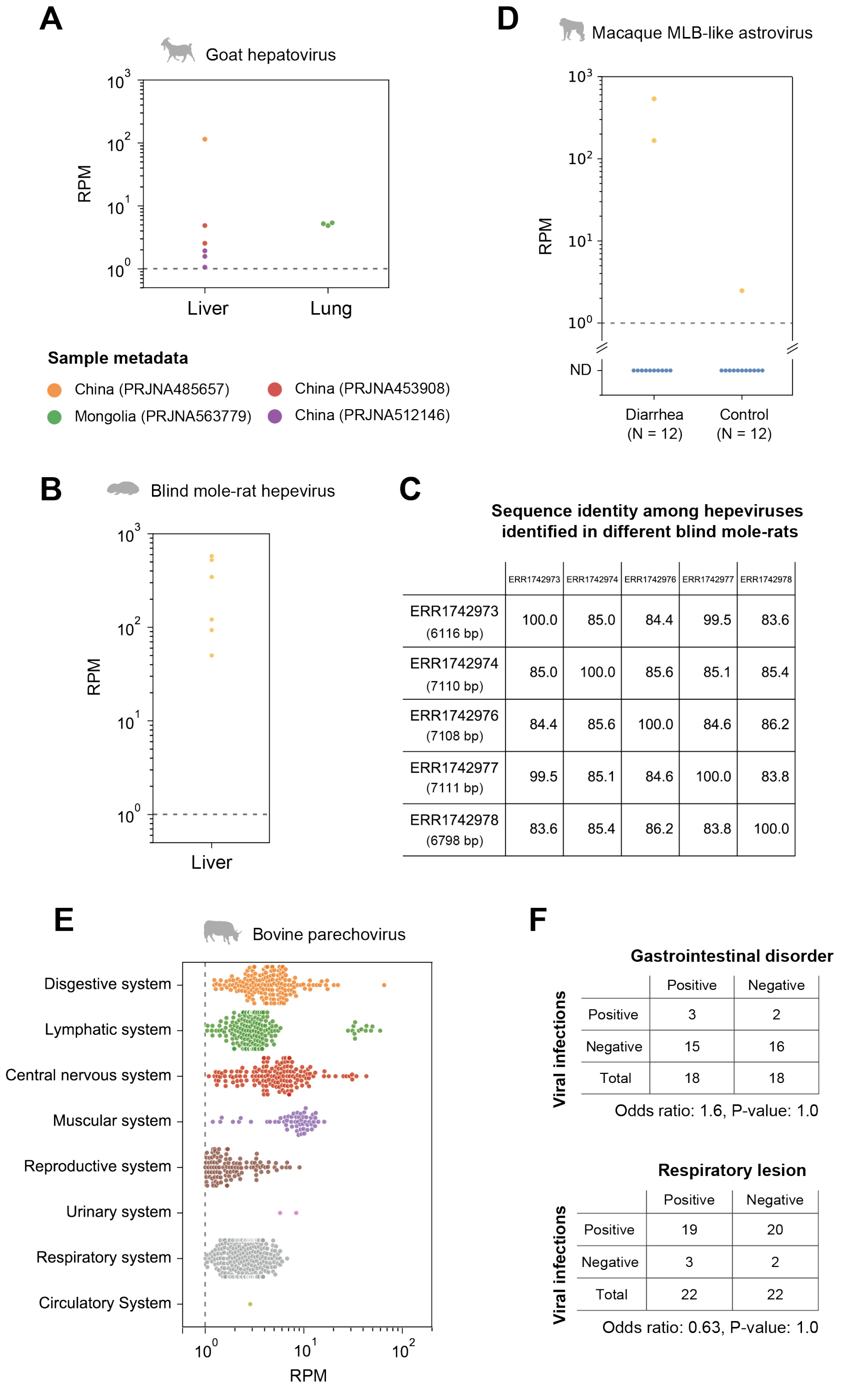

A
KT452730_Hepatovirus_G1_bats
KR703607_Hepatovirus_B1_seals
KT452691_Hepatovirus_H1_hedgehogs
KT452735_Hepatovirus_E1_rodents
KT452742_Hepatovirus_C1_rodents
Goat Hepatovirus
KT452685_Hepatovirus_F1_rodents
KT452637_Hepatovirus_D1_rodents
M14707_Hepatovirus_A1_primates
KT452658_Hepatovirus_I1_shrews
0.0
1.0
2.0
B
Hepatitis E virus
Orthohepevirus A (human, pig and wild boar)
Orthohepevirus A (rabbit)
Orthohepevirus A (cammel)
Orthohepevirus A (human, pig and wild boar)
Orthohepevirus A (human)
unassigned_Hepeviridae_moose_hepatitis_E_KF951328
Galago Hepevirus
ERR1742973_6118
ERR1742977_7111
ERR1742974_6300
Blind mole-rat Hepevirus
Orthohepevirus C (ferret)
Orthohepevirus C (rat)
unassigned_Hepeviridae_kestrel_hepatitis_E_KU670940
Orthohepevirus B (chicken)
unassigned_Hepeviridae_little_egret_hepatitis_E_KX589065
Orthohepevirus D (bat)
0.0
1.0
2.0
C
HAstV-HMO/VA
Mamastrovirus_2 (YP_009052462.1)
Mamastrovirus_10 (NP_795336.1)
Porcine_astrovirus_3 (YP_007003832.1)
Bovine_astrovirus_CH13 (YP_009047248.1)
Mamastrovirus_13 (NP_059946.1)
Mouse_astrovirus_M-52_USA_2008 (YP_004782206.1)
Mamastrovirus_17 (YP_009505809.1)
Qinghai_Himalayan_marmot_astrovirus_1 (YP_009345899.1)
Qinghai_Himalayan_marmot_astrovirus_2 (YP_009345904.1)
Rabbit_astrovirus_TN_2208_2010 (YP_009094054.1)
Mamastrovirus 1
Feline_astrovirus_2 (YP_008519303.1)
Mamastrovirus_3 (YP_009094279.1)
Mamastrovirus_4 (YP_009380534.1)
Canine_astrovirus (YP_009130649.1)
Bastrovirus_7 (YP_009422197.1)
Sichuan_takin_astrovirus (YP_009480538.1)
Bovine_astrovirus_YP (009041464.1)
Bovine_astrovirus_B18_HK (YP_009010951.1)
Bovine_astrovirus_B76-2_HK (YP_009010948.1)
Mamastrovirus_2 (YP_009380532.1)
Porcine_astrovirus_2 (YP_009010980.1)
Dromedary_astrovirus (YP_009162619.1)
Bovine_astrovirus_B170_HK (YP_009010954.1)
Bovine_astrovirus_B76_HK (YP_009010945.1)
Rodent_astrovirus (YP_009448181.1)
Porcine_astrovirus_4 (YP_009010983.1)
Astrovirus_wild_boar_WBAstV-1_2011_HUN (YP_005271209.1)
Murine_astrovirus (YP_006843892.1)
Astrovirus_Er_SZAL6_HUN_2011 (YP_009154707.1)
Astrovirus_MLB3 (YP_006905854.1)
Astrovirus_MLB2 (YP_004934010.1)
Astrovirus_MLB1 (YP_002290968.1)
HAstV-MLB
Macaque MLB-like astrovirus
Porcine_astrovirus_5 (YP_009010970.1)
0.0
2.5
5.0
D
Bovine parechovirus
Parechovirus_E1 (KY645497)
Parechovirus_C1 (HF677705)
Parechovirus_B1 (AF327920)
Parechovirus_A1 (L02971)
Parechovirus_F1 (MG600084)
Parechovirus_D1 (KF006989)
0.0
1.5
2.0
E
Human pegivirus
AF070476_chimpanzee_pegivirus
Pegivirus C
HGU94421_GBV-A_Alab
HGU22303_GBV-A
Pegivirus A
MT085214.1_Pegivirus sp. isolate SoSp-PegV/Tn2013
Tree shrew Pegivirus
KC796076_Bat_pegivirus_PDB-620
Pegivirus G
KC815311_Rodent_pegivirus_CC61
Pegivirus J
KC796080_Bat_pegivirus_PDB-1698
Pegivirus F
KT439329_Human_hepegivirus_AK-790
Pegivirus H
KU351669_Porcine_pegivirus_PPgV_903_Ger_2013
Pegivirus K
GU566735_GBV-D_93
GU566734_GBV-D68
Pegivirus B
KC796088_Bat_pegivirus_PDB-1715
Pegivirus I
KC145265_TDAV-HorseA1
Pegivirus D
KC410872_Equine_Pegivirus-1_C0035
Pegivirus E
0.0
1.0
2.0
3.0
F
Kadipiro virus (human)
Kadipiro virus (rat)
Kadipiro virus (mosquito)
Banna_virus
Liao_ning_virus
0.0
0.5
1.0
1.5

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



Sequence identity among hepeviruses identified in different blind mole-rats

	ERR1742973	ERR1742974	ERR1742976	ERR1742977	ERR1742978
ERR1742973 (6116 bp)	100.0	85.0	84.4	99.5	83.6
ERR1742974 (7110 bp)	85.0	100.0	85.6	85.1	85.4
ERR1742976 (7108 bp)	84.4	85.6	100.0	84.6	86.2
ERR1742977 (7111 bp)	99.5	85.1	84.6	100.0	83.8
ERR1742978 (6798 bp)	83.6	85.4	86.2	83.8	100.0

Gastrointestinal disorder

	Positive	Negative
Positive	3	2
Negative	15	16
Total	18	18

Odds ratio: 1.6, P-value: 1.0

Respiratory lesion

	Positive	Negative
Positive	19	20
Negative	3	2
Total	22	22

Odds ratio: 0.63, P-value: 1.0

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 or organ system. The gray dotted line indicates the criterion used to determine viral infections (RPM: 1.0). The lower panel in (A) represents the sample metadata.

(C) Comparison of nucleotide sequence identity among the hepeviral sequences identified in five different blind mole-rats. The numbers in parentheses in the row indicate the total number of aligned sites between the viral contigs identified in each individual and the blind mole-rat hepevirus identified in ERR1742977.

(D) Quantification of the macaque MLB-like viral infection levels in the patient with diarrhea and control macaque monkeys. The x-axis indicates the diagnosis for the 24 monkeys, and the y-axis indicates the RPM. The average RPM for each individual is plotted because six samples were collected from each individual. The dotted line indicates the criterion used for detecting viral infections (RPM: 1.0). We considered samples with RPMs below the criterion as non-detectable (ND).

(F) Association between the parechovirus infections and symptoms. The tables show the number of RNA-seq data with and without the parechovirus infections in two independent studies, which provide diagnostic information: gastrointestinal disorder (upper panel) and respiratory lesion (lower panel).

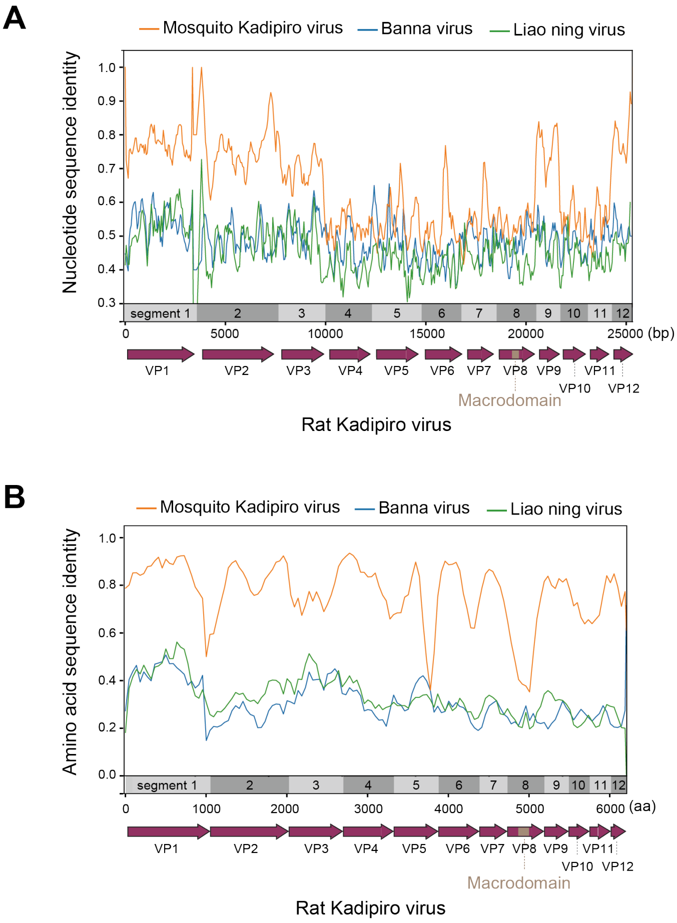
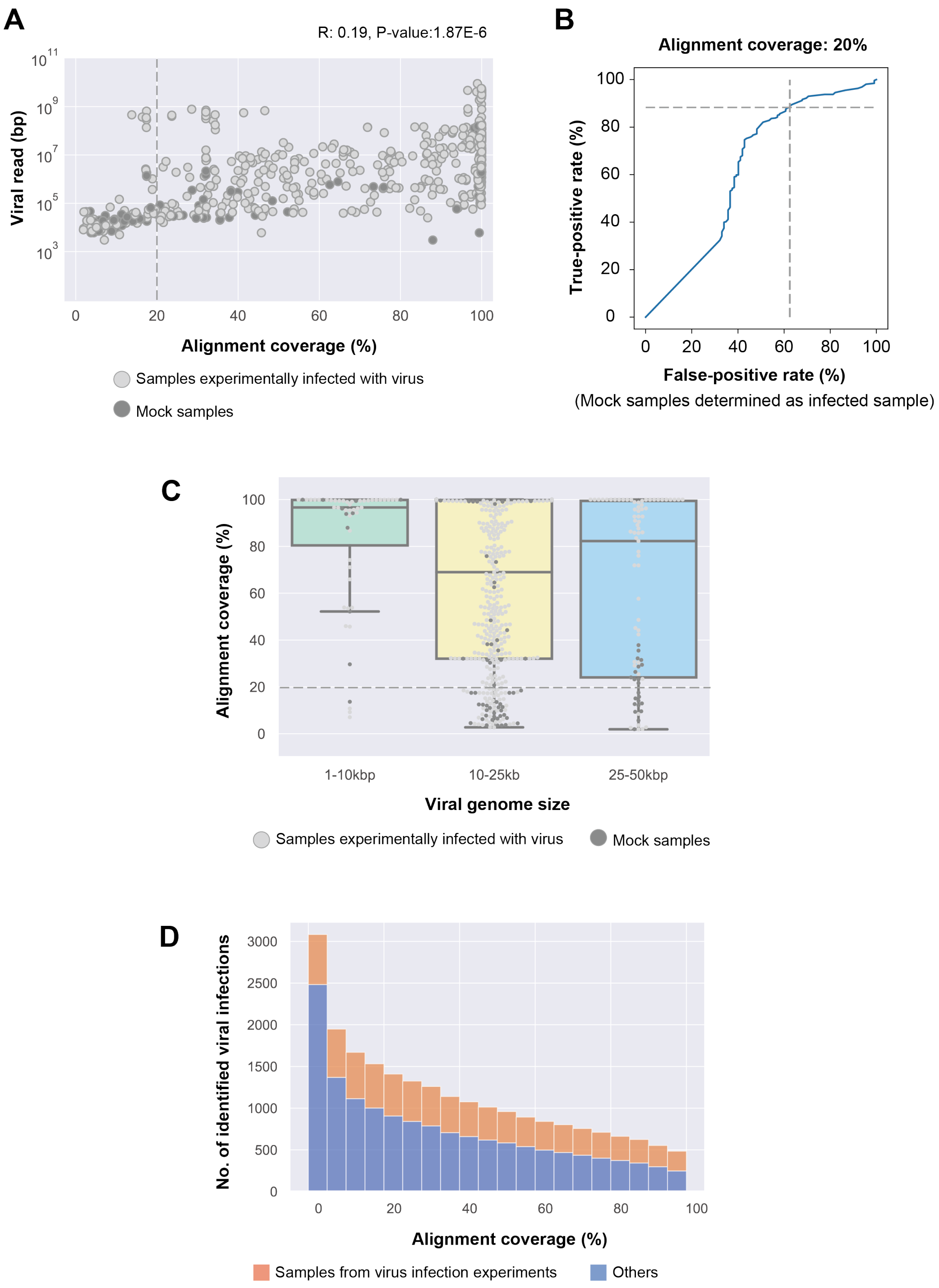


Figure 6. Sequence identity plots between rat Kadipiro virus and other known seadornaviruses.

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.

Supplemental Materials



Supplementary Figure 1

Supplemental Figure 1. Validation of the alignment coverage-based method for detecting viral infections using samples obtained from viral infection experiments.

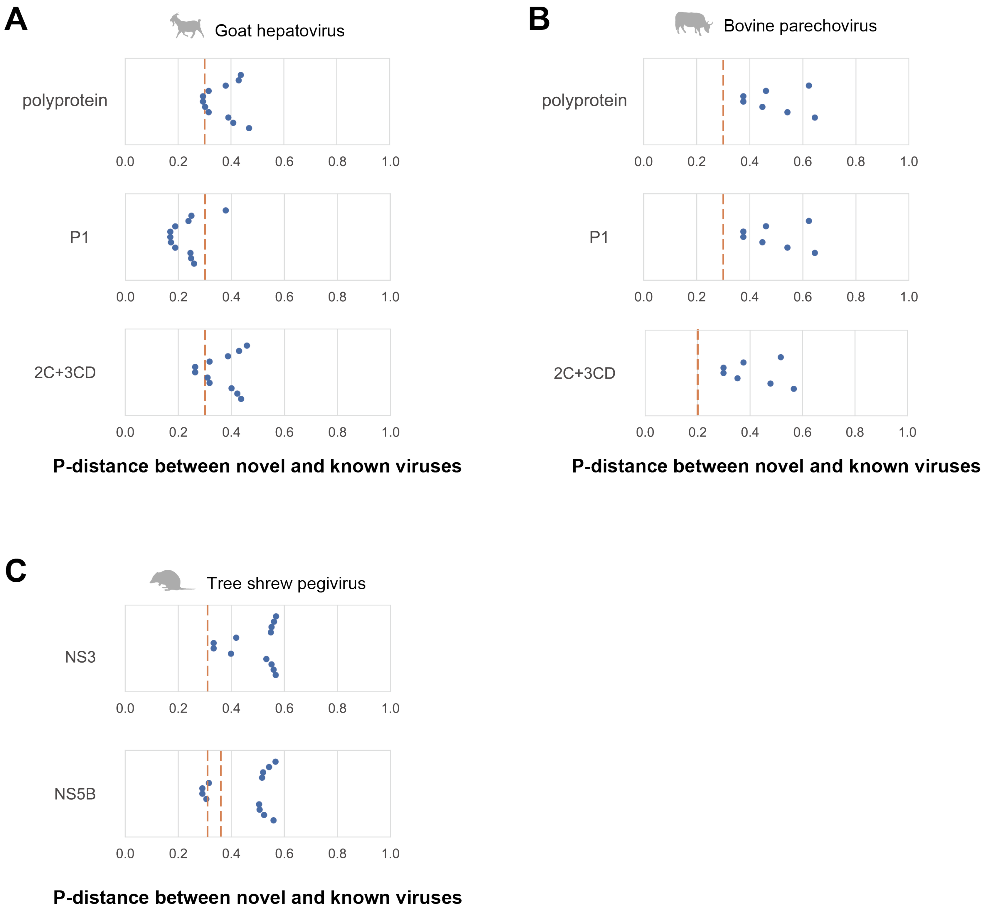
(A) Comparison between the alignment coverage-based method and the viral read-based method using samples obtained from viral infection experiments. The x-axis indicates alignment coverage between viral contigs in each RNA-seq data and the reference viral genome used for the experiments. The y-axis indicates the total read length of the virus family used for the experiment, which was obtained from the NCBI SRA Taxonomy Analysis Tool. Light gray dots indicate samples experimentally infected with viruses, and dark gray dots indicate mock samples. R: Pearson's correlation coefficient. Dotted line indicates 20% alignment coverage.

(B) Changes in the true-positive and the false-positive rates depending on the criteria to determine viral infections. The true-positive rate (y-axis) indicates the number of samples experimentally infected with viruses correctly determined as the infected sample, and the false-positive rate (x-axis) indicates the number of mock samples determined as the infected sample. Dotted line indicates the true-positive rate (88.3%) and the false-positive rate (62.5%) when 20% alignment coverage was used as the criterion (**details in Materials and Methods**).

(C) Detection rate of viral infections depending on the viral genome size. Box plots show the distributions of alignment coverage of the viral genome with 1-10kbp (green), 10-25kbp (yellow), and 25-50kbp (blue). Light gray dots indicate samples infected with viruses experimentally, and dark gray dots indicate mock samples. Dotted line indicates 20% alignment coverage.

(D) The number of detected viral infections depending on the alignment coverage criteria. 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.

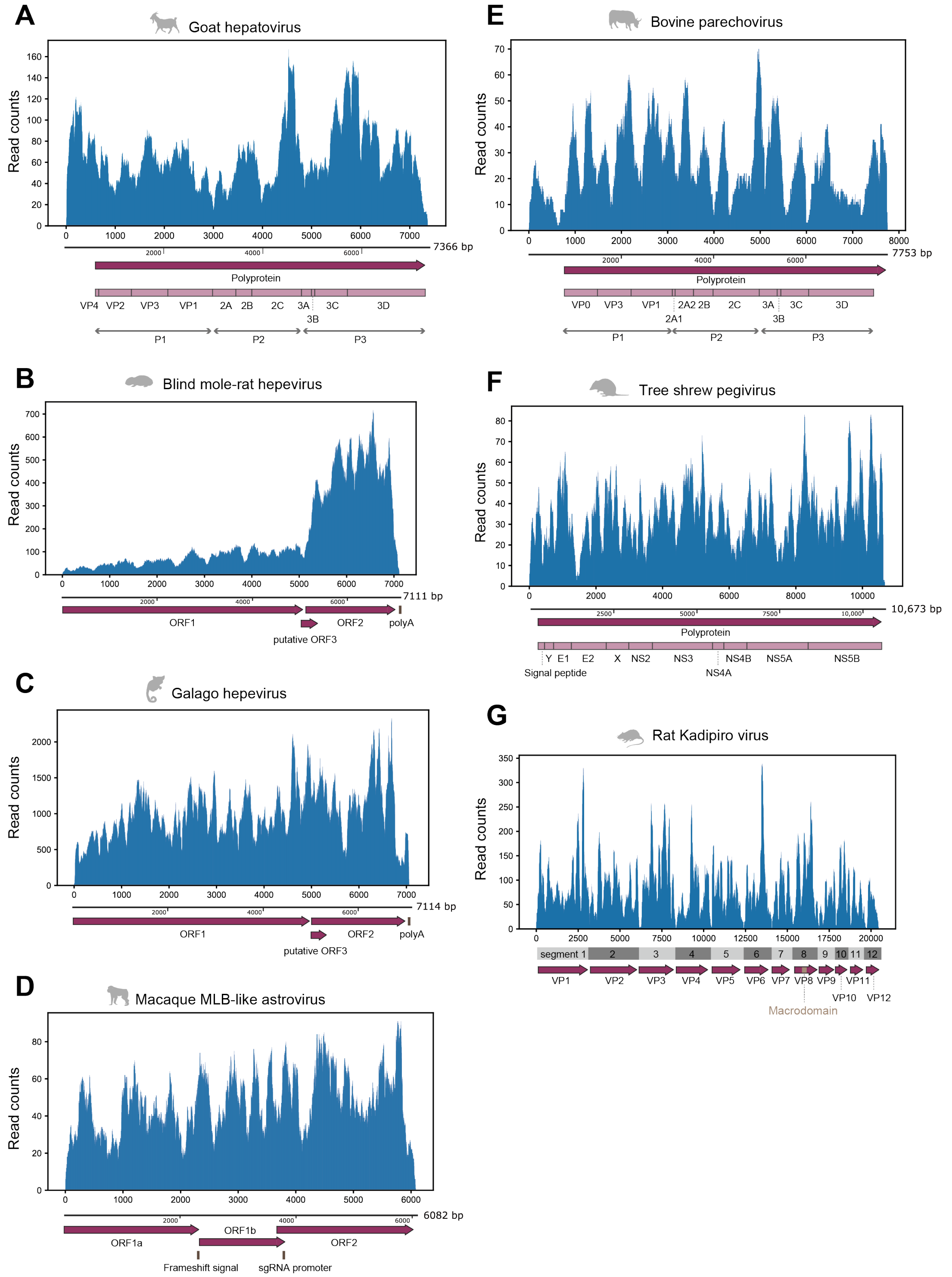


Supplementary Figure 2

972

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

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



Supplementary Figure 3

983

Supplemental Figure 3. Read mapping analysis using RNA-seq data in which the viral sequence was identified.

(A-G) Read distributions mapped to the viral sequence: goat hepatovirus (A), blind mole-rat hepevirus (B), galago hepevirus (C), macaque MLB-like astrovirus (D), bovine parechovirus (E), tree shrew pegivirus (F), and rat Kadipiro virus (G). The upper panel shows the virus genomic positions (x-axis) and read counts at each position (y-axis). The lower panel shows genomic annotations, such as protein-coding regions or signal sequences. Dark purple arrows indicate open reading frames (ORFs) in the viral genome. Light purple boxes show mature proteins predicted based on aligned positions with reference viruses (**details in Materials and Methods**). Brown vertical lines indicate nucleotide sequence features, such as polyadenylation signal (poly-A), ribosomal frameshift signal (frameshift signal), and promoter sequence for subgenomic RNA synthesis (sgRNA promoter). Light and dark gray boxes indicate the segments of rat Kadipiro virus. Segment 8 of rat Kadipiro virus was expected to encode chimeric VP8, containing a macrodomain, shown as a brown box in the dark purple arrow.

Supplemental Dataset 1. List of Sequence Read Archive run accession numbers, genome file, and sequence assembly method.

Supplemental Dataset 2. Information on RNA-seq data from experimental infection with viruses.

Supplemental Dataset 3. Information on possible viral contamination excluded from the totalization.

Supplemental Dataset 4. Information on manual curation for virus-host relationships.

Supplemental Dataset 5. Accession numbers of viral sequences used for phylogenetic analyses, viral genomic annotations, and comparing the International Committee on Taxonomy of Viruses species demarcation criteria.

Supplemental Dataset 6. Information on concatenated seadornaviral sequences.

Supplemental Dataset 7. Sequence Read Archive run accessions used for mapping analyses.

Supplemental Dataset 8. Sample metadata in which the goat hepatoviral infections were detected.

Supplemental Dataset 9. Sample metadata in which the blind mole-rat hepeviral infections were detected.

Supplemental Dataset 10. Sample metadata in which the galago hepeviral infections were detected.

Supplemental Dataset 11. Sample metadata in which the macaque MLB-like astrovirus infections were detected.

Supplemental Dataset 12. Sample metadata in which the bovine parechovirus infections were detected.

Supplemental Dataset 13. Sample metadata in which the tree shrew pegiviral infections were detected.

Supplemental Dataset 14. Bioinformatics tools and their versions used in this study.