Virus Evolution, 2019, 5(1): vez013

doi: 10.1093/ve/vez013
Research article

Gut virome of mammals and birds reveals high genetic diversity of the family *Microviridae*

Hao Wang,[1,†] Yu Ling,[1,†] Tongling Shan,[2,†] Shixing Yang,[1,†] Hui Xu,[3,†] Xutao Deng,[4] Eric Delwart,[4,5] and Wen Zhang[1,*,‡]

[1]Department of Microbiology, School of Medicine, Jiangsu University, 310 Xuefu Road, Zhenjiang, Jiangsu 212013, China, [2]Department of Swine Infectious Disease, Shanghai Veterinary Research Institute, Chinese Academy of Agricultural Sciences, 518 Ziyue Road, Shanghai 200241, China, [3]The Affiliated Hospital of Jiangsu University, 438 Jiefang Road, Zhenjiang, Jiangsu 212001, China, [4]Vitalant Research Institute, 270 masonic avenue, San Francisco, CA 94118, USA and [5]Department of Laboratory Medicine, University of California, 270 masonic avenue, San Francisco, San Francisco CA 94118, USA

[†]These authors contribute equally to this work.

[‡]http://orcid.org/0000-0002-9352-6153
*Corresponding author: E-mail: z0216wen@yahoo.com

Abstract

Nineteen families of phages infecting bacteria or archaea are currently recognized by the International Committee on Taxonomy of Viruses (ICTV). Of these, only two have single-stranded DNA genomes, namely *Inoviridae* and *Microviridae*. The distribution, genetic characteristics, and ecological roles of *Microviridae* remain largely under explored. Here, using viral metagenomics, we investigate the intestinal virome from human and twenty-four species of animals, as well as freshwater samples, containing abundant sequence reads showing similarity to the *Microviridae*. Eight hundred and sixty complete or near complete *Microviridae*-related genomes were generated, showing high levels of co-infections and sequence divergence. Sequence comparison and phylogenetic analysis showed that the *Microviridae* subfamily *Gokushovirinae* was highly prevalent and that some strains may qualify as new subfamilies. This study significantly augments our knowledge of the genetic diversity, genome evolution, and distribution in animal species of members of the family *Microviridae*.

Key words: *Microviridae*; complete genome; phylogenetic analysis; genetic diversity.

1. Introduction

Bacteriophages can influence microbial abundance, composition of microbial communities, and even biogeochemical cycling by interacting with their cellular hosts (Morella et al. 2018). A great deal of research has been performed with phages with double-stranded DNA (dsDNA) genomes (Salmond and Fineran 2015). Recent studies have revealed that members of the *Microviridae* family were found in water and the content of animal and human guts (Desnues et al. 2008; Tucker et al. 2011; Labonte and Suttle 2013; Hopkins et al. 2014; Bryson et al. 2015; Quaiser et al. 2015; Doore and Fane 2016; Guo et al. 2017; Barrientos-Somarribas et al. 2018; Xie et al. 2019), suggesting possible roles in affecting different environments. The properties of some members of the *Microviridae* family have been well elucidated particularly in virion structure and assembly (McKenna et al. 1992; Doore and Fane 2016; Blackburn et al. 2017). This family is currently comprised of two International

Committee on Taxonomy of Viruses (ICTV)-approved subfamilies: *Bullavirinae* and *Gokushovirinae*, and two tentative groups: *Pichovirinae* and *Alpavirinae*, which have not been formally accepted by ICTV (Roux et al. 2012; Adams et al. 2016). The subfamily *Bullavirinae* includes three genera, *Alpha3microvirus*, *G4microvirus*, and *Phix174microvirus*, which are known to infect enterobacteria and have been extensively studied through the archetype of this family, the bacteriophage Phix174, while the *Gokushovirinae* subfamily is commonly referred to as 'parasites of parasites', which infect obligate intracellular parasites within the genera *Chlamydia*, *Bdellovibrio*, and *Spiroplasma* (Brentlinger et al. 2002).

Mammalian bodies support a diverse bacterial community residing in the lower gastrointestinal tract which itself host diverse phages. Using viral metagenomics, we investigated the virome in fecal samples from mammals and birds which revealed a great deal of divergent *Microviridae* sequences. The estimated abundance of these bacterial viruses in different animal species, genome organization, sequence similarity based on VP1 protein which is a classic phylogenetic marker for the classification of *Microviridae* (Hopkins et al. 2014; Quaiser et al. 2015), and phylogenetic diversity are analyzed here.

2. Materials and methods

2.1 Samples and preparation

During 2014–6, 1,795 samples, including 1,546 fecal samples from human and 24 species of animal, 80 cattle (*Bos taurus*) blood, 20 wild rat oral secretion, 27 cattle nasal secretion, 20 nasal secretion from human, 80 cattle genital secretions, 20 wild rat (*Rattus norvegicus*) skin swab, and 2 concentrated fresh water samples were collected from 13 different provinces in China (Fig.1a and Supplementary Table S1). All samples were collected by disposable materials and shipped on dry ice. These samples were collected as part of a Virome Project of China supported by the Ministry of Science and Technology of China aimed at the discovery of emerging viruses in mammals and birds. The samples obtained from humans were collected following informed written consent from each person.

Fecal samples were re-suspended in ten volumes of phosphate-buffered saline (PBS) and vigorously vortexed for 5 min. Fecal supernatants were then collected after centrifugation (10 min, 15,000 × *g*). The tips of swabs were immersed into 1 ml PBS and vigorously vortexed for 5 min and incubated for 30 min in 4 °C. The supernatants were then collected after centrifugation (10 min, 15,000 × *g*). Oyster (*Ostrea gigas*) digestive tissue samples were homogenized, frozen, and thawed three times on dry ice, the supernatants were then collected after centrifugation (10 min, 15,000 × *g*). The whole blood samples were centrifuged (10 min, 15,000 × *g*) for collection of plasma. Five hundred microliters of each supernatant was filtered through a 0.45-μm filter (Millipore) to remove eukaryotic and bacterial cell-sized particles. The filtrates enriched in viral particles were treated with DNase and RNase to digest unprotected nucleic acid at 37 °C for 60 min (Zhang et al. 2014, 2016; Liu et al. 2016).

2.2 Viral metagenomic analysis

Remaining total nucleic acid was then isolated using QIAamp MinElute Virus Spin Kit (Qiagen) according to manufacturer's protocol. The enriched viral nucleic acid preparations of the respective pools or samples were individually subjected to reverse transcription reactions using reverse transcriptase (Super-Script III, Invitrogen) and 100 pmol of random hexamer primer, followed by a single round of DNA synthesis using Klenow fragment polymerase (New England BioLabs), where reverse transcription was included because of the original purpose of detecting DNA as well as RNA viruses.

Overall, 263 libraries were then constructed using Nextera XT DNA Sample Preparation Kit (Illumina) and sequenced using the MiSeq Illumina platform with 250 bases paired ends with dual barcoding for each individual sample or sample pool. The information about each library was shown in Supplementary Table S1. Sample collection and all experiments in the present study were performed with Ethical Approval given by Ethics Committee of Jiangsu University and the reference number is No. UJS2015022.

For bioinformatics analysis, paired-end reads of 250 bp generated by MiSeq were debarcoded using vendor software from Illumina. An in-house analysis pipeline running on a thirty-two-nodes Linux cluster was used to process the data. Reads were considered duplicates if bases 5–55 were identical and only one random copy of duplicates was kept. Clonal reads were removed and low sequencing quality tails were trimmed using Phred quality score ten as the threshold. The unique reads number of each library was shown in Supplementary Table S1. Adaptors were trimmed using the default parameters of VecScreen which is NCBI BLASTn with specialized parameters designed for adapter removal. The cleaned reads were *de-novo* assembled within each barcode using the ENSEMBLE assembler (Deng et al. 2015). Contigs and unassembled reads are then matched against a customized viral proteome database using BLASTx with an *E*-value cutoff of $<10^{-5}$, where the virus BLASTx database was compiled using NCBI virus reference proteome (ftp://ftp.ncbi.nih.gov/refseq/release/viral/) to which was added viral proteins sequences from NCBI nr fasta file (based on annotation taxonomy in Virus Kingdom). Candidate viral hits are then compared to an in-house non-virus non-redundant (NVNR) protein database to remove false positive viral hits, where the NVNR database was compiled using non-viral protein sequences extracted from NCBI nr fasta file (based on annotation taxonomy excluding Virus Kingdom). Contigs without significant BLASTx similarity to viral proteome database are searched against viral protein families in vFam database (Skewes-Cox et al. 2014) using HMMER3 (Eddy 2009; Johnson, Eddy, and Portugaly 2010; Finn, Clements, and Eddy 2011) to detect remote viral protein similarities.

2.3 Acquisition of *Microviridae* genomes

For assembly of the *Microviridae* genomes, the contigs which showed significant BLASTx similarity to *Microviridae* were selected to check whether they are circular genomes in Geneious software version 11.0. Only those genomes that showed overlapping reads at the start and end of the contigs, confirming their circular genomes were retained. The contigs with sequence length >3,000 bp but without circular genome were subjected to further analysis where the individual contig was used as reference for mapping to the raw data of its original barcode using the Low Sensitivity/Fastest parameter in Geneious software version 11.0. The prolonged contigs were manually rechecked for determining whether they were circular genomes. Those contigs with sequence length >3,500 bp and the major capsid protein VP1 and the replication protein VP4 were considered to be nearly complete genomes and included in this study. The contigs which had sequence length >7,000 bp but have no putative protein VP1 or VP4 were considered to be non-*Microviridae* sequences and not included in this study.

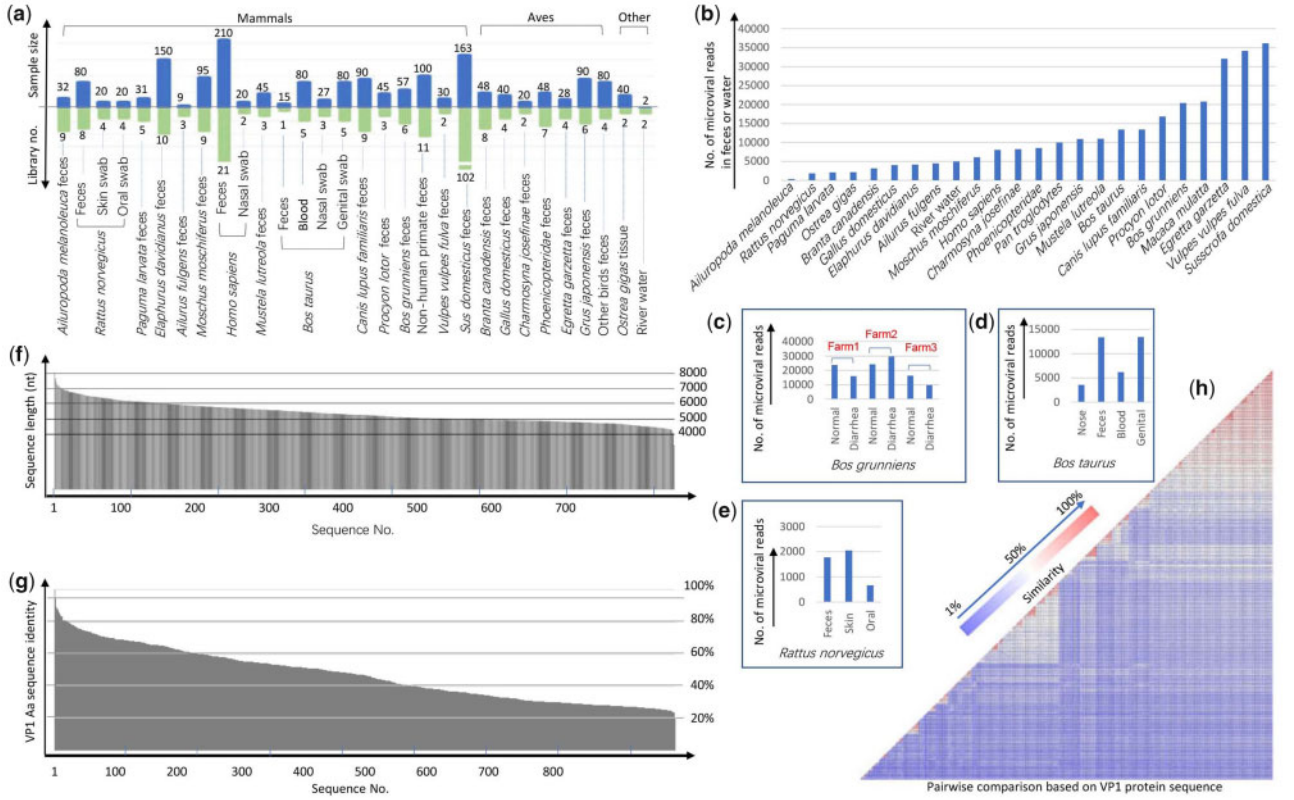

Figure 1. Identification of viral sequences showing similarity to *Microviridae* from virome in human, animals, and other samples. (a) Samples size and library numbers. On top of each bar in the upper graph the sample number is shown. On bottom of each bar in the lower graph the library number is shown. The species name of animal is shown on the bottom. Those samples without marked sample types all belong to fecal samples. (b) The abundance of SSM in fecal sample of different animals and water sample. The number of SSM is normalized to reads per million unique reads. (c) The abundance of SSM in intestinal contents from yaks with different health status in three different farms. Health status normal and diarrheic is shown on the bottom of the bars. On top of each bar the farm numbers are shown. The number of SSM is normalized to reads per million unique reads. (d) The abundance of SSM in different types of samples from cattle. On bottom of each bar the sample types are shown. The number of SSM was normalized to reads per million unique reads. (e) The abundance of SSM in different types of samples from wild rats. On bottom of each bar the sample types are shown. The number of SSM is normalized to reads per million unique reads. (f) Sequence length distribution of the 713 complete *Microviridae* genomes identified in this study. The horizontal axis indicates the sequence numbers and the vertical axis shows the sequence length. (g) Sequence identity distribution based on VP1 amino acid sequence comparison between the 860 *Microviridae* strains in this study and their best matches in BLASTp search, respectively. The horizontal axis indicates the sequence number and the vertical axis shows the amino acid sequence identity. (h) Pairwise sequence comparison of the 860 amino acid sequences of VP1 identified in this study. The percent identities are shown by heat map, where plot colors and color saturation reflect the identity, ranging from low (blue) to high (red).

2.4 Phylogenetic analysis

Phylogenetic analyses were performed based on the predicted amino acid of the major capsid protein VP1 of *Microviridae* and eighteen representative members of the previously confirmed subfamilies, *Bullavirinae* and *Gokushovirinae*, and the other six proposal subfamilies. Sequence alignment was performed using CLUSTAL W with the default settings. Phylogenetic trees with 1,000 bootstrap resamples of the alignment data sets were generated using the Maximum-likelihood method in MEGA7.0. For sequence pairwise comparison, the alignment data were imported into and analyzed by CLC Genomics Workbench 10.0, where the identities between sequences and the corresponding heat maps were generated. Potential viral open reading frames (ORFs) in *Microviridae* genomes were predicted by combining Geneious 8.1 software and NCBI ORF finder. The annotations of these ORFs were mainly based on comparisons to the Conserved Domain Database using RPS-BLAST with an *E*-value cutoff of $<10^{-5}$. For the coding protein sequences from ORFs which had no significant similarity found in Database were annotated as putative proteins.

2.5 Quality control

In order to exclude the possibility of contamination with viral nucleic acids present in the laboratory or from Qiagen nucleic acid extraction kits, twenty samples or pools positive for *Microviridae* whose best matches are related to marine gokushovirus or Eel River basin pequenovirus (Bryson et al. 2015; Creasy et al. 2018) were randomly selected and the nucleic acid were re-extracted using Trizol reagent (Invitrogen). PCR using primers specific to those 20 sequences confirmed their presence in the original biological samples.

3. Results

3.1 Sequence reads of *Microviridae* in different samples

The current study consisted of 1,796 samples, including intestinal content, blood, nasal secretion, genital secretion, and skin swab from human and 24 animal species plus two fresh water samples. All of these samples were pooled or individually used for library constructing and viral metagenomic analysis.

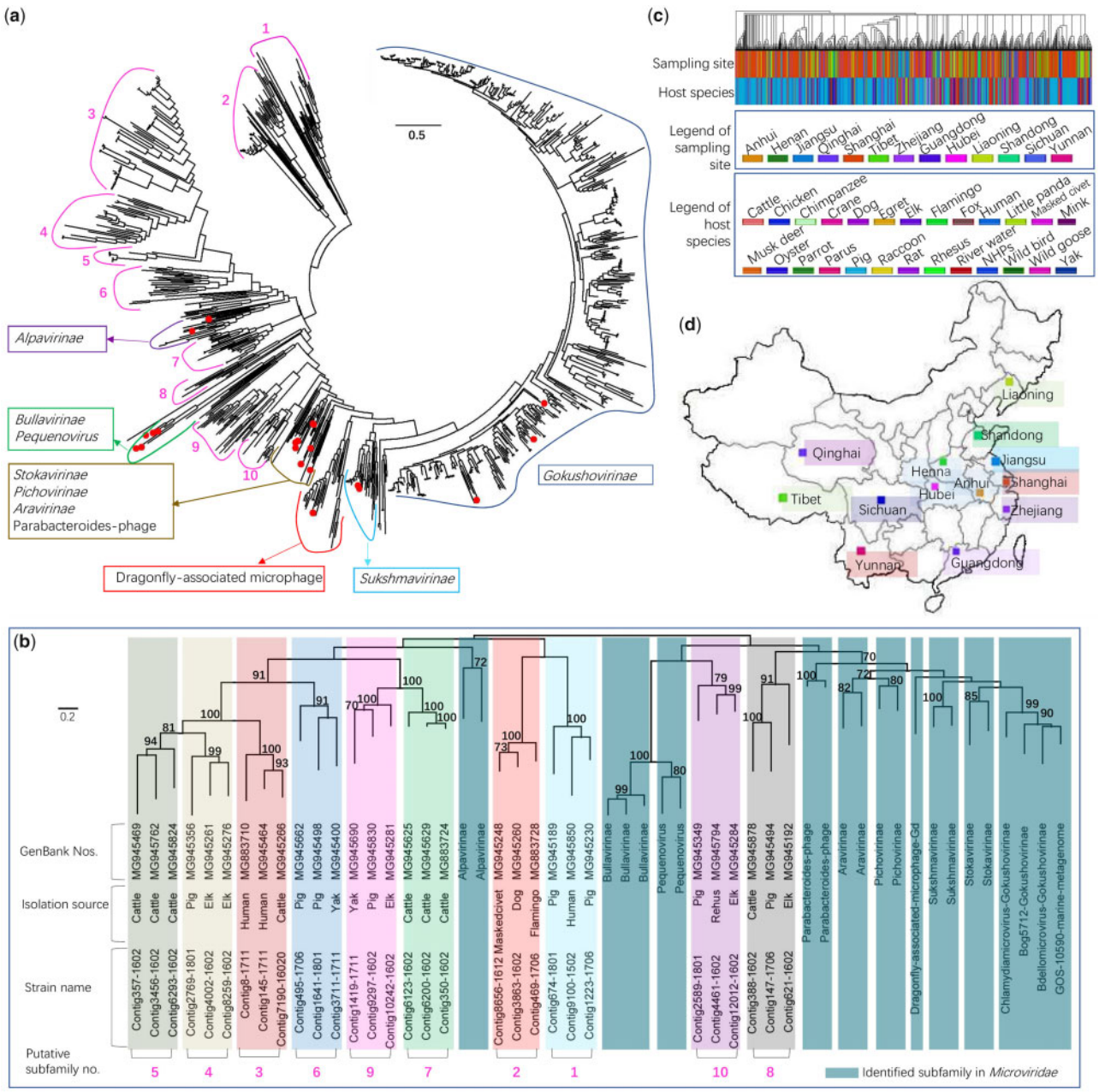

Figure 2. Phylogenetic diversity of *Microviridae*. (a) Phylogenetic tree based on VP1 amino acid sequence. The sequences in this analysis include VP1 amino acid sequences of the 860 *Microviridae* strains and 22 representative members of the previously identified subfamilies. Within the tree, the twenty-two representative members of the previously confirmed subfamilies are shown with red dots. The names of the subfamily are shown beside the corresponding clades. The potential new clades are shown with rose lines and marked with numbers 1–10. The scale bar indicates amino acid substitutions per site. (b) Phylogenetic tree based on VP1 amino acid sequences of the representative strain of the potential new clades and the previously identified subfamilies in *Microviridae*. The previously identified subfamilies and the potential new subfamilies are shaded different colors. GenBank accession numbers, strain names, and isolation sources of each *Microviridae* identified in this study are shown on tree. (c) Phylogenetic analysis based VP1 amino acid sequences the reflecting the correlation between genetic relationship and isolation hosts and geographic location of the 860 *Microviridae* sequence identified in this study. Isolation source and geographic information of each sequences are shown with different colors, respectively. (d) Sampling locations included thirteen different provinces which are shown in different colors in map of China.

Overall, 263 nucleic acid libraries from these samples were generated and sequenced in 8 Illumina MiSeq runs. Sample and library information are shown in Fig.1a and Supplementary Table S1. The viromes corresponding to the 263 libraries/barcodes were then generated, where sequences whose translated amino acid sequences were similar to those proteins of members of the *Microviridae* family were selected for further analysis. The total numbers of sequence reads showing similarity to *Microviridae* (abbreviated as SSM) in each library are listed in Supplementary Table S1. To compare the abundance of SSM in different species of animal or sample type, the number of SSM was normalized to reads per million unique reads (Fig. 1b–e and Supplementary Table S1). As for the fecal samples, giant panda (*Ailuropoda melanoleuca*) showed the lowest abundance of SSM, accounting for 0.032 per cent of the total unique sequence reads, while pig feces showed the highest percentage (i.e. 3.61%)

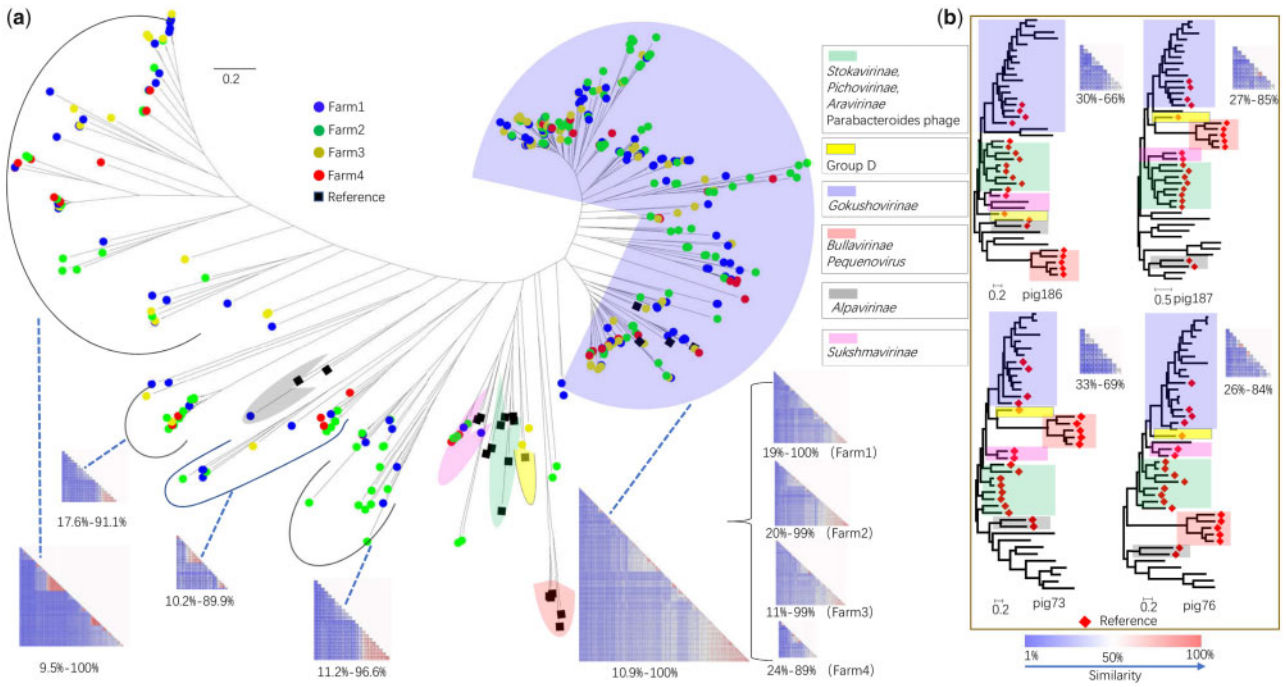

Figure 3. Phylogenetic analysis of *Microviridae* from virome in pigs. (a) Phylogenetic tree based on VP1 amino acid sequence of the 410 *Microviridae* strains from pigs located in 4 different sampling pig farms and 18 representative members of the previously identified subfamilies. Within the tree, the *Microviridae* strains from four different farms are shown with four different color dots, respectively. The twenty-two representative members of the previously confirmed subfamilies in *Microviridae* are marked with black diamonds and the clades in which they are located are shaded with different colors. The subfamily names are shown in text box beside their corresponding shaded colors. The scale bar indicates 0.5 amino acid substitutions per site. Pairwise sequence comparison of the VP1 amino acid sequences of *Microviridae* identified in this study in each clade within this tree was performed. The percent identities are shown beside each clade by heat map, where plot colors and color saturation reflect the identity, ranging from low (blue) to high (red). (b) Phylogenetic tree based on VP1 amino acid sequence of *Microviridae* co-existing in four individual pigs. Four pigs which showed about twenty different genomes co-existing in each virome are included in this analysis. The twenty-two representative members of the previously confirmed subfamilies in *Microviridae* are marked with black diamonds and the clades in which they are located are shaded with different colors. The subfamily names are shown in text box beside their corresponding shaded colors. The scale bar indicates amino acid substitutions per site. As the predominance of *Gokushovirinae* in the approximate twenty *Microviridae* in each of the four pigs, pairwise sequence comparison of the VP1 amino acid sequences of *Gokushovirinae* identified in this study in each clade within these trees was also performed. The percent identities are shown beside each clade by heat map, where plot colors and color saturation reflect the identity, ranging from low (blue) to high (red).

of SSM (Fig. 1b). The comparison of SSM percentage in fecal virome of yaks with different health status and farms suggested that the abundance of SSM was not directly related to diarrhea of yaks in this study (Fig. 1c). Among the different sample types of cattle, feces, and genital secretion contained the highest abundance (1.34%) of SSM (Fig. 1d). Among the feces, skin swab, and oral swab from each of twenty wild rats, skin swab showed SSM percent of 0.22 per cent which is higher than those in feces (0.17%) and oral swab (0.06%) (Fig. 1e).

3.2 Genomes of *Microviridae* in different samples

From these viromes of the samples in this study, 713 complete circular genomes and 147 contigs with complete major capsid protein (VP1) and replication protein (VP4) of *Microviridae* were generated by *de novo* assembly. The number of *Microviridae* genomes from pigs (*Sus scrofa domestica*) was 461, accounting for 53.5 per cent of the genomes, followed by cattle (n = 98), yak (*Bos grunniens*) (n = 53), dog (*Canis lupus familiaris*) (n = 42), and others (Supplementary Fig. S1a). Based on the 713 complete circular genomes of *Microviridae*, the genome sizes ranged from 3,143 to 8,312 bp with 19 genomes <4.1 kb and 128 genomes >6.1 kb (Fig. 1f), significantly exceeding the known *Microviridae* genome size range of 4.4–6.1 kb (King et al. 2012). The characteristic genes encoding the VP1 and replication protein (VP4) were identified in all the 860 complete or near complete genomes in this study. Out of the 713 complete circular genomes, 569 genomes possessed VP1, VP4, VP2 (DNA pilot protein), and VP3 (internal scaffolding protein), where 126 genomes contained the gene order of VP4–VP1–VP2–VP3 while 433 genomes had the gene order of VP4–VP3–VP1–VP2. Two of the complete genomes only contained VP1, VP2, and VP4, and the remaining 152 genomes had unclear genome structure which all contained VP1 and VP4 while lacking VP2 and VP3 instead containing unassigned ORFs at equivalent positions. Among the 569 genomes possessing VP1, VP2, VP3, and VP4, 280 genomes contained VP5 which has function of mediating synthesis of viral ssDNA (Doore and Fane 2016), where 276 genomes had VP5 located behind VP4 and 4 genomes had VP5 before VP4. Gene order of these genomes was shown in Supplementary Table S2 and Supplementary Fig. S1b.

3.3 Divergence of VP1 in genomes of *Microviridae*

The well-conserved major capsid protein VP1 of *Microviridae* can be used as a phylogenetic marker facilitating the classification (Hopkins et al. 2014; Quaiser et al. 2015). By BLASTp search in GenBank based on the VP1 amino acid sequence of the 860 *Microviridae* strains, the sequence identities between our sequenced *Microviridae* and their best matches in GenBank were delineated (Fig. 1g), which indicated that 204 strains showed >60 per cent sequence identities to their best matches in GenBank and the other 656 stains shared <60 per cent identity

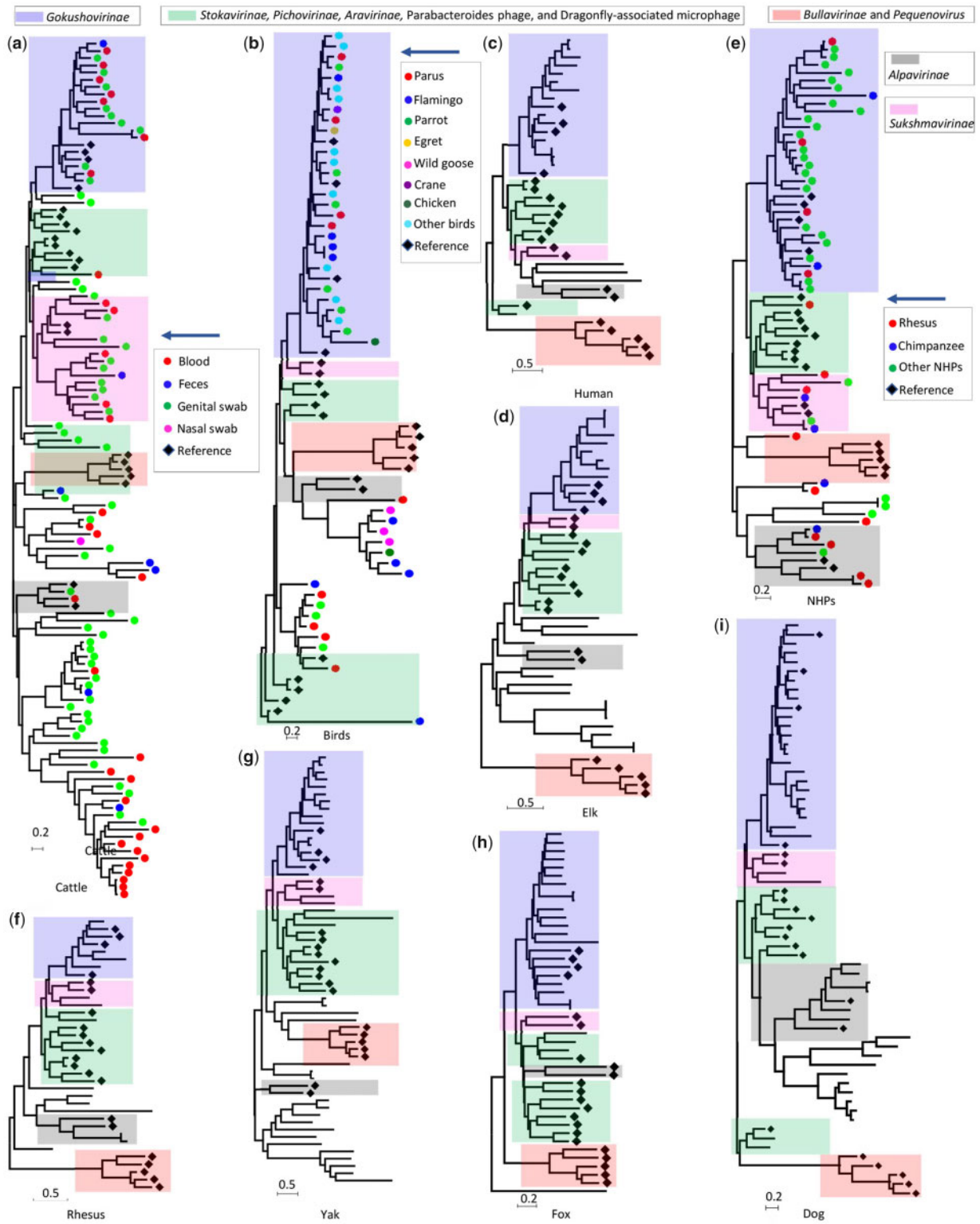

Figure 4. Phylogenetic analysis of *Microviridae* from virome in different species or group of animals. Animals species or group names are shown under each tree. Phylogenetic tree based on VP1 amino acid sequence of the *Microviridae* strains identified in this study and twenty-two representative members of the previously identified subfamilies. Within the tree, the twenty-two representative members of the previously confirmed subfamilies in *Microviridae* are marked with black dots and the clades in which they are located are shaded with different colors. The subfamily names are shown in text box beside their corresponding shaded colors. The scale bar indicates amino acid substitutions per site. In the phylogenetic analysis based on sequences from cattle (panel a), different sample types are marked with different color dots. In the phylogenetic analysis based on sequences from birds (panel b) or non-human primates (panel e), different species of animals are marked with different color dots.

with the best matches including 159 strains <30 per cent identity, suggesting new putative clades in the *Microviridae* family. Based on the names of best matched viruses, 408 of these genomes belonged to species in the subfamily *Gokushovirinae*, accounting for 47.4 per cent of the *Microviridae* strains identified here (Supplementary Fig. S1c and Supplementary Table S2). Pairwise sequence comparison of the 860 amino acid sequences of VP1 identified in this study showed that they shared sequence identities ranging from 10.3 to 100 per cent, where >70 per cent of these identities showed <50 per cent, indicating the highly genetic diversity of these *Microviridae* (Fig. 1h).

3.4 Phylogenetic analysis based on VP1 proteins

To further assess the genetic diversity of the 860 *Microviridae* strains, phylogenetic analysis was performed using the amino acid sequences of the major capsid protein VP1, where 18 representative sequences of already identified *Microviridae* were also included (Fig. 2a). Results indicated that 437 *Microviridae* genomes clustered within clades related to *Gokushovirinae*, 23 genomes felt into the cluster which also included *Stokavirinae*, *Pichovirinae*, *Aravirinae*, and Parabacteroides-phage, 9 strains were related to *Bullavirinae* though showing highly divergence with each other, 17 genomes clustered with the dragonfly-associated microphage which was suggested to be a potentially novel subfamily (Group D) (Quaiser et al. 2015), 27 sequences were clustered within the clade of *Alpavirinae*, the other 347 *Microviridae* genomes did not clustered within any known families but formed several potentially new clades (Fig. 2a). From each of the potential new clades, two or three representative genomes were selected for a new phylogenetic analysis which confirmed that several groups of strains place outside of existing subfamilies in *Microviridae* (Fig. 2b). To investigate whether the 860 *Microviridae* genomes clustered according to their isolation sources and geographic location, metadata of host, and geographic information were added to each strains and phylogenetic analysis was performed, which indicated that overall these *Microviridae* failed to cluster according to animal sources and geographic sampling sites, although a small number of strains from the sample host or sampling site did cluster phylogenetically (Fig. 2c and d).

Phylogenetic analysis based on those 410 strains from pigs located in four different sampling pig farms in Shanghai indicated that these *Microviridae* genomes did not cluster according to farms (Fig. 3a), where about 50 per cent of the strains showed relationship to *Gokushovirinae* while the remaining genomes fell into several potential new subfamilies. Out of the 102 pig feces barcodes, 88 were constructed using individual samples, where different *Microviridae* genomes were detected in a single barcode, suggesting the presence of multiple viral strains in a single animal. Four of the barcodes contained 19–22 different genomes that were used for phylogenetic analysis to investigate the relationship among different *Microviridae* strains co-existing in the same animal. Each of the four pigs showed >10 divergent strains related to *Gokushovirinae*, which also suggested that even in single animal multiple divergent genomes from the same subfamily of *Microviridae* could be present (Fig. 3b). Besides pig, seven of the other groups of animals and human contained a large number of different *Microviridae* genomes which were also subjected to phylogenetic analysis based on each species or group of animals, respectively (Fig. 4a–i). Although *Gokushovirinae* genomes were dominant among members of the *Microviridae* family in most of the animal species, they only accounted for a fraction (17.3%) of the 98 *Microviridae* genomes from cattle.

4. Discussion

The *Microviridae*, being secondary actors in environmental viral communities (Reyes et al. 2012; Scarpellini et al. 2015), is one of the most globally ubiquitous and highly diverse virus families (Tucker et al. 2011; Roux et al. 2012; Hopkins et al. 2014; Quaiser et al. 2015). The discovery of novel bacteriophages is classically dependent on plaque assays of bacterial lawns, such as the historically known bacteriophage T7 (Yin 1993; Abedon 2018), and some single-stranded DNA (ssDNA) phages were also identified through culture methods recently (Nasu et al. 2000; Murugaiyan et al. 2011). As a result, the majority of previously identified phages were isolated from bacteria belonging to easily culturable groups. So far, most of the recognized phages are tailed, dsDNA phages belonging to the order *Caudovirales* (Adams et al. 2016), while information on the ssDNA is still limited, with members of the *Microviridae* family only representing about 2 per cent of the phages in the ICTV database. Many studies have demonstrated the ubiquity of *Microviridae* genomes across habitats (marine, freshwater, wastewater, sediment) and global regions (Antarctic to subtropical), especially those related to the *Gokushovirinae* lineage (Tucker et al. 2011; Roux et al. 2012; Hopkins et al. 2014; Quaiser et al. 2015). With increasing number of novel *Microviridae* genomes, more potential subfamilies have been suggested including *Alpavirinae*, which seem to be mainly associated with human microbiota (Krupovic and Forterre 2011), *Pichovirinae*, which have a broader environmental range (Roux et al. 2012), and *Stokavirinae*, *Aravirinae*, and Group D, which were identified from sphagnum-dominated peatlands (Quaiser et al. 2015). In the present study, using viral metagenomics, 860 complete or near complete genomes showing sequence similarity to *Microviridae* in the virome of different types of samples from human, mammals, and birds, and digestive tract of oyster and fresh water were acquired. These genomes showed considerable variations, even in the sample from a single animal showing multiple co-existence of different genomes. Sequence comparison and phylogenetic analysis based on the major capsid protein, provided insights into the evolution and genetic diversity of the *Microviridae* family, suggesting that although the *Gokushovirinae* subfamily was dominant some new subfamilies might exist. Identification of the corresponding bacterial hosts should help elucidate the ecological and functional significance of these *Microviridae* subfamilies.

Acknowledgements

We thank Ms Jingjiao Li, Mr Xiuguo Hua, Ms Li Cui, Mr Dunwu Qi, Mr Shouxin Li, and Mr Hongxing Shen for their help in sample collection.

Data availability

The raw sequence reads from the metagenomic libraries were deposited in the Short Read Archive of GenBank database and the accession numbers were listed in Supplementary Table S1. The 860 *Microviridae* genomes identified in this study were submitted to GenBank and the accession numbers were listed in Supplementary Table S2.

Supplementary data

Supplementary data are available at *Virus Evolution* online.

Funding

This work was partly supported by National Key Research and Development Programs of China (No. 2017YFC1200201), Jiangsu Provincial Key Research and Development Projects (No. BE2017693), NSFC (No.31872478), and Blood Systems Research Institute.

Conflict of interest: None declared.

References

Abedon, S. T. (2018) 'Detection of Bacteriophages: Phage Plaques', in Harper, D., Abedon, S., Burrowes, B., and McConville, M. (eds.) *Bacteriophages*, pp. 1–32. Cham: Springer International Publishing.

Adams, M. J. et al. (2016) 'Ratification Vote on Taxonomic Proposals to the International Committee on Taxonomy of Viruses (2016)', *Archives of Virology*, 161: 2921–49.

Barrientos-Somarribas, M. et al. (2018) 'Discovering Viral Genomes in Human Metagenomic Data by Predicting Unknown Protein Families', *Scientific Reports*, 8: 28.

Blackburn, B. J. et al. (2017) 'Coat Protein Mutations That Alter the Flux of Morphogenetic Intermediates through the ϕX174 Early Assembly Pathway', *Journal of Virology*, 91: e01384–17.

Brentlinger, K. L. et al. (2002) '*Microviridae*, a Family Divided: Isolation, Characterization, and Genome Sequence of phiMH2K, a Bacteriophage of the Obligate Intracellular Parasitic Bacterium *Bdellovibrio bacteriovorus*', *Journal of Bacteriology*, 184: 1089–94.

Bryson, S. J. et al. (2015) 'A Novel Sister Clade to the Enterobacteria Microviruses (Family *Microviridae*) Identified in Methane Seep Sediments', *Environmental Microbiology*, 17: 3708–21.

Creasy, A. et al. (2018) 'Unprecedented Diversity of ssDNA Phages from the Family *Microviridae* Detected within the Gut of a Protochordate Model Organism (*Ciona robusta*)', *Viruses*, 10: 404.

Deng, X. et al. (2015) An Ensemble Strategy That Significantly Improves De Novo Assembly of Microbial Genomes from Metagenomic Next-Generation Sequencing Data. *Nucleic Acids Research*, 43: e46.

Desnues, C. et al. (2008) 'Biodiversity and Biogeography of Phages in Modern Stromatolites and Thrombolites', *Nature*, 452: 340–3.

Doore, S. M., and Fane, B. A. (2016) 'The *Microviridae*: Diversity, Assembly, and Experimental Evolution', *Virology*, 491: 45–55.

Eddy, S. R. (2009) 'A New Generation of Homology Search Tools Based on Probabilistic Inference', *Genome Informatics*, 23: 205–11.

Finn, R. D., Clements, J., and Eddy, S. R. (2011) 'HMMER Web Server: Interactive Sequence Similarity Searching', *Nucleic Acids Research*, 39: W29–37.

Guo, L. et al. (2017) 'Viral Metagenomics Analysis of Feces from Coronary Heart Disease Patients Reveals the Genetic Diversity of the *Microviridae*', *Virologica Sinica*, 32: 130–8.

Hopkins, M. et al. (2014) 'Diversity of Environmental Single-Stranded DNA Phages Revealed by PCR Amplification of the Partial Major Capsid Protein', *The ISME Journal*, 8: 2093–103.

Johnson, L. S., Eddy, S. R., and Portugaly, E. (2010) 'Hidden Markov Model Speed Heuristic and Iterative HMM Search Procedure', *BMC Bioinformatics*, 11: 431.

King, A. M. Q. et al., eds. (2012) 'Virus taxonomy: classification and nomenclature of viruses', in *Ninth report of the International Committee on Taxonomy of Viruses*. London, UK: Academic Press.

Krupovic, M., and Forterre, P. (2011) '*Microviridae* Goes Temperate: Microvirus-Related Proviruses Reside in the Genomes of Bacteroidetes', *PLoS One*, 6: e19893.

Labonte, J. M., and Suttle, C. A. (2013) 'Metagenomic and Whole-Genome Analysis Reveals New Lineages of Gokushoviruses and Biogeographic Separation in the Sea', *Frontiers in Microbiology*, 4: 404.

Liu, Z. et al. (2016) 'Identification of a Novel Human Papillomavirus by Metagenomic Analysis of Vaginal Swab Samples from Pregnant Women', *Virology Journal*, 13: 122.

McKenna, R. et al. (1992) 'Atomic Structure of Single-Stranded DNA Bacteriophage ΦX174 and Its Functional Implications', *Nature*, 355: 137–43.

Morella, N. M. et al. (2018) 'The Impact of Bacteriophages on Phyllosphere Bacterial Abundance and Composition', *Molecular Ecology*, 27: 2025–38.

Murugaiyan, S. et al. (2011) 'Characterization of Filamentous Bacteriophage PE226 Infecting *Ralstonia solanacearum* Strains', *Journal of Applied Microbiology*, 110: 296–303.

Nasu, H. et al. (2000) 'A Filamentous Phage Associated with Recent Pandemic *Vibrio parahaemolyticus* O3:K6 Strains', *Journal of Clinical Microbiology*, 38: 2156–61.

Quaiser, A. et al. (2015) 'Diversity and Comparative Genomics of *Microviridae* in Sphagnum-Dominated Peatlands', *Frontiers in Microbiology*, 6: 375.

Reyes, A. et al. (2012) 'Going Viral: Next-Generation Sequencing Applied to Phage Populations in the Human Gut', *Nature Reviews Microbiology*, 10: 607–17.

Roux, S. et al. (2012) 'Evolution and Diversity of the *Microviridae* Viral Family through a Collection of 81 New Complete Genomes Assembled from Virome Reads', *PLoS One*, 7: e40418.

Salmond, G. P. C., and Fineran, P. C. (2015) 'A Century of the Phage: Past, Present and Future', *Nature Reviews Microbiology*, 13: 777–86.

Scarpellini, E. et al. (2015) 'The Human Gut Microbiota and Virome: Potential Therapeutic Implications', *Digestive and Liver Disease*, 47: 1007–12.

Skewes-Cox, P. et al. (2014) 'Profile Hidden Markov Models for the Detection of Viruses within Metagenomic Sequence Data', *PLoS One*, 9: e105067.

Tucker, K. P. et al. (2011) 'Diversity and Distribution of Single-Stranded DNA Phages in the North Atlantic Ocean', *The ISME Journal*, 5: 822–30.

Xie, X.-T. et al. (2019) 'Prevalence of Fecal Viruses and Bacteriophage in Canadian Farmed Mink (*Neovison vison*)', *MicrobiologyOpen*, 8: e00622.

Yin, J. (1993) 'Evolution of Bacteriophage T7 in a Growing Plaque', *Journal of Bacteriology*, 175: 1272–7.

Zhang, W. et al. (2014) 'Faecal Virome of Cats in an Animal Shelter', *The Journal of General Virology*, 95: 2553–64.

—— et al. (2016) 'Viral Nucleic Acids in Human Plasma Pools', *Transfusion*, 56: 2248–55.