npj | biofilms and microbiomes

Article

Published in partnership with Nanyang Technological University

https://doi.org/10.1038/s41522-024-00554-0

Massive expansion of the pig gut virome based on global metagenomic mining

Check for updates

Jiandui Mi[1], Xiaoping Jing[2], Chouxian Ma[3], Yiwen Yang[4], Yong Li[4], Yu Zhang[5], Ruijun Long[2] & Haixue Zheng[6]

The pig gut virome plays a vital role in the gut microbial ecosystem of pigs. However, a comprehensive understanding of their diversity and a reference database for the virome are currently lacking. To address this gap, we established a Pig Virome Database (PVD) that comprised of 5,566,804 viral contig sequences from 4650 publicly available gut metagenomic samples using a pipeline designated "metav". By clustering sequences, we identified 48,299 viral operational taxonomic units (vOTUs) genomes of at least medium quality, of which 92.83% of which were not found in existing major databases. The majority of vOTUs were identified as *Caudoviricetes* (72.21%). The PVD database contained a total of 2,362,631 protein-coding genes across the above medium-quality vOTUs genomes that can be used to explore the functional potential of the pig gut virome. These findings highlight the extensive diversity of viruses in the pig gut and provide a pivotal reference dataset for forthcoming research concerning the pig gut virome.

The gut microbiota is a complex microbial ecosystem that plays crucial roles in pig health, nutrient metabolism, and productivity[1–4]. Pig intestinal virus infections are prevalent in the global pig industry, but the many diverse viruses that inhabit the pig gut can greatly affect the structure and function of the gut microbial ecosystem[5]. Controlling the occurrence of these diseases solely through vaccines or medication is difficult. Thus, understanding the pig gut virome is essential for improving pig health and enhancing production efficiency. However, despite their ubiquity, knowledge of the diversity of the virome in the pig gut ecosystem is limited, and most virome genomes fail to be found in existing genome databases. Furthermore, most virome databases are established based on human gut metagenomic samples[6–8], and there is a lack of a specialized virus database for pigs[9]. Therefore, a comprehensive database of the virome from the pig gut microbiota is a prerequisite for characterizing virus diversity, understanding host-virus interactions, and resolving the functions of viral genomes.

Currently, there is a wealth of metagenomic data available that offers a unique opportunity to discover viral genomes[10]. Despite being generated through untargeted methods without virus particle enrichment, these datasets still contain a significant number of viral genomes. Several databases, such as GOV2[11], IMG/VR4[12], GVD[7], MGV[8], GPD[6], and RVD[13], have been established to facilitate the analysis of viromes from various environments using metagenomic data. Since the establishment of these databases, the number of publicly available pig gut microbiota datasets has rapidly increased[9,14–21].

To make use of the existing resources and provide a comprehensive, global view of the pig gut virome, we developed a comprehensive metav analysis pipeline to examine 4650 metagenomic samples. We optimized several software tools to improve processing speed and end-to-end output. The Pig Virome Database (PVD) of the gut was established, containing 5,566,804 viral contig sequences estimated to be >50% complete, representing 48,299 vOTU genomes. Our analysis revealed a diverse and complex pig gut virome, with a high number of unique vOTUs (92.83%) compared to other databases. Moreover, we identified several potential novel viral species in the pig gut. These findings enhance our understanding of the pig gut virome, and provide insights into the complexity of gut ecosystems, emphasizing the importance of further research in this field.

Results

The DNA viruses from the pig gut microbiome

In this study, our objective was to create a comprehensive pig virome database (PVD) of the gut utilizing next-generation sequencing of metagenomic samples and the development of metav, a virus detection pipeline

[1]State Key Laboratory for Animal Disease Control and Prevention, College of Veterinary Medicine, Lanzhou University, Lanzhou, China. [2]State Key Laboratory of Grassland and Agro-Ecosystems, International Centre for Tibetan Plateau Ecosystem Management, College of Ecology, Lanzhou University, Lanzhou, China. [3]Independent Researcher, Changsha, China. [4]College of Animal Science, South China Agricultural University, Guangzhou, China. [5]Guangdong Provincial Research Center for Environment Pollution Control and Remediation Materials, College of Life Science and Technology, Jinan University, Guangzhou, China. [6]State Key Laboratory for Animal Disease Control and Prevention, College of Veterinary Medicine, Lanzhou University, Lanzhou Veterinary Research Institute, Chinese Academy of Agricultural Sciences, Lanzhou, China. e-mail: mijiandui@163.com; longrj@lzu.edu.cn; haixuezheng@163.com

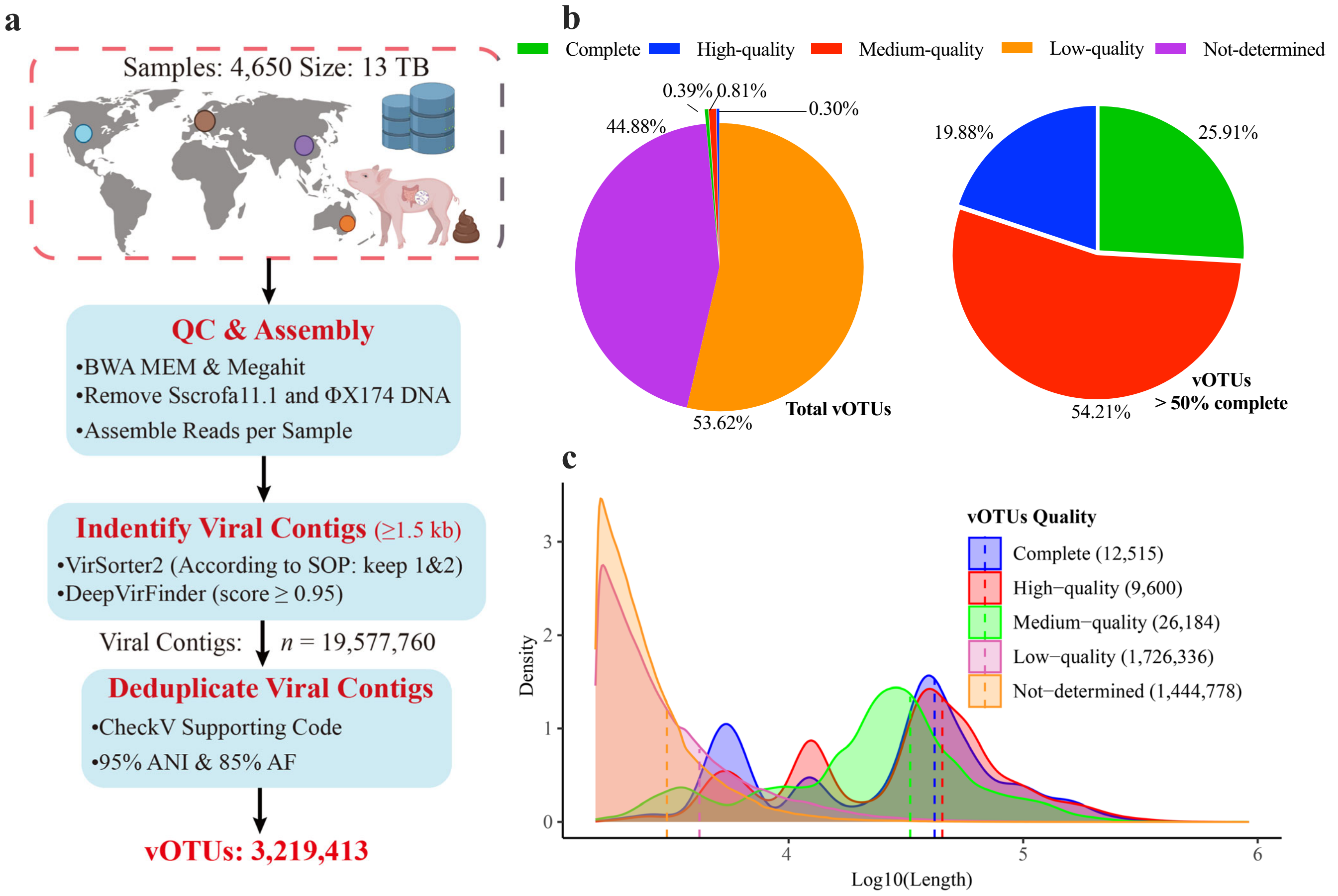


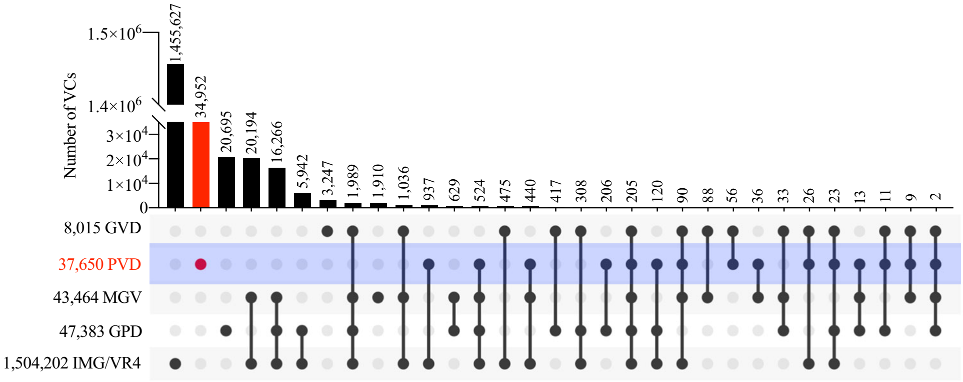
Fig. 1 | Summary of viral genomes recovered from pig gut metagenomes. **a** Overview of the PVD database viral discovery pipeline. **b** Proportion of vOTU genomes based on quality tiers. **c** Distribution of density and number for different vOTU qualities: complete (*n* = 12,515), high quality (*n* = 9600), medium quality (*n* = 26,184), low quality (*n* = 1,726,336), and not determined (*n* = 1,444,778).

that integrates established methods and signatures such as VirSorter2 and DeepVirFinder for viral sequence identification[22,23]. Our findings indicated that the choice of assembler, specifically Megahit versus metaSPAdes[24,25], had a minimal effect on the recovery of viruses, which was consistent with prior research[8]. We used the Megahit assembler and selected -k-list 39, 59, and 79 after testing different Kmer lists based on computational efficiency. We applied metav to 4650 fecal and gut content samples from various regions worldwide, including the USA, China, Europe, and Australia, and identified 19,577,760 unique, single-contig viral genomes exceeding 1.5 kb in length, which allowed us to comprehensively capture DNA viruses in the pig gut (Fig. 1a and Supplementary Fig. 1). By clustering viral contigs using 95% ANI and 85% AF, we identified 3,219,423 viral operational taxonomic units (vOTUs) at the species level (Fig. 1a). In addition, we evaluated vOTU genome quality using CheckV and found 48,299 vOTU genomes (corresponding to 5,566,804 viral contig sequences) that were at least 50% complete[26], including 12,515 complete vOTU genomes (corresponding to 2,099,394 viral contig sequences) (Fig. 1b, c, Supplementary Table 1, and Supplementary Fig. 2). Notably, although only 1.50% of the total vOTU genomes were more than 50% complete, they corresponded to 28.43% of the contig sequences with >50% completeness of the total contig genomes (Fig. 1c). Moreover, the size of most vOTU genomes with completeness exceeding 50% was greater than 10 kb, indicating their potential to provide more reliable results[8,27].

Newly established and expanded pig gut viral diversity

To ensure comparability and consistency with established quality standards for genomic comparisons[28,29], we focused on 48,299 vOTU genomes whose completeness exceeded 50%. We used CheckV on the IMG/VR4, MGV, GVD, and GPD datasets to remove low-quality sequences, retaining vOTU genome fragments with more than 50% completeness. Following previous studies (Camarillo-Guerrero et al. [6]; Nayfach et al.)[6,8], we combined all vOTU genomes >50% from these four datasets with those from the PVD database and clustered them at the species level using a query coverage of 0.9 and an identity of 0.7. Our results showed that the PVD database significantly enhanced the diversity of viral genomes from pig gut microbiota metagenomic samples. Of the 47,650 species-level vOTUs in the PVD database, 34,952 (92.83%) did not cluster with any vOTU genomes from the other datasets (Fig. 2). The generalist IMG/VR4 database contains many more viruses because it does not contain exclusively mammalian gut samples. Of the 1,504,202 vOTU genomes from the IMG/VR4, 1,455,627 (96.77%) were unique compared to those from other databases (Fig. 2). However, the human gut metagenome samples used to construct the GVD, MGV, and GPD databases represented a greater proportion of viral genomes in the IMG/VR4 database than pig gut viruses. Specifically, GVD, MGV, and GPD had 4712 (58.79%), 43,464 (95.19%), and 25,377 (53.56%) sequences from human gut metagenomes, respectively, while GPD, GVD, and MGV had 20,695 (43.68%), 3247 (40.51%), and 1910 (4.39%) unique sequences compared to the other databases, respectively (Fig. 2). Our findings highlighted the importance of using comprehensive and high-quality databases to improve our understanding of viral diversity in different environments. The improved detection of viral reads in whole metagenomes and expanded coverage of virus-host connections in the PVD database make it particularly useful for studying pig gut viruses.

Fig. 2 | Clustering and comparison with four existing datasets of vOTUs that were above medium-quality. The vOTU genomes from the PVD (n = 37,650) catalog were compared with other virus genomes with >50% completeness from four databases: IMG/VR4 (n = 1,504,202), MGV (n = 43,464), GPD (n = 47,383), and GVD (n = 8015) at the species level with the parameters query coverage 90, and identity 70.



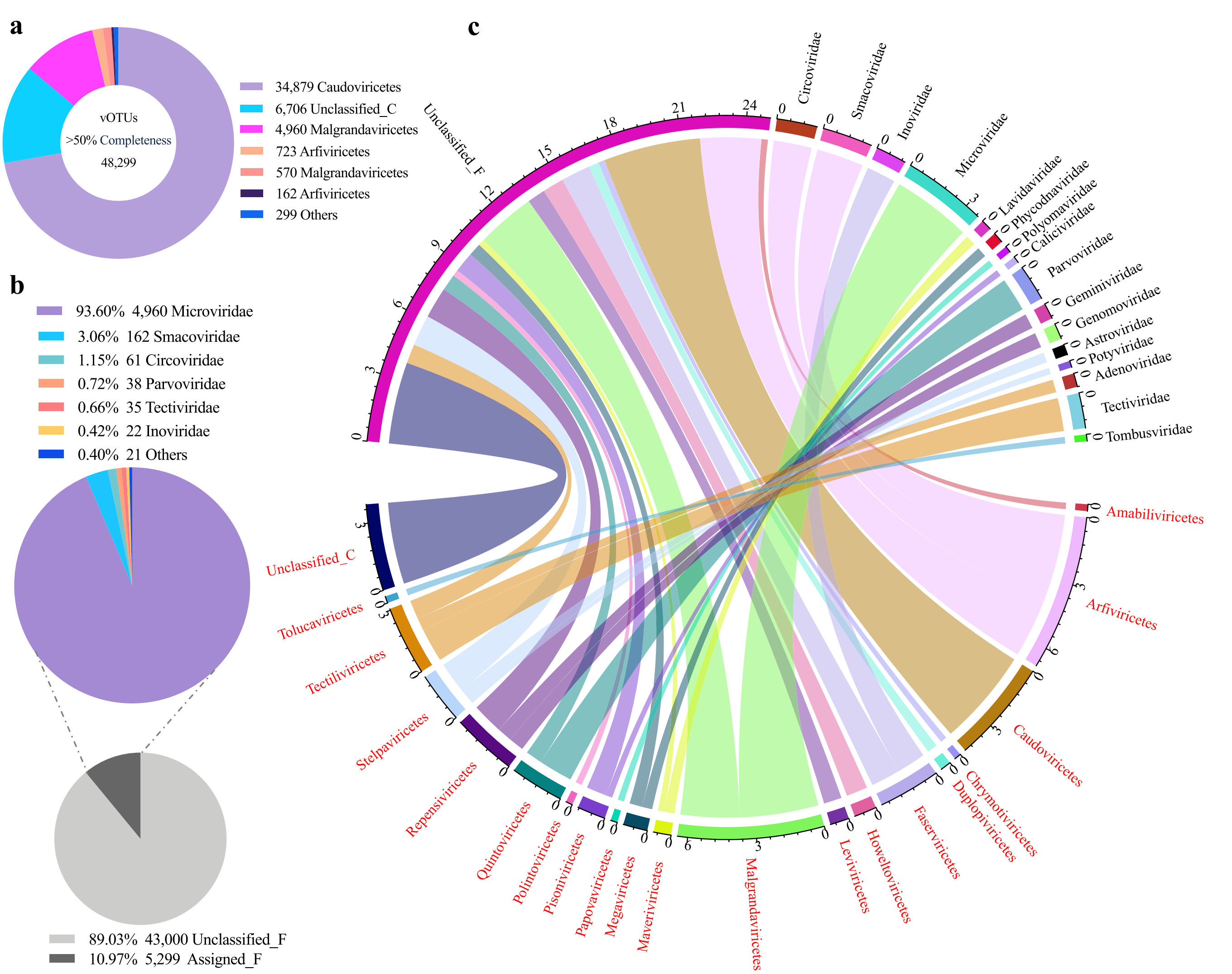


Fig. 3 | The taxonomy and composition of vOTUs with >50% completeness. a Composition of vOTUs at the class level. **b** Composition of vOTUs at the family level. **c** Circle chart that depicts the vOTU taxonomy at both the class and family levels. The names of the classes are shown in red font, and those of families are shown in black font.

Taxonomic annotation

We utilized the new ICTV classification, which abolished the paraphyletic morphological families *Podoviridae*, *Siphoviridae*, and *Myoviridae* and the order *Caudovirales* and replaced them with the class *Caudoviricetes* to categorize all tailed bacterial and archaeal viruses featuring icosahedral capsids and a double-stranded DNA genome[30]. After using the new ICTV classification, the majority of vOTUs (65.36%) were identified as *Caudoviricetes* (Fig. 3a and Supplementary Table 2). However, only 10.97% of the medium-quality vOTU genomes could be annotated at the family level using the new ICTV database (Fig. 3b, c). These findings suggested significant gaps in knowledge regarding pig gut virus taxonomy, which is also a significant challenge for analyzing viromes from different environments[8,31,32].

Host prediction and temperate identification

Accurately predicting virus hosts is a critical step towards understanding virus-host interactions and utilizing them to manipulate the gut microbiota ecosystem or design innovative phage tools[33,34]. In this study, we used the iPHoP tool, which integrates multiple methods, to reliably predict the host taxonomy of medium-quality vOTUs at the genus level[35]. As expected, Firmicutes and Bacteroidetes, the most abundant bacterial phyla in the gut microbiome, were found to be common hosts for viruses at the phylum level (Fig. 4a, b). At the genus level, *Prevotella* and *Bacteroides* (belonging to Bacteroidetes) and *Vescimonas* and *Faecousia* (belonging to Firmicutes) were the most commonly assigned hosts (Fig. 3a). Additionally, some viruses were predicted to inhabit *Lactobacillus*, *Ruminococcus, Faecalibacterium*, and *Blautia*, which are known to play important roles in pig productivity and feed efficiency[36,37].

Previous study has shown that using incomplete viral contigs for lifestyle prediction can lead to an overestimation of the number of virulent viruses[8]. Thus, we used only the above high-quality (>90% completeness) vOTU genomes to predict the virulent or temperate phages with two software programs, BACPHLIP and PhaTYP (Fig. 4c). BACPHLIP showed that the majority of the above high-quality (>90% completeness) vOTU genomes (64%) were predicted to be virulent (Fig. 4c). PhaTYP, a python library for bacteriophage prediction with a Bidirectional Encoder Representations from Transformer (BERT)-based model, was used to confirm the lifestyles of the same set of high-quality vOTUs (>90% completeness, $n = 22{,}115$). The results showed that the temperate, virulent, and unknown categories accounted for 62%, 34%, and 4%, respectively, of all predictions. However, other studies have reported that virulent viruses accounted for 65%~91% of the total viruses in the human gut environment, which might be due to the inclusion of low-quality vOTU[8,38]. These discrepancies suggest that further research should employ high-quality vOTUs to predict the lifestyles of viruses more reasonably.

Phylogenomic analyzes of viruses

Viruses that infect vertebrates can cause serious diseases in pigs, leading to significant economic losses for the swine industry. For example, porcine circovirus can cause porcine circovirus-associated disease (PCVAD)[39,40]. In our metagenomic samples, we did not detect *Asfarviridae*, a large encapsulated double-stranded DNA virus[41], possibly because the samples were collected from healthy or noninfected pigs. To explore the diversity of pig gut viruses, we constructed a phylogenetic tree based on 265 genomes with >50% completeness from the PVD (Fig. 5). The majority of viruses from the PVD represented new lineages across the tree. *Smacoviridae* and *Circoviridae* were the most diverse families found, primarily due to the broad phylogenetic distribution of the vOTUs belonging to these groups, which are circular replication-associated protein (Rep)-encoding single-stranded DNA viruses[42–44]. Moreover, we found porcine circovirus-like viruses that have been associated with porcine diarrheal disease[43]. Additionally, we found 206 viral genomes with >50% completeness in which the host was dominated by *Methanobacteriaceae* and *Methanomethylophilaceae*, consistent with the archaeal virome found in the human gut (Supplementary Fig. 3)[45]. Compared with recently published collections of viruses from diverse environments and the human gut[8,46], our analysis of the PVD resulted in a substantial expansion of pig gut viral diversity.

Functional capacity of the gut virome

To investigate the potential roles of the pig gut virome, we identified 2,362,631 protein-coding genes across 48,299 of the above mentioned medium-quality vOTU genomes from our current study. To explore the

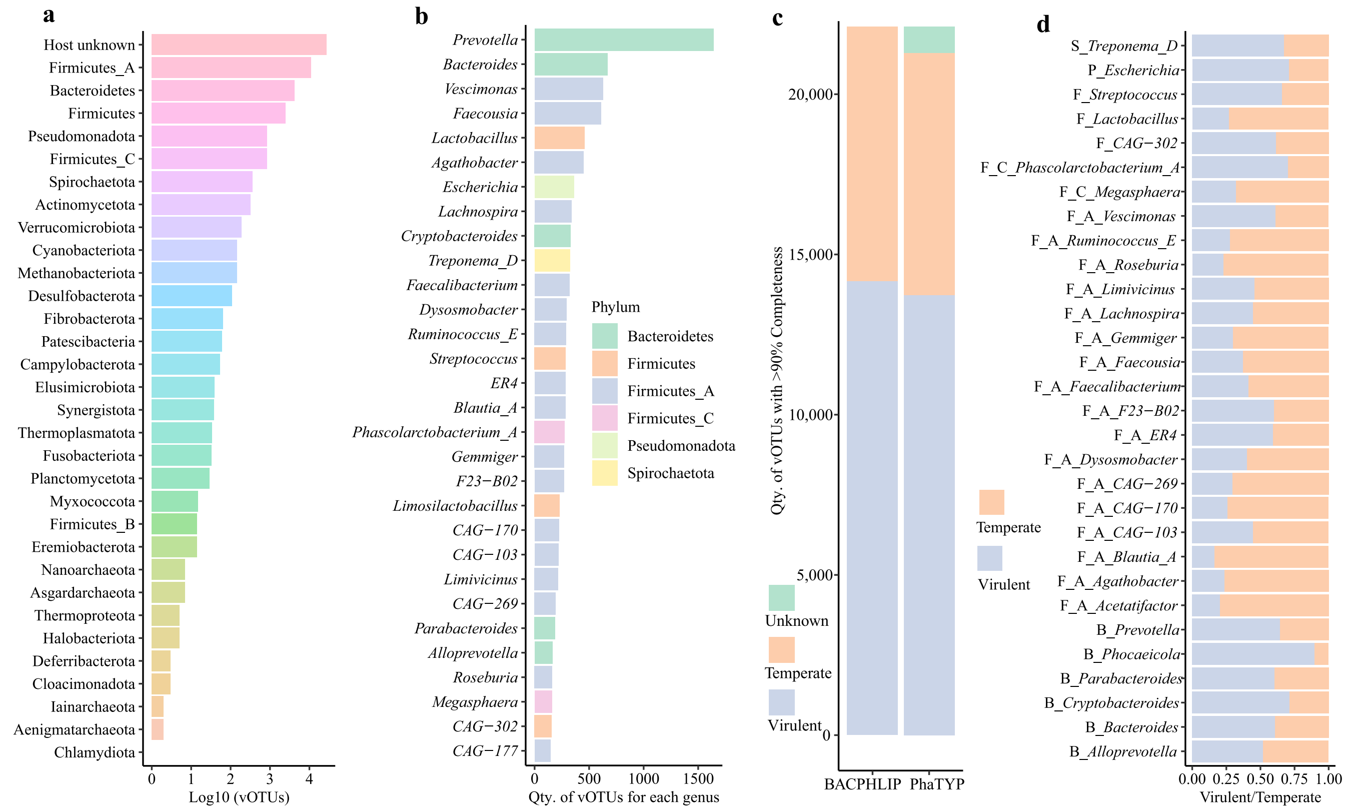


Fig. 4 | Host prediction and temperate identification of vOTUs. a Number of vOTUs with >50% completeness for each predicted host at the phylum level. **b** Number of vOTUs with >50% completeness for the top 30 predicted hosts at the genus level. **c** The ratio of virulent and temperate phages for total vOTUs with >50% completeness identified by PhaTYP and BACPHLIP. **d** The ratio of virulent and temperate phages for the top 30 vOTUs with >50% completeness identified by BACPHLIP at the genus level. The letters “S”, “P”, “F”, “F_C”, “F_A”, and “B” before the italicized genus names represent the Spirochaetota, Pseudomonadota, Firmicutes, Firmicutes_C, Firmicutes_A, and Bacteroidetes phyla, respectively.

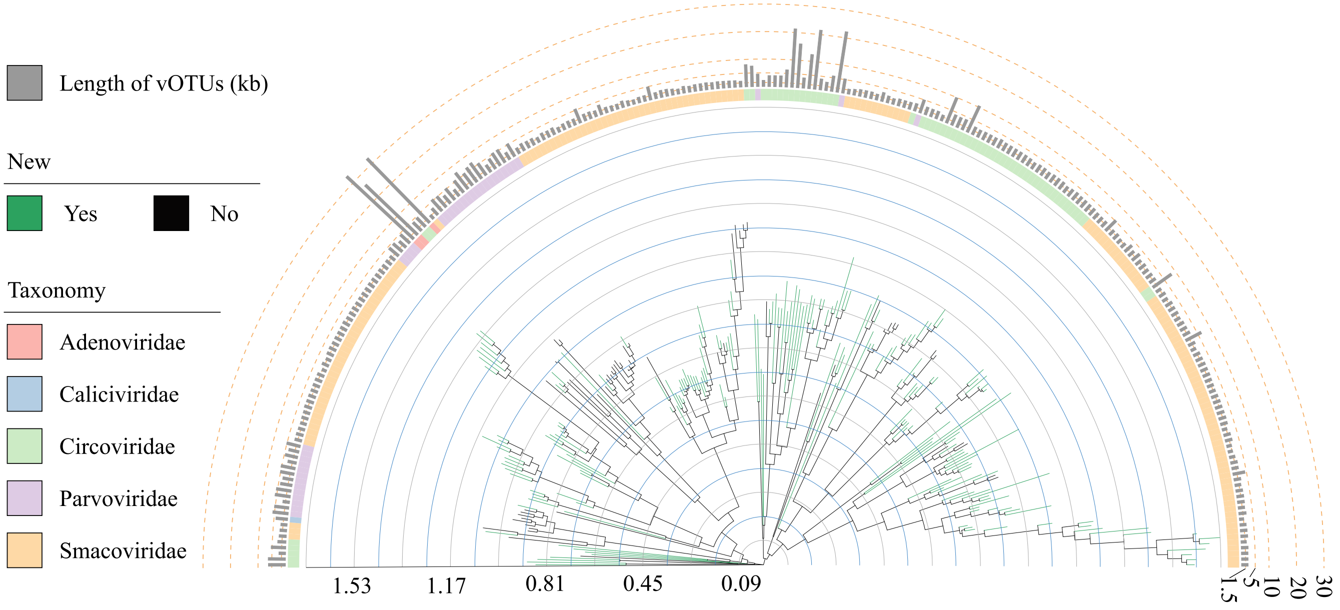


Fig. 5 | Phylogenetic tree of vOTUs with vertebrates as hosts. A phylogenetic tree was constructed from 265 viral genomes derived from the PVD. The tree was plotted using iTOL. Branch color indicates whether a lineage is represented by previously polished databases (IMG/VR4, MGV, GPD, and GVD) (black) or is unique to the PVD database (green). The different colors in the middle ring represent classifications at the family level. The outer ring displays the length (kb) for each vOTU.

functional potential of the pig gut virome, we used eggNOG-mapper v.2 for comparison against eggNOG 5.0 using default parameters[6,47]. Overall, 60% (1,416,705/2,362,631) of the vOTU genes matched the database (Fig. 6a). However, 73% of these genes were not assigned any biological function, highlighting our limited understanding of the pig gut virome. Among the matched genes, the largest protein clusters were single-stranded DNA-binding proteins ($n = 35{,}964$), followed by proteins involved in other typical viral functions, such as DNA polymerase and replicative DNA helicase (Fig. 6a). With respect to potential metabolic functions, we detected a total of 53,090 genes, with the top five categories being nucleotide metabolism ($n = 11{,}016$), amino acid metabolism ($n = 9492$), carbohydrate metabolism ($n = 9036$), metabolism of cofactors and vitamins ($n = 6652$), and energy metabolism ($n = 4533$) (Supplementary Fig. 4).

We then compared these genes to databases such as SARG, MGEs, VFDB, the quorum sensing, and BacMet databases using Diamond's BLASTp tool, with parameters set to --evalue1e-5, --query-cover 85, and --id 80 (Fig. 6b and Supplementary Tables 3–7). However, we only identified 69 protein-coding genes in 55 viral contigs ($55/48{,}299 = 0.11\%$) when checking the SARG database for antibiotic resistance genes (ARGs) (Supplementary Table 3). Among them, there are 10 contigs carrying 2 genes and 2 contigs carrying 3 resistance genes (Supplementary Table 3). The top two categories were macrolide-lincosamide-streptogramin (MLS) ($n = 21$) and multidrug ($n = 20$) (Fig. 6b). We observed that the parameters of query cover and identity had a significant impact on the number of ARGs detected. Various studies have used different parameters to detect ARGs carried by viruses, often below the 80% identity threshold used in our study and recommended by Yin et al. [48–52]. This could lead to inaccuracies and overestimations in the reported number of ARGs[49,50,52]. For example, by setting the parameters of –query-cover 85 and –id 60 in BLASTp, we obtained 4564 hits, which is 66 times greater than the previous setting. To compare results across different studies and establish more appropriate parameters for detecting virus-carried ARGs, comprehensive evaluations of parameters should be conducted in the future (Billaud et al.[53]; Enault et al., 2017)[53,54]. We utilized a previously established database to detect mobile genetic elements (MGEs) in pig gut vOTUs[55]. A total of 52 protein-coding genes were found and encompassed two dominant categories: transposase ($n = 21$) and integrase ($n = 15$) (Fi. 6b and Supplementary Table 4), which are known to play important roles in the transposition of ARGs[21,56]. However, we found only one viral genome containing both ARGs ($n = 2$) and MGEs ($n = 1$) (Fig. 6c). Moreover, most ARG-positive vOTUs found in pig metagenomes were not active[53]. Taken together, our results suggest that the pig gut virome carries a small number of ARGs and might contribute minimally to the gut resistome.

We also explored the distribution patterns of virulence factor genes (VFGs) carried by the virome (Fig. 6b and Supplementary Table 5). The results showed that VFGs were dominated by adherence ($n = 50$) and immune modulation ($n = 34$), consistent with previous studies[57]. Notably, we detected a small number of BacMet resistance protein genes ($n = 48$) in the pig gut vOTUs (Fig. 6b and Supplementary Table 6). The majority of metal resistance genes in the pig gut virome were associated with arsenic (As) ($n = 7$) and copper (Cu) ($n = 6$). Cu is extensively used as a feed additive for pigs and has a significant impact on the dissemination of antibiotic resistance genes (ARGs) in manure and soil[58]. We also examined vOTU proteins against the Quorum Sensing of Human Gut Microbes (QSHGM) database[59]. We identified only 74 protein genes of pig gut vOTUs that utilized quorum sensing languages, with the top three languages being HAQs, AI-2, and CAI-1 (Fig. 6b and Supplementary Table 7).

Discussion

In the present study, we conducted a comprehensive analysis of pig gut viral genomes by mining publicly available metagenomic samples ($n = 4650$) using our integrated pipeline designated metav (https://github.com/mijiandui/metav). The software tools in this pipeline were optimized to improve the speed and accuracy of metagenomic analysis and enable comparisons between different microbial ecosystems. We identified 19,577,760 draft-quality viral genomes, representing an estimated 3,219,413 species-level vOTUs. After applying CheckV, the PVD database was established, which included 5,566,804 viral contigs with more than 50% completeness, forming 48,299 species-level vOTUs. This is equivalent to the finding in the MGV ($n = 54{,}118$) and GPD ($n = 46{,}480$) human gut virome databases[6,8]. However, a pair-to-pair clustered comparison between IMG/VR4, GPD, MGV, GVD, and PVD showed that PVD still contained 34,952 (92.83%) unique vOTUs, making it the largest resource of extensive pig gut viral genomes to our knowledge. Although this study focused on DNA viruses, other studies have explored RNA viruses[5,10,60–63]. Thus, future

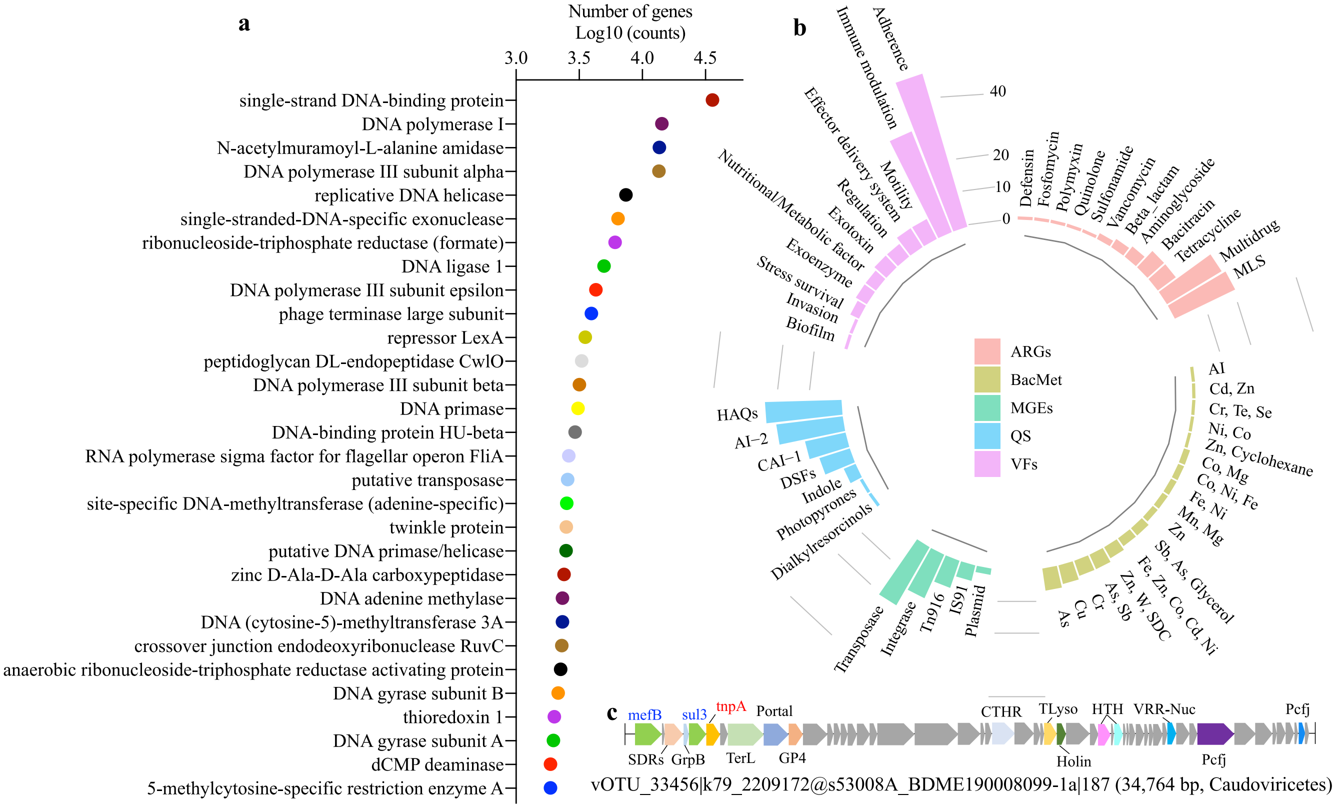


Fig. 6 | Functional landscape of vOTUs that were above medium quality. **a** Functional annotations for the 30 largest protein clusters. **b** Protein-coding viral genes were identified against a structured antibiotic-resistance gene (SARG) database, mobile genetic elements (MGEs) database, virulence factor database (VFDB), quorum sensing databases, and BacMet (Antibacterial Biocide & Metal Resistance Genes) with BLASTp using Diamond with the following specific parameters: --evalue1e-5 –query-cover 85 –id 80. **c** The gene structure of one vOTU containing ARGs and MGEs. The red font represents MGEs, while the blue font represents ARGs.

investigations could utilize metatranscriptomic data from pig gut microbiota samples to investigate RNA viromes.

To rapidly recover viral genomes from metagenomics and metatranscriptomics data, ICTV has updated its taxonomy list, which has led to significant changes compared to the previous version[30]. geNomad is significantly faster than similar tools and can process large datasets, accurately assigning identified viruses to the latest ICTV taxonomy release[64]. However, this modification has made it difficult to compare the taxonomic results of this study with those of previous studies. To address this issue, it would be useful, in the future, to integrate several large-scale viral genome databases and create a unified, updated, and standardized catalog that is compatible with the new ICTV taxonomy. It is also essential to establish a unified and standardized taxonomy classification pipeline or platform in the future[65].

The results of our study provided significant insights into the composition and function of the pig gut virome. Our analysis revealed a diverse and complex pig gut virome, consisting of numerous viral operational taxonomic units (vOTUs). In addition, we have identified several potentially novel viral species and found evidence of multiple viral species existing in the pig gut. These discoveries highlight the importance of continued research in this field, as they could have significant implications for our understanding of viral evolution, pathogenesis, and host-pathogen interactions. Moreover, our findings suggest that the existence of multiple viral species is common in the pig gut, which has important implications for future studies investigating the impact of the pig gut virome on pig health and productivity and the potential for viral transmission to other animals and humans. However, this study did not identify certain viruses, such as the African swine fever double-stranded DNA virus, that have a significant impact on the pig industry. The potential reason is that the samples collected for this study were from healthy pigs. In the future, building comprehensive pig intestinal virome databases derived from pigs affected by different pathogens (including viruses), and conducting a comparative analysis with the viromes of healthy pigs, will provide essential data support for the prevention and control of corresponding pig diseases. Given the lessons learned from the COVID-19 pandemic, such as the potential for emerging infectious diseases to adversely affect human health, there is a need for increased awareness of zoonoses[66,67]. Therefore, we propose a "One Health" framework that emphasizes the importance of studying the "ecosystem-animal-livestock-human" pathogen system using metagenomic technology. This approach should include combining rich metadata, such as climate change data, to reveal the sources, hosts, transmission, and evolutionary mechanisms of known and unknown pathogens[68].

In summary, our study lays the groundwork for further research into the pig gut virome. Our identification of new viral species and evidence of coexistence highlight the intricacies of gut ecosystems and emphasize the importance of further investigation in this area. Our findings provide the basis for a more comprehensive understanding of viral and microbial ecology in the pig gut and are expected to facilitate the monitoring and maintenance of pig health.

Methods

Assembly and viral identification in the gut contents and feces of pigs

We developed "metav", a pipeline for identifying viruses from raw metagenomics generated via next-generation sequencing. This analysis was comprised of three main steps: 1) quality control of the raw reads using fastp v.0.23.2 (with parameters --detect_adapter_for_pe and --dont_eval_duplication -w 16), followed by the removal of the host (pig genome: Sscrofa11.1) contamination using BWA v.0.7.17 with parameters (-k 31 -p -S -K

200000000)[69,70]; 2) assembly of the data using Megahit v.1.2.9 (with -k-list 39, 59, and 79)[25]; and 3) viral identification using VirSorter2 v.2.2.3 and DeepVirFinder v.2.0[22,23]. Viral sequences were identified with the criteria Keep1 (viral_gene >0) and Keep2 (viral_gene = 0 AND (host_gene = 0 OR score >= 0.95 OR hallmark >2)) based on VirSorter2's SOP[71] and *q*-value > =0.95 using DeepVriFinder. We collected metagenomic data from various sources, including SRP188615/PRJNA526405[16,17], CNP0000824[14], PRJEB11755[20], PRJNA788462[17], PRJNA775062[19], PRJEB22062[18], PRJCA009609[9], PRJEB44118[72], and 70 samples collected in our laboratory. In total, 4,650 samples were used in our study to extract the viral contigs. Each sample was processed using the above pipeline, and then all the viral contigs were combined for subsequent analysis. To ensure the reliability of the results, all subsequent analyzes were conducted using vOTUs of at least medium quality (>50% completeness).

Viral contigs cluster and quality control

We applied CheckV v.1.0.1 (database v.1.4) to assess the quality of all viral sequences[26]. All viral contigs were clustered into species-level vOTUs based on 95% average nucleotide identity (ANI) and 85% alignment fraction (AF) using pairwise ANI calculations from the CheckV repository's rapid genome clustering supporting code. All-vs-all local alignments were performed with the BLAST+ package v.2.13.0, with the parameters (-max_target_seqs 10000). Pairwise ANI values were calculated by combining local alignments between sequence pairs using anicalc.py script. UCLUST-like clustering was carried out with the aniclust.py script using the recommended parameters (95% ANI + 85% AF) from MIUVIG. The sequences of vOTUs resulting from clustering were extracted from the viral contigs, and their quality was re-evaluated using CheckV.

Viral taxonomy annotation

Viral taxonomy identification is challenging due to the lack of a specific marker gene for viral sequences and the presence of a large amount of 'dark matter' in various environments. A new ICTV taxonomy was released in 2022 and abolished the previous large proportion of the order *Caudovirales* and the families *Myoviridae*, *Siphoviridae*, and *Podoviridae*[30]. We applied geNomad v.1.2.0[64] to obtain the taxonomy of vOTUs above 50% completeness. The new ICTV taxonomy database (v.214) was used.

Functional annotation, host prediction, and lifestyle prediction

All protein-coding genes of vOTUs were predicted using prodigal-gv v.2.9.0[73]. Genes were annotated based on Dimond searches against protein databases, including eggNOG 5.0 and VOGDB (http://vogdb.org)[47], using EggNOG-mapper with default parameters (Cantalapiedra et al.)[74]. The structured antibiotic resistance gene (SARG) database[48], antibacterial biocide and metal resistance genes (BacMet) database[75], mobile genetic elements (MGEs) database[55], virulence factor database (VFDB)[76], and quorum sensing database[59] were searched against with BLASTp using Diamond v.2.0.15.153 with the specific parameters: --evalue 1e-5 –query-cover 85 –id 80[77]. iPHoP v1.3.3[35] was used for the vOTU host prediction with the Aug_2023 release database and default parameters. We used BACPHLIP v.0.9.6[78] and PhaTYP[79], which use bidirectional encoder representations from transformers (BERT), to determine whether the vOTUs were likely to be temperate or virulent.

Comparison to other viral reference databases

In this study, the vOTUs from the PVD were compared against four reference databases: IMG/VR v.4.0[12], GVD v.2.0[7], MGV v.1.0[8], and GPD v.1.0[6]. To improve computational efficiency and facilitate cross-database comparison, all sequences were filtered to the vOTU species level and above 50% completeness according to the CheckV results before clustering between different databases. The sequences were then combined and clustered using the supporting code in the CheckV repository with the following BLASTn-specific parameters: evalue 1e-5, max-target-seqs 10,000, query coverage 90, and identity 70. We extracted the VCs between different databases, and if the sequences clustered together, it indicated that the vOTU was shared between two databases; otherwise, it meant that they were different. The comparison results were visualized using the UpSetR package[80].

Phylogenetic analyzes

We generated phylogenetic trees for the vOTUs with completeness above 50% with taxonomy, host vertebrates, and archaea. First, we performed multiple sequence alignment for each type of vOTU sequence using MAFFT v.7.515 with the "--auto" model[81] for each type of vOTU sequence. Second, we inferred a concatenated nucleic acid sequence phylogeny from the multiple sequence alignment using FastTree v2.1.11 with the parameters "-nt -gtr"[82]. Finally, the tree was midpoint rooted and visualized using iToL[83].

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Access to the full vOTUs genomes from PVD is provided at https://doi.org/10.6084/m9.figshare.22671076.v1, and above medium-quality vOTUs genomes is also provided at https://doi.org/10.6084/m9.figshare.25895290. Any further requests for data should be directed to the corresponding authors.

Code availability

Supporting code is provided at https://github.com/mijiandui/metav.

Received: 17 December 2023; Accepted: 16 August 2024;
Published online: 29 August 2024

References

1. Chen, C. et al. Prevotella copri increases fat accumulation in pigs fed with formula diets. *Microbiome* **9**, 175 (2021).
2. Hu, J. et al. Core-predominant gut fungus *Kazachstania slooffiae* promotes intestinal epithelial glycolysis via lysine desuccinylation in pigs. *Microbiome* **11**, 31 (2023).
3. Wang, G. et al. *Lactobacillus reuteri* improves the development and maturation of fecal microbiota in piglets through mother-to-infant microbe and metabolite vertical transmission. *Microbiome* **10**, 211 (2022).
4. Yang, H. et al. ABO genotype alters the gut microbiota by regulating GalNAc levels in pigs. *Nature* **606**, 358–367 (2022).
5. Shkoporov, A. N. et al. Viral biogeography of the mammalian gut and parenchymal organs. *Nat. Microbiol.* **7**, 1301–1311 (2022).
6. Camarillo-Guerrero, L. F., Almeida, A., Rangel-Pineros, G., Finn, R. D. & Lawley, T. D. Massive expansion of human gut bacteriophage diversity. *Cell* **184**, 1098–1109 e9 (2021).
7. Gregory, A. C. et al. The gut virome database reveals age-dependent patterns of virome diversity in the human gut. *Cell Host Microbe* **28**, 724–740 e8 (2020).
8. Nayfach, S. et al. Metagenomic compendium of 189,680 DNA viruses from the human gut microbiome. *Nat. Microbiol.* **6**, 960–970 (2021).
9. Wu, R. et al. Early life dynamics of ARG and MGE associated with intestinal virome in neonatal piglets. *Vet. Microbiol.* **274**, 109575 (2022).
10. Edgar, R. C. et al. Petabase-scale sequence alignment catalyses viral discovery. *Nature* **602**, 142–147 (2022).
11. Gregory, A. C. et al. Marine DNA viral macro- and microdiversity from pole to pole. *Cell* **177**, 1109–1123 e14 (2019).
12. Camargo, A. P. et al. IMG/VR v4: an expanded database of uncultivated virus genomes within a framework of extensive functional, taxonomic, and ecological metadata. *Nucleic Acids Res* **51**, D733–d743 (2023).

13. Yan, M. et al. Interrogating the viral dark matter of the rumen ecosystem with a global virome database. *Nat. Commun.* **14**, 5254 (2023).
14. Chen, C. et al. Expanded catalog of microbial genes and metagenome-assembled genomes from the pig gut microbiome. *Nat. Commun.* **12**, 1106 (2021).
15. Gaio, D. et al. Hackflex: low-cost, high-throughput, Illumina Nextera Flex library construction. *Microb. Genom.* **8**, 000744 (2022).
16. Gaio, D. et al. Post-weaning shifts in microbiome composition and metabolism revealed by over 25,000 pig gut metagenome-assembled genomes. *Microb. Genom.* **7**, 000501 (2021).
17. Gaire, T. N. et al. The impacts of viral infection and subsequent antimicrobials on the microbiome-resistome of growing pigs. *Microbiome* **10**, 118 (2022).
18. Luiken, R. E. C. et al. Farm dust resistomes and bacterial microbiomes in European poultry and pig farms. *Environ. Int.* **143**, 105971 (2020).
19. Tao, S., Zou, H., Li, J. & Wei, H. Landscapes of enteric virome signatures in early-weaned piglets. *Microbiol. Spectr.* **10**, e0169822 (2022).
20. Xiao, L. et al. A reference gene catalogue of the pig gut microbiome. *Nat. Microbiol.* **1**, 16161 (2016).
21. Zhang, S. et al. Dissemination of antibiotic resistance genes (ARGs) via integrons in *Escherichia coli*: A risk to human health. *Environ. Pollut.* **266**, 115260 (2020).
22. Guo, J. et al. VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. *Microbiome* **9**, 37 (2021).
23. Ren, J. et al. Identifying viruses from metagenomic data using deep learning. *Quant. Biol.* **8**, 64–77 (2020).
24. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. metaSPAdes: a new versatile metagenomic assembler. *Genome Res.* **27**, 824–834 (2017).
25. Li, D., Liu, C. M., Luo, R., Sadakane, K. & Lam, T. W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674–1676 (2015).
26. Nayfach, S. et al. CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nat. Biotechnol.* **39**, 578–585 (2021).
27. Khan Mirzaei, M. et al. Challenges of studying the human virome – relevant emerging technologies. *Trends Microbiol.* **29**, 171–181 (2021).
28. Bowers, R. M. et al. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat. Biotechnol.* **35**, 725–731 (2017).
29. Roux, S. et al. Minimum information about an Uncultivated Virus Genome (MIUViG). *Nat. Biotechnol.* **37**, 29–37 (2019).
30. Turner, D. et al. Abolishment of morphology-based taxa and change to binomial species names: 2022 taxonomy update of the ICTV bacterial viruses subcommittee. *Arch. Virol.* **168**, 74 (2023).
31. Breitbart, M. et al. Metagenomic analyses of an uncultured viral community from human feces. *J. Bacteriol.* **185**, 6220–6223 (2003).
32. Liang, G. & Bushman, F. D. The human virome: assembly, composition and host interactions. *Nat. Rev. Microbiol.* **19**, 514–527 (2021).
33. Kauffman, K. M. et al. Resolving the structure of phage-bacteria interactions in the context of natural diversity. *Nat. Commun.* **13**, 372 (2022).
34. Khan Mirzaei, M. & Deng, L. New technologies for developing phage-based tools to manipulate the human microbiome. *Trends Microbiol.* **30**, 131–142 (2022).
35. Roux, S. et al. iPHoP: An integrated machine learning framework to maximize host prediction for metagenome-derived viruses of archaea and bacteria. *PLoS Biol.* **21**, e3002083 (2023).
36. Bergamaschi, M. et al. Gut microbiome composition differences among breeds impact feed efficiency in swine. *Microbiome* **8**, 110 (2020).
37. Ramayo-Caldas, Y. et al. Phylogenetic network analysis applied to pig gut microbiota identifies an ecosystem structure linked with growth traits. *ISME J.* **10**, 2973–2977 (2016).
38. Nishijima, S. et al. Extensive gut virome variation and its associations with host and environmental factors in a population-level cohort. *Nat. Commun.* **13**, 5252 (2022).
39. Díaz, C., Celer, V. & Frébort, I. The main DNA viruses significantly affecting pig livestock. *J. Vet. Res.* **65**, 15–25 (2021).
40. Klangprapan, S., Weng, C. C., Huang, W. T., Li, Y. K. & Choowongkomon, K. Selection and characterization of a single-chain variable fragment against porcine circovirus type 2 capsid and impedimetric immunosensor development. *ACS Omega* **6**, 24233–24243 (2021).
41. Chen, Y. et al. Structure of the error-prone DNA ligase of African swine fever virus identifies critical active site residues. *Nat. Commun.* **10**, 387 (2019).
42. Abbas, A. A. et al. *Redondoviridae*, a family of small, circular DNA viruses of the human oro-respiratory tract associated with periodontitis and critical illness. *Cell Host Microbe* **25**, 719–729.e4 (2019).
43. Liu, X. et al. Emergence of porcine circovirus-like viruses associated with porcine diarrheal disease in China. *Transbound. Emerg. Dis.* **68**, 3167–3173 (2021).
44. Varsani, A. & Krupovic, M. *Smacoviridae*: a new family of animal-associated single-stranded DNA viruses. *Arch. Virol.* **163**, 2005–2015 (2018).
45. Li, R., Wang, Y., Hu, H., Tan, Y. & Ma, Y. Metagenomic analysis reveals unexplored diversity of archaeal virome in the human gut. *Nat. Commun.* **13**, 7978 (2022).
46. Al-Shayeb, B. et al. Clades of huge phages from across Earth's ecosystems. *Nature* **578**, 425–431 (2020).
47. Huerta-Cepas, J. et al. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res* **47**, D309–D314 (2018).
48. Yin, X. et al. ARGs-OAP v3.0: antibiotic-resistance gene database curation and analysis pipeline optimization. *Engineering* **27**, 234–241 (2023).
49. Chen, M. L. et al. Viral community and virus-associated antibiotic resistance genes in soils amended with organic fertilizers. *Environ. Sci. Technol.* **55**, 13881–13890 (2021).
50. Hu, J. et al. Characterizing the gut phageome and phage-borne antimicrobial resistance genes in pigs. *Microbiome* **12**, 102 (2024).
51. Karkman, A., Parnanen, K. & Larsson, D. G. J. Fecal pollution can explain antibiotic resistance gene abundances in anthropogenically impacted environments. *Nat. Commun.* **10**, 80 (2019).
52. Moon, K. et al. Freshwater viral metagenome reveals novel and functional phage-borne antibiotic resistance genes. *Microbiome* **8**, 75 (2020).
53. Billaud, M. et al. Analysis of viromes and microbiomes from pig fecal samples reveals that phages and prophages rarely carry antibiotic resistance genes. *ISME Commun.* **1**, 55 (2021).
54. Enault, F. et al. Phages rarely encode antibiotic resistance genes: a cautionary tale for virome analyses. *ISME J.* **11**, 237–247 (2017).
55. Pärnänen, K. et al. Maternal gut and breast milk microbiota affect infant gut antibiotic resistome and mobile genetic elements. *Nat. Commun.* **9**, 3891 (2018).
56. Karvelis, T. et al. Transposon-associated TnpB is a programmable RNA-guided DNA endonuclease. *Nature* **599**, 692–696 (2021).
57. Liang, J. et al. Identification and quantification of bacterial genomes carrying antibiotic resistance genes and virulence factor genes for

aquatic microbiological risk assessment. *Water Res* **168**, 115160 (2020).

58. Li, N. et al. Cu and Zn exert a greater influence on antibiotic resistance and its transfer than doxycycline in agricultural soils. *J. Hazard. Mater.* **423**, 127042 (2022).
59. Wu, S. et al. Machine learning aided construction of the quorum sensing communication network for human gut microbiota. *Nat. Commun.* **13**, 3079 (2022).
60. Chen, Y. M. et al. RNA viromes from terrestrial sites across China expand environmental viral diversity. *Nat. Microbiol.* **7**, 1312–1323 (2022).
61. Dominguez-Huerta, G. et al. Diversity and ecological footprint of Global Ocean RNA viruses. *Science* **376**, 1202–1208 (2022).
62. Neri, U. et al. Expansion of the global RNA virome reveals diverse clades of bacteriophages. *Cell* **185**, 4023–4037.e18 (2022).
63. Zayed, A. A. et al. Cryptic and abundant marine viruses at the evolutionary origins of Earth's RNA virome. *Science* **376**, 156–162 (2022).
64. Camargo, A. P. et al. Identification of mobile genetic elements with geNomad. *Nat. Biotechnol*. https://doi.org/10.1038/s41587-023-01953-y (2023).
65. Bolduc, B. et al. iVirus 2.0: Cyberinfrastructure-supported tools and data to power DNA virus ecology. *ISME Commun.* **1**, 77 (2021).
66. Banerjee, S. & van der Heijden, M. G. A. Soil microbiomes and one health. *Nat. Rev. Microbiol.* **21**, 6–20 (2022).
67. Ko, K. K. K., Chng, K. R. & Nagarajan, N. Metagenomics-enabled microbial surveillance. *Nat. Microbiol.* **7**, 486–496 (2022).
68. Carlson, C. J. et al. Climate change increases cross-species viral transmission risk. *Nature* **607**, 555–562 (2022).
69. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
70. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
71. Guo, J., Vik, D., Pratama, A., Roux, S., Sullivan, M. *Viral sequence identification SOP with VirSorter2 V.3. dx.*https://doi.org/10.17504/protocols.io.bwm5pc86. (2021).
72. Zhang, Y. et al. Porcine gut microbiota in mediating host metabolic adaptation to cold stress. *NPJ Biofilms Microbiomes* **8**, 18 (2022).
73. Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinforma.* **11**, 119 (2010).
74. Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Mol. Biol. Evol.* **38**, 5825–5829 (2021).
75. Pal, C., Bengtsson-Palme, J., Rensing, C., Kristiansson, E. & Larsson, D. G. J. BacMet: antibacterial biocide and metal resistance genes database. *Nucleic Acids Res.* **42**, D737–D743 (2013).
76. Liu, B. et al. 2022: a general classification scheme for bacterial virulence factors. *Nucleic Acids Res* **50**, D912–D917 (2021).
77. Buchfink, B., Reuter, K. & Drost, H. G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat. Methods* **18**, 366–368 (2021).
78. Hockenberry, A. J. & Wilke, C. O. BACPHLIP: predicting bacteriophage lifestyle from conserved protein domains. *PeerJ* **9**, e11396 (2021).
79. Shang, J., Tang, X. & Sun, Y. PhaTYP: predicting the lifestyle for bacteriophages using BERT. *Brief. Bioinform.* **24**, bbac487 (2022).
80. Conway, J. R., Lex, A. & Gehlenborg, N. UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics* **33**, 2938–2940 (2017).
81. Katoh, K., Misawa, K., Kuma, K. I. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**, 3059–3066 (2002).
82. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree: Computing large minimum evolution trees with profiles instead of a distance matrix. *Mol. Biol. Evol.* **26**, 1641–1650 (2009).
83. Letunic, I. & Bork, P. Interactive Tree of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res* **49**, W293–w296 (2021).

Acknowledgements

This study was supported by the Major Science and Technology Project of Gansu Province (23ZDNA007) and the International Collaboration 111 Program (BP0719040). We would like to express our gratitude to the Supercomputing Center of Lanzhou University for their valuable support in the computation works.

Author contributions

J.M. designed the study, plotted the figures, and wrote the manuscript. J.M., Y.L., and Y.Z. collected the metagenomic data. Y. Y. provided suggestions. J.M., and C.M. performed the metagenomic analysis. X.J., R.L., and H.Z. contributed to the scientific discussion and preparation of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41522-024-00554-0.

Correspondence and requests for materials should be addressed to Jiandui Mi, Ruijun Long or Haixue Zheng.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.