

Bioinformatic approaches to pig gut phagobiome (phageome) analysis: tools, applications, and current state of research

Anna Miastowska¹, Adrian Augustyniak^{2,3}, Bartłomiej Grygorcewicz⁴, Paweł Nawrotek¹✉

¹Department of Microbiology and Biotechnology, Faculty of Biotechnology and Animal Sciences,
West Pomeranian University of Technology in Szczecin, 70-311 Szczecin, Poland

²Department of Chemical and Process Engineering, Faculty of Chemical Technology and Engineering,
West Pomeranian University of Technology in Szczecin, 71-065 Szczecin, Poland

³Center for Advanced Materials and Manufacturing Process Engineering (MATPRO),
West Pomeranian University of Technology in Szczecin, 71-065 Szczecin, Poland

⁴Department of Genomics and Forensic Genetics, Faculty of Medicine, Pomeranian Medical University in Szczecin,
70-111 Szczecin, Poland
pawel.nawrotek@zut.edu.pl

Received: February 24, 2026

Accepted: September 9, 2026

Abstract

The pig intestine contains a diverse bacteriophage community that helps regulate the gut microbiome and host health. Growing interest in phage therapy reflects its potential as a cost-effective, targeted approach to modulating microbial communities and treating diseases, particularly those involving antibiotic-resistant bacteria. This review proposes an alternative operational definition of the phagobiome that integrates the phageome with host responses and summarises current knowledge of swine gut phages. It includes well-studied *Caudoviricetes* as well as less abundant groups such as crAss-like phages and jumbo phages carrying clustered, regularly interspaced short palindromic repeat spacers. Differences across pig breeds and age groups are outlined, highlighting data gaps. Current knowledge of gut phages and their interactions with the host remains limited, prompting the development of new bioinformatic tools. The paper also critically compares bioinformatic tools for phageome analysis, supporting method selection and future research on core phages shared within pig populations and their potential applications.

Keywords: swine, pig gut bacteriophage, machine learning, bioinformatics, faecal virome transplantation.

Introduction

Diseases caused by antibiotic-resistant microorganisms are difficult to treat with the available drugs. Horizontal gene transfer within the gut microbiome can convert naturally occurring bacteria into antibiotic-resistant strains through the acquisition of antibiotic-resistance genes (ARGs) (59, 86). These challenges have stimulated the search for alternative treatments, such as phage therapy (applied) or clustered, regularly interspaced short palindromic repeat (CRISPR)-based antimicrobials (largely experimental) (48).

Studies (46, 102) have suggested faecal virome transplantation (FVT) as a promising novel therapeutic approach. In principle, limiting transplants to specially selected phages open possibilities for their application in

phage therapy or microbiome engineering (105). To achieve that, effective methods for their taxonomic classification and understanding their biological function must be established. The gut phagobiome, defined here as the assemblage of bacteriophages and prophages within a given biological environment, is functionally integrated with the microbiome and may influence its stability, activity, and interactions with the host (75). Its complexity, combined with the rapid development of artificial intelligence methods in human and veterinary medicine, has led to the search for bioinformatic tools that enable sequence-based taxonomic identification of gut phages, including similarity and machine learning-based approaches; prediction of phage-host interactions; and assessment of phage-host interactions within the entire gut microbial community (18, 137).

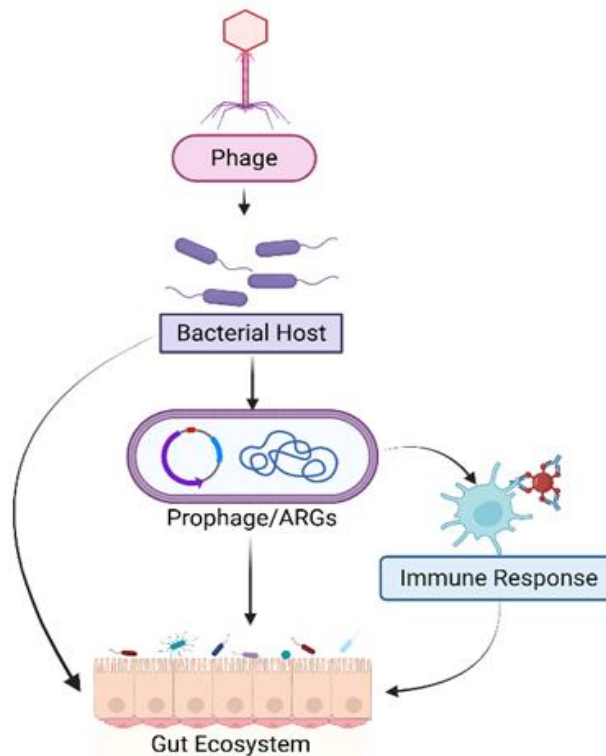


Fig. 1. Diagram illustrating the role of the phagobiome in the host's gut. ARG – antibiotic resistance gene

Although “phageome” is the commonly used term, “phagobiome” (Fig. 1) more accurately reflects the functional integration of bacteriophages with the microbiome and the host, and the authors therefore propose this term (42, 107).

This study presents a comprehensive review of the gut phagobiome and the bioinformatic tools used to analyse it, covering each stage of analysis and highlighting dedicated tools. It is intended to guide researchers seeking to expand knowledge of swine phageomes, and reviews phage identification, taxonomic classification, host–phage interaction prediction, functional characterisation and data visualisation for their application in therapy, and microbiome engineering.

Publications landscape

In recent years, interest in the mammalian gut virome, including that of pigs, has increased significantly. To illustrate how research in the pig gut phageome has developed over time, a literature review was conducted using the Scopus, PubMed and Web of Science databases. A graph representing the number of publications per year was created in RStudio (84) using the ggplot2 package in R (85, 123) (Fig. 2).

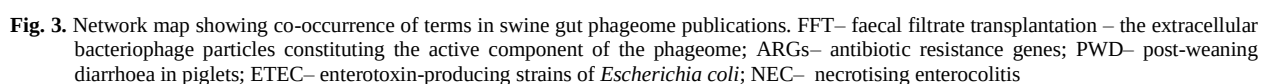
A total of 397 publications were identified across the three databases. Details are presented in Table 1. Data were consolidated in RStudio, and duplicates were removed based on titles using the dplyr package and manual editing (124). Publications were filtered using query keywords associated with pigs, gut, and phage-

related terms (“pigs” OR “swine” OR “piglets” AND “gut” AND “phage” OR “FVT” OR “virome” or “phageome” OR “phagome” OR “phagobiome” OR “bacteriophage”). After merging and removing duplicates, 244 records were obtained.

Table 1. Number of publications in selected databases

Database	Number of publications
Web of Science	240
Scopus	88
PubMed	69
Total before deduplication	397
Total after deduplication	244

In addition, to visualise how key topics related to the gut phageome of swine have changed over the years, a co-occurrence map was created. The VOSviewer software was used to conduct an analysis of publications from 1947–2025 to establish a co-occurrence network (28) (Fig. 3). The resulting co-occurrence network contains 78 terms, after threshold adjustments. The map indicates that topics appearing before 2018 primarily focused on food, animals, strains and humans. Over the years, terms related to treatment, viruses, antibiotic resistance, metagenomics and diseases have become more frequent. Around 2022, there was an increase in interest in topics such as FFT (faecal filtrate transplantation; free-phage fraction of faecal filtrate), gut virome, and phage cocktail.



Phagobiome in pigs: role of gut phages

Phages have been suggested to influence the immune system of mammals by modulating inflammation, innate and adaptive immunity, cell growth, and bacterial clearance (102, 109). In some cases, they can interact directly with bacteria on the mucosal surface, transferring genes related to polysaccharide, carbohydrate, and toxin metabolism. Gut phages colonise the mucosal layer, forming a barrier that prevents bacterial adhesion to the intestinal wall (15). As determined by the life cycle of viruses, intestinal phages can be functionally classified as virulent, temperate (12), pseudolysogenic or associated with chronic infection (111). Virulent phages multiply within host cells, producing virions and causing cell lysis (94). Temperate phages can undergo both a lytic and a lysogenic cycle, in which the phage enters a dormant state by integrating into the host genome to form a prophage (94, 102). Such phages can transfer bacterial genes, influencing the host metabolism. Pseudolysogeny refers to a state in which a phage genome remains non-integrated within the host cell until conditions allow entry into either the lytic or lysogenic cycle. Chronic infection is characterised by continuous virion production and release without lysis of the bacterial host cell, ensuring phage propagation (111). Some prophages present inside bacterial cells can prevent infection by other viruses through a process called superinfection exclusion (SIE), which prevents reinfection of a bacterium by competing phages. In certain systems, prophages can influence bacterial mobility, biofilm formation, adhesion and conjugation, thereby potentially affecting bacterial virulence and overall function. This shows that phages capable of SIE may therefore be effective in phage therapy (21). In this context, SIE constitutes a dynamic regulatory mechanism through which prophages modulate bacterial phenotypes in response to phage–phage interactions.

Phagobiome in pigs: phagobiome assembly within the microbiome

The development of the pig microbiome is influenced by many factors, such as the environment, contact with other animals, diet, breed and age (114), with the possibility of these factors being interrelated. Early changes in the microbiome may be associated with the development of the gastrointestinal tract in piglets. The greatest shifts are observed during the weaning phase, after which the microbiome tends to stabilise in two weeks (27). Diet-related changes in the microbiome have the greatest impact at around three months of age, when the diet transitions from maternal milk to plant carbohydrates, resulting in an increase in dietary carbohydrates and a decrease in protein intake (27, 114). Changes in the abundance and composition of gut microbiome shape the diversity and dynamics of phage populations.

In a study by Hu *et al.* (42), conducted on 112 pigs from seven breeds, the greatest differences in viral operational taxonomic unit (vOTU) diversity were observed between Chinese native pigs (Tibetan miniature, Laiwu, Shaziling, Congjiang miniature,

Huanjiang miniature and Ningxiang) and commercial pigs (Duroc × (Landrace × Yorkshire)), reflecting fundamental differences in the composition and diversity of their bacterial hosts. Gut phage populations in finishing pigs indicated higher alpha diversity compared with those in weaned piglets. Analyses showed age-related changes in the pig gut microbiome, with phage diversity increasing with growth stage. The finishing swine phage population and that of the weaned piglets were observed to cluster together, which may explain differences in phage populations. Interaction among pigs can facilitate the exchange of microorganisms, reinforcing the presence of certain phages associated with specific bacterial species and influencing their abundance and population dynamics. Weaned piglets, particularly the Tibetan miniature breed, had a higher proportion of core phages. In contrast, finishing pigs, especially the commercial breed, had a lower proportion of core phages and a higher proportion of unique phages, suggesting dynamic restructuring of the gut phageome during the pig's lifespan (42).

Despite changes in the microbiome during piglet growth and development, it remains influenced by maternal transmission of it as the initial microbiome (106). This maternal microbiota composition shapes the efficiency of bacterial species colonisation, modifying microbial diversity, and may influence the composition of the phage community in the pig gut.

Phagobiome in pigs: characterisation of gut bacteriophages

The majority of identified tailed phages in pigs belong to the class *Caudoviricetes* (42, 71), which includes the families most commonly appearing in pigs: *Ackermannviridae*, *Straboviridae*, *Peduviridae*, *Zierdtviridae*, *Drexelviriidae* and *Herelleviridae*. The previous classification for these phages was in the order *Caudovirales*, and the families *Podoviridae*, *Siphoviridae* and *Myoviridae* were included (42).

A study (100) reported that *Caudoviricetes* phages in the pig colon were the most abundant in several identification experiments. In addition to bacteriophages, eukaryotic viruses were also present, with *Parvoviridae* predominating in the mucosa of the small intestine, while *Circoviridae*, *Astroviridae*, *Caliciviridae* and *Parvoviridae* predominated in the large intestine. Analyses of shared viral contig sets showed that bacteriophages of the family *Microviridae* were present in the cecum and colon. In intestinal metagenomes, eukaryotic viruses and bacteriophages can be difficult to distinguish bioinformatically because different sampling and sequencing strategies capture different viral subsets. Additionally, the location and presence of prophages integrated into bacterial genomes cause phage and eukaryotic signal to co-occur in the same contig sets, because of chimeric contigs or annotation inaccuracies.

A meta-analysis conducted by Mi *et al.* (71) showed that the most common hosts for phages of the class *Caudoviricetes* are bacteria belonging to the phyla *Firmicutes* and *Bacteroidetes*, including genera such as

Prevotella, *Bacteroides*, *Vescimonas* and *Candidatus Faecousia*. Although Mi *et al.* use the traditional phylum names, the currently accepted nomenclature refers to these phyla as *Bacillota* and *Bacteroidota*. For consistency, the updated taxonomy will be used throughout this work. Phage host-prediction bioinformatic analyses combined with Mi *et al.*'s meta-analysis indicated that some were likely to infect bacteria involved in efficient nutrient metabolism, such as *Blautia*, *Faecalibacterium*, *Lactobacillus* and *Ruminococcus*, potentially indirectly influencing pig productivity and health. Furthermore, the absence of *Asfarviridae* in the analyses derived from healthy pigs indicates that these viruses are not part of the core pig phageome but rather represent occasional or invasive species (71).

In recent publications (116, 131), viruses similar to human cross-assembly phages (crAssphages) were detected in pigs. These phages co-occur with *Bacteroidota* gut communities, which have been associated with influencing metabolism, weight gain and immune function. The intestinal crAssphage was used as a marker confirming contamination with human faeces in publications investigating the source of wastewater pollution (61). In studies (1, 69), crAssphage was also detected in samples containing animal faeces. According to the classification of the International Committee on Taxonomy of Viruses (ICTV), phages similar to crAssphages are divided into clusters: alpha (*Intestiviridae*), beta (*Steigviridae*), gamma (*Crevaviridae*) and delta (*Suoliviridae*). In addition to the official classification, there are two other groups: epsilon and zeta (118). Studies conducted on pigs have detected the highest number of phages belonging to the beta cluster, and smaller numbers belonging to the alpha, delta and zeta clusters (116, 131). In humans, most crAssphages belong to the *Intestiviridae* family (116). The hosts for pig crAssphages are most often bacteria of the phylum *Bacteroidota*, less often *Firmicutes*. Within these taxa, species of the genus *Prevotella* dominate, with *Parabacteroides* and UBA4372 occurring at lower abundances. The second of those is an unclassified bacterial genus often associated with the order *Bacteroidales* and the family *Rikenellaceae* which is dominant in the rumen of buffalo and positively correlated with their immunity and metabolism. In humans, the primary hosts of crAssphages are bacteria of the genera *Bacteroides*, *Prevotella* and *Parabacteroides*. Pig crAssphages, by interacting with *Prevotella* and *Bacteroides*, can influence fat deposition in animals. The species *Prevotella copri*, which forms the core of the intestinal microbiome, is found in greater amounts in fattening pigs (116).

The intestinal microbiome of pigs also includes prophages, i.e. viruses integrated into the genome of bacterial hosts. Prophages encoding putative auxiliary metabolic genes (AMGs) such as *cobA*, *cobS*, and *cobT* can influence the synthesis of vitamin B12, which is essential for proper metabolism. Another feature of some phages is the presence of SIE genes, which protect the host from infection by other phages. They can also transfer ARGs and virulence factor genes (VFGs)

between bacteria through horizontal gene transfer. The assessment of the frequency of transfer of these genes conducted by Wei *et al.* (118) showed that the highest number of prophages carrying both ARGs and VFGs was detected in *Escherichia coli*, followed by lower numbers in *Citrobacter portucalensis* and *Klebsiella pneumoniae*. The most frequently identified antibiotic resistance genes were associated with multidrug, aminocoumarin and nitroimidazole resistance. This shows that the selection exercised in faecal microbiota transplantation (FMT) for bacterial composition may be insufficient and should also consider the phages embedded into bacterial genomes, as well as the free phage fraction. Concerning virulence factors, genes related to flagella and capsule formation and lipopolysaccharides were the most common. Virulent phage genes targeting specific hosts can be evaluated for their ability to infect pathogenic bacterial strains, to verify their suitability for phage therapy (118). Furthermore, Wei and Chen (117) reported detection of lytic phages that, based on CRISPR spacer matching, are capable of infecting *E. coli* and *Klebsiella pneumoniae*. In the study cited earlier by Wei *et al.* (118), crAss-like prophages from the order *Crassvirales* were found, belonging mainly to the zeta cluster, but also to the alpha and beta. The detection of crAss-like phages in samples and crAssphages in faeces in pigs indicates the presence of crAssphages in the intestines of these animals (1, 69, 116, 131).

A study by Yu *et al.* (131) demonstrated the presence of jumbo phages in pigs. Jumbo phages, distinguished by their extremely large genomes (>200 kb) and large virions, play a key role in environmental microbiology and potential therapeutic applications (132). The discovered jumbo phages belonged to the class *Caudoviricetes* and were found to encode genes involved in bacterial defence modulation, infection, assembly, lysis and phage replication. They were also found to encode genes that increase the immune resistance of their bacterial hosts. Two of the identified jumbo phages encoded CRISPR spacers targeting bacterial genomes reconstructed from metagenomes and spacers against other phages present in the analysed sample. In addition, several CRISPR-Cas systems based on the Cas9, Cas10 and Cas14 proteins were discovered in some jumbo phages. The short 153-amino-acid sequence of Cas14 (one of the smallest known CRISPR effectors) makes this protein a useful gene editing tool, especially for delivery systems with limited payload capacity (131). Despite its short length, Cas14 can perform targeted cuts of single-stranded DNA (13). The presence of the Cas14 in jumbo phage genomes implies that these phages may have developed mechanisms to interact with competing phages, suggesting phage-phage competition.

Phagobiome in pigs: phage therapy and microbiome engineering

The transfer of the entire microbiota from the donor's faeces to a recipient, FMT, change the microbiome beneficially. So far, FMT has been shown to be effective in treating *Clostridioides difficile* infections.

The use of FMT in livestock has reduced susceptibility to diarrhoea and increased average weight gain in piglets. Alongside the intended therapeutic effects of FMT, there are also risks. These include the potential for the introduction of pathogenic or antibiotic-resistant strains to the recipient. In an attempt at mitigation of these difficulties, modifications to FMT have been proposed that focus on obtaining the phage fraction (46).

This obtainment of the phage fraction is FVT, proposed by Jankowski *et al.* (46), and it involves administering 0.22-µm-filtered faecal filtrate from a healthy donor to animals. The faecal virome, which mainly contains gut bacteriophages, is a complex but crucial element of the gut microbiome and is thought to contribute to gut ecosystem stability. Some fractions, *i.e.* those with a beneficial composition of intestinal phages that are active against specific pathogenic microorganisms and do not have undesirable characteristics such as the ability to transfer ARGs or VFGs, can support microbiota homeostasis by maintaining bacterial diversity and thereby enabling the more effective removal of pathogens (20, 94). The results of experimental studies indicated that bacterium-free phage suspension may reduce the incidence of necrotic enteritis in piglets (126). Phages can influence the host positively, but also negatively (Fig. 4).

The goal of phage therapy is to introduce selected phages that specifically target certain bacterial species or strains responsible for particular diseases. Phages can be effective in treating certain infections, making them potentially useful for treating gut pathogens in pigs, such as *Campylobacter* spp., *E. coli*, *Listeria* spp. and *Salmonella* spp., either as a supplement or an alternative to antibiotics. Improved methods for treating pigs translate into better animal health and increased safety of animal food products that could otherwise be a source of infection (48).

Research (53, 130) confirmed the effectiveness of phage therapy in the treatment of chronic wound infections, indicating its use in cases where standard

antibiotic therapy is not effective. Furthermore, studies (78, 82) indicate that phages can be applied not only therapeutically but also prophylactically, potentially to reduce *Staphylococcus aureus* infection by preventing biofilm-related bacterial colonisation on medical surfaces and wounds. Phages can be used in therapies as vectors delivering modified DNA to bacterial cells. This approach is important for the removal or inactivation of antibiotic resistance genes, which indicates the use of phage therapy to increase the effectiveness of antibiotics in treating bacterial infections (119). The CRISPR-Cas9 system can be used as an experimental genome editing tool. Modified phages infect bacteria and deliver CRISPR-Cas9, resulting in the removal of plasmids conferring antibiotic resistance or the disruption of essential genes, thereby weakening or killing the bacteria; however, this approach currently remains at the proof-of-concept stage (14).

Beyond therapeutic applications, phages can also be used as precise modulators within the microbiome. The microbiome represents a highly dynamic and complex ecosystem of coexisting organisms, each creating niches and environmental barriers that shape populations. Microbiome engineering typically introduces single, specially designed organisms rather than whole microbial communities, which may fail to thrive and dominate by the prevention of several ecological barriers, including niche availability, interaction with resident microorganisms and environmental conditions such as pH, temperature, nutrients and oxygen levels. These ecological properties limit the long-term success of targeted engineering in microbiome ecology (2).

The success and safety of phage therapy depend on its high specificity and ability to attack and kill target bacteria, thus avoiding disruption of the entire gut microbiota (122) and no less on other factors such as phage preparation and storage conditions, transport, route of administration, individual phage tolerance to environmental factors and multiplicity of infection (80).

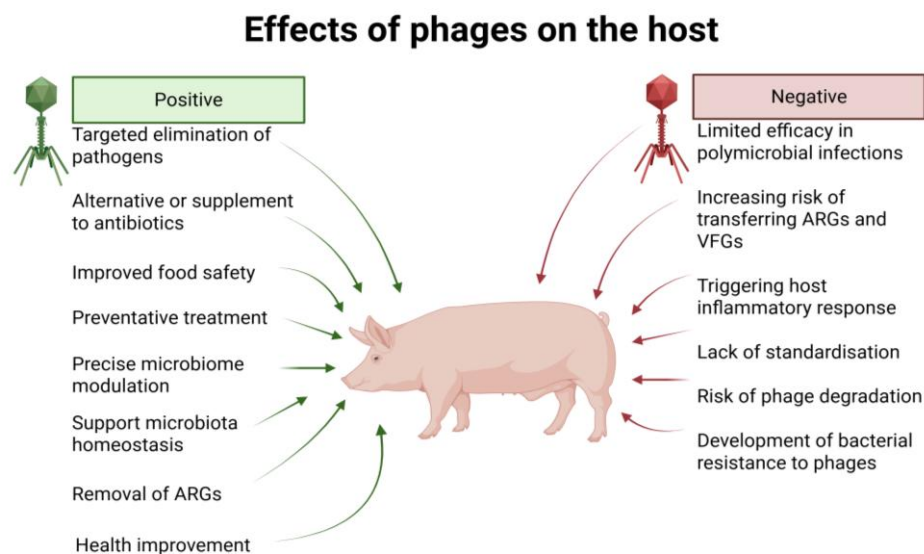


Fig. 4. Key effects of phages on the pig host

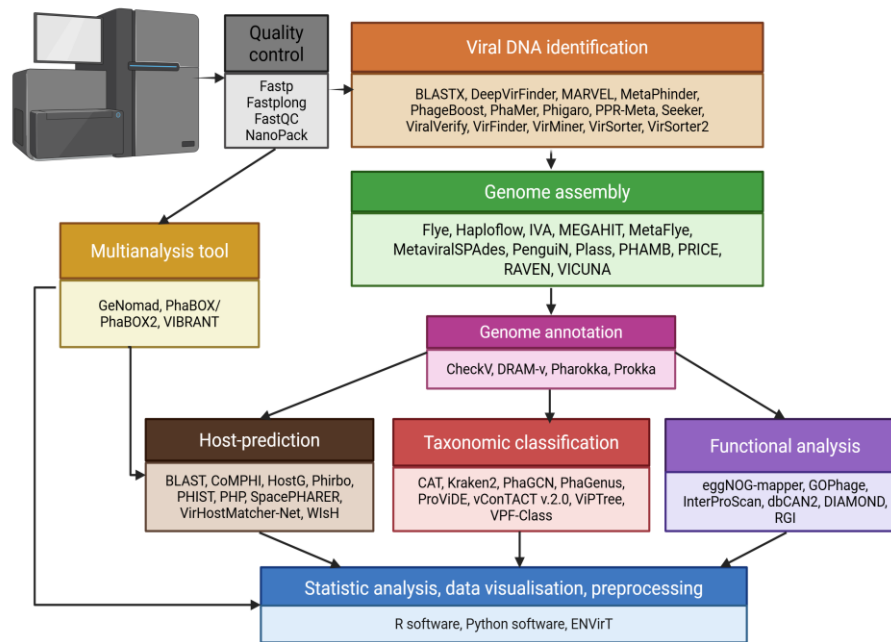


Fig. 5. Selected tools used in phagobiome analyses with assigned analysis stages

Phage-based therapy has several limitations. Because of their high specificity, phages target only selected bacterial genera and species, which restricts their possibilities in treating diseases caused by multiple pathogens. Some phages enter a lysogenic cycle instead of inducing bacterial lysis, which can inhibit other lytic phages and increase the risk of transferring virulence factors and antibiotic resistance genes. Phage-induced lysis of Gram-negative bacteria can result in the release of endotoxins, notably lipopolysaccharides (LPS), and may trigger inflammatory or immune responses. The complex process of phage therapy preparation, together with the lack of clear regulation and standardised protocols for phage isolation, purification, dosing and routes of administration, complicates the therapy's standardisation, quality control and efficiency evaluation. Furthermore, bacteria can develop resistance to phages through defence mechanisms, including CRISPR-Cas systems, which reduces treatment effectiveness (64). Phages may also be rapidly degraded or eliminated by the host's immune system, depending on the route of administration. Oral administration induced only a weak immune response, related to secretory phage-specific IgA, which decreased after termination of phage exposure, allowing phages to be transmitted through the gastrointestinal tract. In contrast, systemic administration can lead to the induction of IgG antibody production only at the systemic level and IgA antibody at both systemic and secretory levels (68).

Phages as tools for targeted microbiome modulation are currently used in *in vitro* systems and animal models, with limited clinical data. Most applications of phage therapy have focused on the elimination of specific pathogenic bacteria rather than broad microbiome reshaping, and further well-designed

randomised controlled clinical trials are needed to validate their use as microbiome-modulating agents (87).

Bioinformatic tools

Analysis of the pig gut phagobiome from metagenomic data requires a multi-stage pipeline because of the lack of universal markers and high fragmentation of contigs. The high diversity of phages and the absence of universal phylogenetic markers or conserved regions as found in bacteria make the analysis complex and highly dependent on assembly quality, contig length and the availability of reference genomes. This necessitates the use of bioinformatic tools. The stages of phageome characterisation are raw data processing, sequence identification and assembly, genome annotation, taxonomic classification, functional analysis, detection of phage–host interaction and statistical analyses (6). Different tools are used in different stages of these analyses (Fig. 5). A brief description of the tools and their sources with references is available in Supplementary Table S1.

Bioinformatic tools: data quality control

Raw data obtained from sequencing (e.g. Illumina and Nanopore) can be trimmed and filtered to remove contaminants, low-quality reads, adapters and technical artifacts (121). For this purpose, tools such as fastp and fastplong can be used (24, 25), the first being for short-read data, including those from Illumina, and the second being a fast FASTQ preprocessor designed for long-read sequencing data, such as from Nanopore. These tools enable quality and length filtering, read trimming and output splitting (24, 25). Another tool, FastQC, checks

the quality control on raw short-read data, and a complementary one, NanoPack, does so for long reads, providing summary statistics and diagnostic plots describing read quality (7, 26).

Bioinformatic tools: viral DNA identification

Following the initial filtering stage, viral sequences are classified from the reads. The low ratio of viral biomass to total DNA causes viral sequences to be underrepresented in bulk metagenomes, and background contamination from eukaryotic or bacterial host-derived DNA can mask this DNA (49). The risk exists of accidentally removing viral sequences, and precautions such as host-filtering or soft masking are needed. Phages that integrate into the host genomes, viruses attached to cells, or those in a pseudo lysogenic state are particularly challenging to properly identify. However, prophages can be further detected and removed if needed using dedicated programs such as Phigaro and PhageBoost. Phigaro uses an algorithm based on hidden Markov models (HMMs) and prokaryotic virus orthologous groups to predict open reading frames (ORFs) belonging to prophages (103). PhageBoost uses generalised ML models that focus on differentiating viral gene features from background (101).

Integrated prophages may carry important bacterial traits, which can be considered part of phagobiome, the microbiome, or a mixture of both. Viral genome identification relies on different approaches, including homology-based methods, *k*-mer-based strategies and machine learning models (41).

Sequence similarity-based methods identify viral DNA based on the homology of phage proteins to data in reference databases (121). Typical tools in this category are basic local alignment search tool, translated nucleotide version (BLASTX) and VirSorter, the first of which only achieves high accuracy for known viruses and works slowly on larger sets (4, 36, 91). The limitation of these methods is the frequent lack of data in databases for poorly known viruses, which may lead to these viruses being overlooked.

Methods based on *k*-mers compare repeating nucleotide patterns of length *k* derived from reference databases with the sample sequence (67). Using fragments rather than the entire genome accelerates analysis. These methods can handle genome rearrangements, but matching effectiveness depends on the *k*-mer length (121). One tool employing the *k*-mer method is Kraken2 (125), the performance of which depends on how well the relevant genomes are represented in the database and was noted to do well on mock datasets but require substantial memory resources.

Machine learning (ML) approaches combine the ability to process large datasets with pattern recognition. Models are trained to identify features such as *k*-mer composition, codon bias related to the usage of specific codon pairs and oligonucleotide frequency patterns, especially cytosine–guanine phosphodiester bond bias (74). Machine learning-based tools have the potential to identify new viral species.

Some tools combine more than one approach. DeepVirFinder and PPR-Meta (phage and plasmid recognizer for metagenomes) use deep learning (DL) models along with *k*-mer features; VirFinder relies on *k*-mer and machine learning; Seeker combines DL and homology-based methods; and VIBRANT (virus identification by iterative ANnotation) (54), VirSorter2 and ViralVerify integrate homology-based approaches with ML (40, 127). DeepVirFinder is an improved version of VirFinder, achieving higher prediction accuracy for short viral sequences, which makes it particularly useful for metagenomic data. It bases its modelling on convoluted neural networks (CNNs) (88). PPR-Meta calculates the likelihood that metagenomic sequence originates from phage, plasmid or chromosome, also using CNNs to improve prediction accuracy for short fragments. In comparisons with VirFinder, cBar, PlasFlow and VirSorter, PPR-Meta archived higher precision and accuracy (30). Seeker distinguishes bacterial from phage genomes directly from sequence data, achieving moderate precision and recall. It is based on long short-term memory, a type of recurrent neural network which makes possible the analysis of complete DNA sequences rather than short fragments. Seeker can capture distant dependencies within genomic sequences (11).

MetaPhinder identifies metagenomic contigs by using BLAST searches against a reference database, combining the advantages of homology-based and *k*-mer approaches (52). A specialised tool is MARVEL (metagenomic analyses and retrieval of viral elements) is an ML based method for detecting viruses from the order *Caudovirales* (currently classified as the class *Caudoviricetes*), allowing the identification of phage sequences in metagenomic datasets (5).

VirMiner can be applied to raw reads and includes quality control steps. It demonstrates moderate precision and sensitivity and uses a random forest algorithm trained on metagenomic data from the gut microbiome (135).

VirSorter2 predicts a wide range of viral groups, including dsDNA, ssDNA phages and virophages, and also RNA viruses. It achieved high accuracy detection of viruses, including those that are underrepresented in reference databases (39). Viral RNA in DNA metagenomes is marginal but can be refined using specific sequencing conditions and data.

PhaMer is based on a Transformer model that learns proteins' compositions using a dictionary of protein clusters and their positional information within sequence. It then distinguishes bacterial from phages contigs and has been shown to outperform other methods, particularly in the identification of novel viruses (99).

Several recent benchmarking studies have compared performance of viral identification tools, including DeepVirFinder, Kraken2, MetaPhinder, PPR-Meta, Seeker, VIBRANT, ViralVerify, VirFinder, VirSorter, VirSorter2 and PhaMer (41, 99, 127). In summary, these tests indicated that tool performance depended on the length of the contigs, the type of data, the purpose of the

analysis and the novelty of the viruses. Homology-based methods tend to perform better for viruses present in reference database, whereas machine learning approaches are more suited for detecting novel viruses (127).

Bioinformatic tools: genome assembly

Several genome assembly strategies are currently pursued. The first is *de novo* assembly, implemented in two ways: based on de Bruijn graphs, or based on overlap graphs as part of the overlap-layout-consensus (OLC) approach. In the OLC method, reads are represented as nodes in a graph, and edges are created based on overlapping fragments of entire sequences. The assemblers of this type used for single viral populations are Vicuna, PRICE (paired-read iterative contig extension) and IVA (iterative virus assembler). In contrast, for de Bruijn graphs, sequences are connected through overlapping *k*-mers. Problems with *de novo* assemblers arise from uneven coverage and population variability, which are common in viral and metagenomic datasets. Modern, second-generation assemblers based on modified de Bruijn graphs, such as MEGAHIT (60) and MetaSPAdes (metagenomic St. Petersburg genome Assembler), implement strategies to simplify the graph, by dividing it into subgroups and removing low coverage *k*-mers and local graph structures (89).

The IVA tool is a *de novo* single population virus assembler made for paired-end reads from short-read platforms such as Illumina (44). This approach can result in fragmented assemblies when reconstructing genomes with terminal repeats. Working on paired reads, PRICE increases the size of contigs created from metagenomic data (92). Vicuna is software for short, highly variable reads (128). Haploflow is a genome assembler for short-read sequences, which uses a flow algorithm based on the de Bruijn graph to resolve viral strains (31). Strain reconstruction remains challenging and assemblies are often fragmented or incomplete.

The MetaviralSPAdes pipeline is an extension of MetaSPAdes that aims to reconstruct contigs representing entire viral genomes instead of mere fragments. Such approaches may be key to recovering the genomes of previously unassembled viruses. The program consists of several tools: ViralAssembly, which creates contigs from metagenomic graphs; ViralVerify, which checks whether contigs are of viral origin using a naive Bayes classifier and is capable of detecting large and diverse sets of viral sequences notably faster than BLASTX and achieving higher recall than VirSorter, VirFinder and VirMiner (8); and ViralComplete, which compares the obtained genomes with reference databases to assess whether the entire virus genome or only a fragment has been received. This tool enabled the identification of new viruses, such as phages from the crAssphage family that were previously misclassified as bacterial contigs, as well as other unknown viruses (8).

The PHA MB (PHAgEs from metagenomics binning) tool developed by Johansen *et al.* (50) is a framework for binning phage genomes from

assembled metagenome data. The method combines contig binning with a machine learning-based classifier that distinguishes viral from non-viral genome bins. This framework enables large-scale reconstruction of high-quality viral genomes directly from bulk metagenomes (50).

An approach optimised for viral assembly is the overlap-based assembler Penguin (protEiN-GUIDed nucleotide assembler) developed by Jochheim *et al.* (49). This program assembles viral genomes and 16S ribosomal RNA (rRNA) genes from metagenomic data to distinguish between strains. Unlike de Bruijn-graph assembly programmes such as MEGAHIT or SPAdes, Penguin relies on read-overlap detection combined with a Bayesian statistic, which allows it to separate closely related strains that are often collapsed or fragmented by graph-based approaches. In test datasets, Penguin reconstructed a significantly higher number of complete and near-complete viral genomes and 16S rRNA sequences, particularly under conditions of high strain diversity and low coverage. Although Penguin is more computationally demanding than de Bruijn-graph tools like MEGAHIT or SPAdes, it shows better scalability than other overlap-based tools such as Haploflow and SAVAGE (strain-aware virAl gEnome assembly), which require longer runtime or fail to complete assemblies on large, strain-rich metagenomic datasets. Usage of the Linclust linear clustering algorithm reduces runtime and memory requirements, making Penguin practical for large and complex metagenomic datasets, including those generated by long-read sequencing platforms such as Nanopore (49).

Plas (protein-level assembler) operates with protein sequences derived from short-read sequencing data. It translates reads in all six frames, improving memory efficiency and making it optimal for complex metagenomic or metatranscriptomic datasets (104).

Flye is a *de novo* assembler designed for single-genome assembly from long sequencing reads. It works on a repeat graph, where edges represent genomic segments that may be unique or repetitive, and nodes are repeated boundaries. Unlike de Bruijn-graph-based assemblers, which require exact *k*-mer matches, Flye uses approximate matches, making it well-suited to long reads and tolerant of sequencing error (57). MetaFlye is an extension to Flye that supports metagenomic assembly of long-read sequencing data. It adds repeat-graph modifications that enable it to handle an uneven abundance of species, high levels of repeats in genomes and large data sets typical of metagenomes. In particular, metaFlye uses repeat structure and coverage information to distinguish between different genomes (56).

Another tool for long reads is the *de novo* genome assembler Raven. It operates on the OLC approach and utilises minimisers in combination with MinHash to effectively detect sequencing read overlaps and generate an assembly graph for efficient usage of time and memory (110).

Bioinformatic tools: genome annotation

The next step involves interpreting the obtained sequences and searching for genes, protein domains and virion structure-related elements. Approaches used include sequence homology-based methods, *ab initio* strategies based on statistical methods such as HMMs and other probabilistic models and hybrid methods. Gene predictions combine intrinsic sequence features used in *ab initio* approaches with extrinsic data derived from similarity to viral proteins (29). The tools for this purpose include Prokka, PHANOTATE and GeneMarkS.

Pharokka is an improved version of Prokka (19, 95), combining gene annotation with functional analysis, and is dedicated to phage analysis. It searches for transfer RNA, transfer-messenger RNA and CRISPR in the sequences. It predicts coding sequences based on two methods: PHANOTATE or Prodigal (pROkaryotic dynamIc programming genefinding aLgorithm). The first predicts genes in viral genomes by accounting for gene packing, start codons and short ORFs, while the second is designed for rapid prediction in large metagenomes. Coding sequences are compared to the PHROGs (protein homologous groups) database based on phage protein homology, the CARD (comprehensive antibiotic resistance database) for identifying antibiotic resistance genes, and the VFDB (virulence factor database) for detecting virulence factors (19). However, detection of ARGs and VFs in phage genomes reconstructed from metagenomes carries an elevated risk of false positives due to bacterial contamination.

The previously mentioned VIBRANT is a multitool framework that predicts bacterial and archaeal viruses, annotates viral functions, identifies genes and determines genome completeness. It integrates neural network models with databases such as Pfam (for protein families), VOGDB (Virus Orthologous Groups Database) and the Kyoto Encyclopedia of Genes and Genomes (KEGG) (54). This framework performs well on novel viruses and diverse metagenomes, achieving good predictive performance in benchmark tests (41, 127).

As an extension of DRAM (distilled and refined annotation of metabolism), DRAM-v is a tool designed for annotating viral genomes, particularly for finding AMGs (96). It operates on viral contigs, generated from VirSorter. Prodigal is used for predicting ORFs, and the resulting amino acid sequences are searched for in multiple databases and web servers, including KEGG, Pfam, UniRef90 (clustered sequences from UniProt), dbCAN (database for carbohydrate-active enzyme annotation) and VOGDB (96).

For identifying the closest reference genomes and assessing the completeness of viral genomes, CheckV is a tool widely used. It also detects host contamination, resulting from integrated proviruses. This pipeline compares viral sequences with the IMG/VR (integrated microbial genomes/virus and retrovirus) database (the largest database with reference isolated DNA of viruses

and computationally identified viral contigs (79) and is designed to work with single-contig viral genomes. The tool sets a uniform length threshold and analyses all viral contigs longer than it. Complete or circular genomes can be identified based on the presence of direct terminal repeats and by mapping paired-end sequencing reads. The pipeline is computationally efficient but is not recommended for bulk metagenomic datasets (76).

Taxonomic classification

Another element of phagobiome analysis is determining the family or order to which a given phage belongs. Further classification is limited, and the achievable taxonomic resolution depends on genome completeness. Classification should follow the accepted nomenclature. A virus can be assigned to an existing genus with an established classification, or to a new genus, which requires the formation of a new taxonomic hierarchy that must be approved by the ICTV (15). The ICTV virus classification is constantly changing, as new species are discovered, previously unclassified viruses are categorised, and some families are redefined and removed (136).

Virus classification tools are based on the analysis of viral marker proteins and gene-sharing similarity networks, using HMMs and machine-learning methods, as an advanced alternative to simple BLAST-based comparison for viruses from particular taxa (6, 136). Alignment-based tools include VPF-Class (viral protein family classifier) and CAT (contig annotation tool), network-based tools include vConTACT (viral ConTig automatic cluster taxonomy) and learning-based tools comprise PhaGCN (phage graph convolutional network (GCN)) and PhaGenus (38).

Approaches employing BLAST have several limitations: poorly characterised viruses may be missing from the reference database, and phages themselves are also highly variable and diverse. ProViDE (program for viral diversity estimation) supports BLAST by assigning viruses to different taxonomic levels based on matches to the protein database using virus-specific parameters, but it cannot overcome the problem of a lack of representation in databases (35, 89).

GenomeNet provides the ViPTree web server for creating phage proteomic trees by using genome similarity computed by tBLASTx (translated nucleotide cross-comparison version), enabling classification with reference viruses and their prokaryotic or eukaryotic hosts. Some advantages of this approach include the ability to analyse diverse genomes together, reduced sensitivity to genome rearrangements and the possibility of performing classification based on entire genomes rather than fragments. Unfortunately, this means that it cannot be used for most metagenomic contigs (77).

Artificial intelligence models also achieve satisfactory results. PhaGenus, developed by Guan *et al.* (38), constructs sets of protein clusters from reference genomes, where each group becomes a "word". When an unknown sequence is entered, PhaGenus predicts

ORFs, translates them into proteins and assigns each protein to one cluster. The transformer model learns which proteins appear in the phage genome and in what order. PhaGenus classifies at the genus level. In benchmark tests on RefSeq phage contigs, it achieved a higher prediction rate and comparable or higher classification accuracy than CAT and Kraken2 across a range of sequence lengths, including short contigs (2 kb–20 kb) (38).

In a study by Zhu *et al.* (136), four tools for taxonomic classification were compared based on data from *Caudoviricetes* phages. One of them was vConTACT v. 2.0, utilising the whole genome for phage classification by constructing a network and clustering sequences into a taxon based on genome distance (44). It proved to be the most effective tool for classifying new viruses, although its performance decreased on shorter sequences. Another of the evaluated tools, PhaGCN, classified phages at the family level, while vConTACT v. 2.0, CAT and MMseqs2 (many-against-many sequence searching) categorised viruses to the genus level. On data from the RefSeq database, MMseqs2 assigned 46% of species correctly and outperformed CAT which only assigned 28% correctly (72). Among these tools, PhaGCN was the most accurate, especially for contigs longer than 3,000 base pairs; however, shorter sequences were not supported by this tool (136).

With a design suited to long sequences, CAT uses an ORF-based algorithm, allowing classification of contigs and metagenome assembled genomes (MAGs) (112). The VPF-Class tool combines taxonomic classification with host prediction by assigning viral proteins to viral protein families from large metagenomic datasets (83). Rather than a taxonomic classifier, MMseqs2 is a sequence similarity search engine, which can be incorporated into classification pipelines. It clusters protein and nucleotide sequences using a profile database rather than BLAST, providing faster performance (73).

Newly discovered virus sequences can also be preliminarily assessed for their closest taxonomic relationship by comparison with reference genomes using TaxaBLAST, an application provided by the ICTV. TaxaBLAST is intended for preliminary comparisons and does not constitute a formal taxonomic classification (15).

Functional analysis

Understanding the functions of phage proteins can have potential applications in phage therapy, but accurate functional annotation remains a challenge, as many proteins are hypothetical or classified as having no known function. Viral genes obtained from sequencing are analysed for the presence of proteins with structural, enzymatic and host-modulating functions, including those related to virulence, antibiotic resistance, or bactericidal activity (42, 54). Databases used for functional analyses include Gene Ontology (GO), which provides information on the molecular function of gene

products, the biological processes they participate in and their cellular localisation; and KEGG Orthology (KO), which enables the assignment of genes to metabolic pathways and reaction networks, describing genome functions related to metabolism, transport and signalling (40). It should be noted that these databases are mainly built on cellular organisms, and a large proportion of phage proteins lack annotations in GO or KO, limiting coverage for viral genomes.

Similarity-based methods, artificial intelligence models and orthology-based approaches are used in these analyses (43, 54), but their efficiency remains uncertain for most predicted functions. The tools of research include VIBRANT (54), eggNOG- (evolutionary genealogy of genes: non-supervised orthologous groups) mapper (43), InterProScan (16), DIAMOND (double-index alignMent of next-generation sequencing data) (22) and dbCAN2 (database for carbohydrate-active enzyme aNnotation, version 2) (133).

Methods based on sequence similarity such as DiamondScore and DiamondBlast rely on the assumption that similar sequences have similar functions. This assumption often does not work for phage proteins, as a significant proportion of them have no homologues with known functions. Such genes are commonly referred to as ORFans and are defined as open reading frames that have no detectable homologues in sequence databases. Analyses have shown that approximately 38.4% of phage ORFs have no homologues in other phages, and approximately 30.1% have no detectable homologues in either viral or prokaryotic genomes (129).

DIAMOND is a sequence similarity search tool used to compare protein sequences against a reference database such as KEGG (22). The dbCAN2 server submits metagenomes or proteomes of assembled viruses for annotation against the CAZyme (carbohydrate-active enzymes) database. It relies on HMMs constructed for CAZyme families, as well as DIAMOND and Hotpep, which search for conserved peptide motifs from CAZyme families in the database (133). The RGI (resistance gene identifier) software predicts antibiotic resistance genes from nucleotide or protein sequences from metagenomics data, based on homology and single-nucleotide polymorphism models from CARD (3).

EggNOG-mapper is based on orthology rather than homology, on the premise that genes originating from a single ancestor through vertical divergence have similar biological functions. Phages often evolve through mosaic and horizontal gene transfer, making functional predictions based on orthology less reliable. The tool compares the sequence to proteins in its database using HMMs for rare or new species, or DIAMOND for well-known species, and balances speed and memory consumption. It then verifies whether the best-matching protein is indeed an orthologue of the given sequence. Based on orthology assignment, it

determines functional annotations described in GO, KEGG and COG (the clusters of orthologous genes protein database). In comparative analyses, eggNOG-mapper proved more accurate than BLAST and InterProScan on well-characterised model organism proteomes (43). Its performance on phage proteins can be inferior because of the underrepresentation of viral sequences in reference databases.

InterProScan searches protein sequences against the InterPro database, predicting their function, finding domains and sites and classifying proteins into the families to which they belong (16). Many phage proteins lack recognisable domains in InterPro, which limits the useability of this tool.

In the study by Guan *et al.* (37), the authors presented a new program, GOPhage, dedicated to the analysis of typical phage proteins, including newly discovered ones. It uses the ESM2 protein language model, which creates a vector describing proteins based on their amino acid sequences, capturing features related to structure, biochemistry and interaction patterns. The phage genome is then transcribed into "sentences", composed of proteins arranged in a specific order. A transformer model is trained to analyse the protein order in the genome and determine their functions based on relationships between neighbouring proteins. Based on probability, an assignment is made to one of three protein types each with a Gene Ontology term: molecular functions, biological processes or cellular components. Tests performed on phage proteins and genomes from the class *Caudoviricetes* in the National Center for Biotechnology Information RefSeq database showed the efficiency of GoPhage in the functional classification of phage proteins (37).

Host prediction

Host prediction methods use whole-sequence similarity (alignment-based), sequence composition (alignment-free), machine learning and CRISPR spacer and tRNA information (137, 138). Current approaches predict only with high uncertainty, especially below the species level. Available tools include homology-based methods such as PHIST (phage–host interaction search tool) (138) and Phirbo (phage–host interaction ranking BLAST output) (137); alignment-free solutions such as WIsH (who is the host?) (32); CRISPR-based tools like SpacePHARER (CRISPR Spacer Phage-Host pAiRs findER) (134) and machine learning approaches such as VirHostMatcher-Net (115) and PHP (Prokaryotic virus Host Predictor) (66).

Alignment-based methods use the homology between viral and host genomic or protein sequences, which arises from their co-evolution, and their efficiency depends on the availability of well-represented genomes in databases. Homology signals can come from horizontal gene transfer, integrated prophages or shared metabolic genes, and are detected using tools such as BLAST (66). Their performance is affected by genome variability, resulting in high accuracy but low recall.

Methods based on CRISPR- exploit the fact that some hosts retain fragments from infections in their genomes. These methods involve matching viral sequences to spacers encoded in host genomes. These approaches may not be perfectly accurate because not all hosts will contain CRISPR spacers corresponding to viruses present in the analysed sample (138).

Alignment-free prediction methods use statistical and probability models to compare host-dependent sequence signatures, including *k*-mer composition, codon usage and oligonucleotide frequencies (17). Several tools have been developed to improve the performance of homology-based methods, such as VirHostMatcher-Net, which also integrates multiple signals (137) and Phirbo (138).

VirHostMatcher-Net integrates CRISPR information, host–virus, and virus–virus sequence similarity, including BLAST-derived scores and network-based associations, and can make predictions based on short viral contigs or whole genomes. It has been successfully used to predict crAss-like phages (115). The other machine-learning implementation, PHP, uses a Gaussian model that captures the frequency of *k*-mer shared between viruses and their hosts, predicting host likelihood based on probability. It can analyse complete or fragmentary sequences and outperformed VirHostMatcher and WIsH in benchmarking tests by Lu *et al.* (66).

SpacePHARER predicts hosts by comparing proteins encoded by CRISPR spacers with phage proteins from metagenomic data, using the homology-based protein software search tool MMseqs2, which achieved higher sensitivity and faster performance than BLASTN (134). This approach is limited in applicability only to hosts of which the genomes contain CRISPRs in reference databases.

WIsH is an alignment-free tool that predicts a host using a pre-trained Markov model. It showed good speed and accuracy on long contigs, but lower accuracy for short contigs compared to alignment-based methods (32).

Phirbo extends BLAST by considering not only the three top host candidates suggested by BLAST, but the entire ranked list, which it further analyses for sequence matching. In a benchmark test against BLAST and WIsH, Phirbo achieved the highest accuracy among the tree tools across different taxonomic levels, including the genus, family and species. However, its overall performance was lower at the species level, than genus and family (137).

In a publication by Zielezinski *et al.* (138), the authors introduced PHIST, a host-prediction method based on *k*-mers, and compared it with Phirbo, BLASTN, WIsH, PILER-CR (a CRISPR repeat-finding tool based on pile construction) and SpacePHARER. PHIST was proven to be faster and more accurate than the others.

In recent years, artificial intelligence models dedicated to viral discovery and host prediction have

emerged, such as HostG. HostG uses a GCN in which viruses and hosts are represented as nodes, and edges encode similarity relationships between them. Edges are created based on the similarity of proteins between viruses or the similarity of DNA sequences between viruses and hosts. Each node in the graph has a feature vector generated by a CNN based on the raw sequence, enabling the extraction of sequence patterns. Graph convolutional networks predict hosts by propagating information from nodes with known hosts to unannotated viral nodes. An additional feature of the model is the ability to set a threshold for predictions to filter out low-confidence assignments. HostG and the other classifiers VirHostMatcher-net, WIsH, PHP, HoPhage (host of phage), RaFAH (random forest assignment of hosts), vHULK (viral host unveiling kit) and VPF-Class, were evaluated relatively on simulated and real data. Shang and Sun, the evaluators (98), indicated that HostG outperformed all other tested tools.

The validity of combining machine learning models with other methods in host prediction is further supported by the development of hybrid programs, such as CoMPHI (composite model for phage–host interaction). This hybrid creates multiple feature encodings from viral and host nucleotide and protein sequences and incorporates alignment-based scores into an ML model to improve predictions (17). Despite that, it does not solve the fundamental problem of host prediction failure when the host is not represented in the genome reference database.

Statistical analysis

Statistical analyses can be used to distinguish between different phage types, for example, by identifying core phages present across most populations versus those occurring only in individual organisms, as well as to study age-related changes in the phageome (42).

The data pertaining to phageomes are often sparse and high-dimensional, containing many zeros arising from rare taxa, information gaps due to undersampling, or compositional effects. This structure creates a challenge for differential abundance methods and requires the use of logarithmic or other transformations to account for excess zeros when identifying taxa that differ between groups (63).

In phage and virus analyses, viral contigs are typically grouped into vOTUs or viral clusters, similarly to how MAGs are grouped, to assess diversity and abundance in samples. Phage data are compositional in nature, and before performing statistical analyses, it is recommended to apply logarithmic transformations such as centred, additive or isometric log-ratio. Total-sum scaling is a method used to normalise and correct differences in sequencing depth in a sample (93, 113).

Principal Component Analysis (PCA) or Principal Coordinate Analysis (PCoA) is used to reduce the size of dimensions. Applying PCA allows the identification of taxa that contribute more significantly to the variation appearing among samples (10).

Viral community differences can also be measured using alpha- and beta-diversity indices. Alpha diversity quantifies variation within individual samples and captures abundance-based diversity of viral contig clusters using Shannon and Simpson indices. Beta diversity measures variation between samples using distance metrics such as Bray-Curtis. Read counts need to be normalised by contig length and sequencing depth, because contig clusters in the virome sample vary in length and completeness (90).

The analysis of similarities statistical test compares the similarity between and within groups based on the distance matrices, and permutational multivariate analysis of variance (PERMANOVA) is a permutation test that is used for comparing groups of microbial communities. A major limitation of PERMANOVA is its sensitivity to differences in within-group dispersion, which are often mistaken for differences in centroids (55). Analysis of compositions of microbiomes is a differential abundance test that uses logarithmic transformation to identify taxa that differ between groups, but its performance can be affected by sparsity and compositional bias and requires additional data transformation (70). Visualisation methods such as dendrograms and heat-maps help illustrate similarity and differences between groups (33).

For metavirome analyses, methods like ENViT (a method for estimating the richness of novel viral mixtures) have been developed to estimate viral abundance and average genome length, using the distribution to fit a virus population model. The ENViT algorithm provides the estimation without dependence on reference databases, making it suitable for poorly described viruses. Compared with the PHACCS (phage communities from contig spectrum) and CatchAll methods, it achieved more accurate results (47).

Visualisation of results

The achieved results can be visualised using plots that show the taxonomic composition and functional diversity of samples (6, 33, 102, 120).

The R programming language, together with a broad collection of packages, is one of the most used environments for data visualisation and statistical analysis. The ggplot2 package allows users to choose different graph types, such as ternary plots with ggtern, network graphs with ggraph and trees and cladograms with ggtree (120). Currently, there are many libraries facilitating data visualisation and processing, diversity analysis, statistical analyses, biomarker discovery, network and correlation creation, functional analyses and modelling in machine learning (120). Another programming language often used for data visualisation, preprocessing and statistical analysis is Python, which offers libraries such as Matplotlib, Seaborn, SciPy, scikit-learn, and pandas (51).

Analysis and tool limitations

Despite the availability of numerous tools and environments, fully end-to-end systems dedicated to phageome research are still limited. Integrated platforms such as PhaBOX (a toolbox for phage analysis and characterisation) (97) or geNomad (23) cover selected aspects of analysis, while having their limitations. Many studies, therefore, still require combining a variety of separate, more suited tools into custom workflows. The increasing amount of new data poses challenges in terms of memory usage and longer processing time. As a result, there is a growing demand for a more comprehensive and compact platform that offers a broader range of steps in phageome analyses, while optimising computational requirements. Table 2 summarises the main limitations of current viral bioinformatics methods and highlights directions for future development.

In recent years, dedicated web servers capable of performing comprehensive analysis have been developed. PhaBOX, recently updated to PhaBOX2 (a pipeline that combines modules with varying levels of accuracy and inherent limitations) is implemented in R and Python and integrates the functionality of several different tools: PhaMer for phage contig identification, PhaGCN for taxonomic classification, CHERRY (computational metHod for accuratE pRediction of virus-prokaRotic interactions using a graph encoder-decoder model for host prediction, and PhaTYP for determination of phage lifestyle type (lytic or lysogenic). It also renders a visualisation of the obtained results. Phage identification and its life cycle prediction are based on a linguistic model in which a transformer learns sequence patterns. For classification and host prediction, a graph based on DNA and proteins is used, along with a GCN for characterised and undescribed features. CHERRY is a newer form of HostG. Both end-to-end analyses and the execution of only selected steps are possible in PhaBOX (97). In a study by Hu *et al.* (42), this tool was used to identify, classify and functionally annotate the pig gut phageome from metagenomic data, including host prediction, lifecycle determination and detection of ARGs.

The literature lacks direct comparisons of the performance of PhaBOX with those of other tools, as few tools are as advanced as PhaBOX. Tests of its individual modules separately show varying results. In a benchmark PhaGenus evaluation, PhaGCN showed limited usefulness, as it is restricted to family-level classification and does not support taxonomic assignment at the genus level (136). HostG, an earlier version of the CHERRY module, demonstrated high effectiveness in phage host prediction (98). In tests conducted by its author, PhaMer performed well in the identification of novel viruses (99). Overall, the performance of PhaBOX modules is generally strong; however, there are some limitations, as illustrated by PhaGCN, which highlights that PhaBOX may not be suitable for every analytical problem.

There is also a second, similarly advanced tool, geNomad, created by Camargo *et al.* (23). This program is used for the identification of viral and plasmid genomes in nucleotide sequences, as well as for taxonomic classification and gene functional annotation. In addition, geNomad can detect proviruses integrated into host genomes based on a conditional random field model. It operates on sets of protein markers and classification models based on neural networks, allowing mobile genetic elements to be distinguished. For classification, geNomad uses a hybrid approach, including alignment-free and gene-based models (23).

The usefulness of functional analysis tools for phages is reduced significantly by the presence of ORFs. As new genomes are sequenced, the number of ORFans tends to increase (129). Metagenome studies indicate that 40–90% of viral sequences cannot be assigned to any function or even matched to known viral sequences. Uncharacterised viral genes are referred to as ‘viral dark matter’. Even in reference databases, a significant proportion of proteins are labelled as hypothetical or unknown. Furthermore, host assignment for new phages is often based on statistical methods such as CRISPR spacer matching, rather than experimental validation. The genetic variability and rapid evolution of viruses, together with incomplete reference databases, highlight the limitations of functional annotation methods (58).

Sources of uncertainty in phagobiome analyses may arise at the very beginning of analyses and increase at subsequent stages. At the sample preparation stage, multiple displacement amplification, used to increase viral DNA prior to sequencing, tends to introduce systematic error by increasing the amplification of small circular single-stranded DNA virus genomes and under-representing sequences with extremely high guanine-cytosine content (81). Similarly, differences between the outcomes of metagenomic sequencing of virus-like particle-enriched faecal samples and of bulk sequencing showed only partial coverage of the detected genomes. Virus-like particle methods may artificially increase the representation of certain viruses because of structural differences in the cell walls of bacterial hosts, leading to a shift in results toward viruses with low prevalence, low abundance or those infecting Gram-positive hosts (62).

During the assembly of genomes from metagenomic data, uneven sequencing coverage and high genomic diversity of viral populations lead to contig fragmentation. The degree of fragmentation is strongly dependent on the genome’s length and the sequencing coverage distribution.

Assembly can result in numerous short contigs with varying coverage, reducing the accuracy of genome assignment (34).

Computational host prediction methods are constrained by fragmented viral assemblies and limited reference host datasets, which makes predicting the actual hosts of newly discovered viruses unreliable (108).

Table 2. Summary of limitations and development direction for virome analysis

Analysis stage	Limitation	Development directions
Viral DNA identification	Limited representation of poorly characterised viruses in reference databases can lead to low sensitivity to novel and highly divergent viruses in similarity-based approaches	Expansion of virus reference databases; deep learning models trained on wider datasets
Genome assembly	Uneven coverage and population variability in viral and metagenomic datasets (<i>de novo</i> assembly); assembly of genomes with terminal repeats often results in fragmented assemblies; strain-level reconstruction remains challenging	Hybrid assemblers (short and long reads); using approximate rather than exact <i>k</i> -mer alignment
Genome annotation	Prone to false positives due to bacterial contamination; machine-learning-based annotations depend on the quality and representativeness of training datasets	Integration of records from multiple databases for cross-validation; improvement of contamination detection
Taxonomic classification	Lack of universal viral markers; resolution depends on genome completeness; genomic diversity and rearrangements limit classification accuracy; classification based on entire genomes is not applicable in bulk metagenomes	Standardised virus taxonomy; graph, trees and network-based classification
Functional analysis	Large proportion of phage proteins remain hypothetical or uncharacterised; large numbers of ORFans and ‘viral dark matter’; rapid evolution and incomplete reference databases give uncertain results	Expansion of virus-specific functional databases; improvement of protein family clustering; integration of machine learning-based structural prediction
Host prediction	High uncertainty, especially below the species level; affected by genome variability and database bias; CRISPR-based methods limited by the absence of CRISPRs in host and reference databases; high false-positive rate	Development of hybrid approaches combining multiple signals (CRISPR, sequence similarity, <i>k</i> -mer composition and network-based inference)
Statistical analysis	Data sparsity and multidimensionality with excess zeros due to rare taxa; undersampling; compositional effects	Development of statistical methods specific for virome
Multianalysis	Individual modules may perform worse than specialised tools; less flexibility and transparency; heavy memory usage	Design of modular pipelines enabling tool substitution, benchmarking and further optimisation

CRISPR – clustered, regularly interspaced short palindromic repeats

Conclusion

Phage studies involve a multistep analysis that requires different approaches, making the characterisation of phagobiomes challenging. Creating and properly handling metavirome datasets is crucial for the identification of putative novel viral lineages or vOTUs and known phages in samples. However, in bulk metagenomes, a major challenge is the low proportion of viral sequences, which requires reliable methods for efficiently analysing the phage component of the microbiome. The choice of the most appropriate tool follows the type of data (e.g. the environmental origin of the sample and the sequence divergence), their quality (e.g. the presence of prophages and host contamination), and their size (e.g. contig length, sequencing depth and viral genome completeness). The choice must also consider the purpose of the analysis (e.g. phage identification, functional characterisation or host prediction) and the available computational resources (e.g. RAM size and number of processor threads). The conclusion is therefore that there is no single optimal process that would be suitable for all virus data, although several of the presented tools show better performance than others and demonstrate potential for phage studies in their present form and for development in the future. Despite the assorted nature of the tools required through the steps of multi-stage analysis for phagobiome characterisation, in most cases, tools using ML methods and hybrid solutions were proved to be the most effective. These approaches generally provide higher accuracy and sensitivity, while also reducing computational times. Their performance, however,

depends on the quality and representativeness of the training data, and the inherent uncertainty of metagenomic viral data means that the results should be interpreted with caution.

CRedit Authorship Contribution Statement: **Anna Miastowska:** research concept and design, collection and assembly of data, data analysis and interpretation, writing the article, critical revision of the article, final approval of the article. **Adrian Augustyniak:** research concept and design, data analysis and interpretation, writing the article, critical revision of the article, final approval of the article. **Bartłomiej Grygorcewicz:** data analysis and interpretation, critical revision of the article, final approval of the article. **Paweł Nawrotek:** research concept and design, data analysis and interpretation, writing the article, critical revision of the article, final approval of the article.

Conflict of Interests Statement: The authors declare that there is no conflict of interests regarding the publication of this article.

Financial Disclosure Statement: This research received no external funding.

Animal Rights Statement: None required.

References

1. Ahmed W., Payyappat S., Cassidy M., Besley C., Power K.: Novel crAssphage marker genes ascertain sewage pollution in

- a recreational lake receiving urban stormwater runoff. *Water Res* 2018, 145, 769–778, doi: 10.1016/j.watres.2018.08.049.
2. Albright M.B.N., Louca S., Winkler D.E., Feeser K.L., Haig S.J., Whiteson K.L., Emerson J.B., Dunbar J.: Solutions in microbiome engineering: prioritizing barriers to organism establishment. *ISME J* 2022, 16, 331–338, doi: 10.1038/s41396-021-01088-5.
 3. Alcock B.P., Huynh W., Chalil R., Smith K.W., Raphenya A.R., Wlodarski M.A., Edalatmand A., Petkau A., Syed S.A., Tsang K.K., Baker S.J.C., Dave M., McCarthy M.C., Mukiri K.M., Nasir J.A., Golbon B., Imtiaz H., Jiang X., Kaur K., Kwong M., Liang Z.C., Niu K.C., Shan P., Yang J.Y.J., Gray K.L., Hoad G.R., Jia B., Bhando T., Carfrae L.A., Farha M.A., French S., Gordzevich R., Rachwalski K., Tu M.M., Bordeleau E., Dooley D., Griffiths E., Zubyk H.L., Brown E.D., Maguire F., Beiko R.G., Hsiao W.W.L., Brinkman F.S.L., Van Domselaar G., McArthur A.G.: CARD 2023: expanded curation, support for machine learning, and resistome prediction at the Comprehensive Antibiotic Resistance Database. *Nucleic Acids Res* 2023, 51, D690–D699, doi: 10.1093/nar/gkac920.
 4. Altschul S., Madden T.L., Schäffer A.A.: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997, 25, 3389–3402, doi: 10.1093/nar/25.17.3389.
 5. Amgarten D., Braga L.P.P., da Silva A.M., Setubal J.C.: MARVEL, a tool for prediction of bacteriophage sequences in metagenomic bins. *Front Genet* 2018, 9, 304, doi: 10.3389/fgene.2018.00304.
 6. Andrade-Martínez J.S., Camelo Valera L.C., Chica Cárdenas L.A., Forero-Junco L., López-Leal G., Moreno-Gallego J.L., Rangel-Pineros G., Reyes A.: Computational tools for the analysis of uncultivated phage genomes. *Microbiol Mol Biol Rev* 2022, 86, e00004-21, doi: 10.1128/mmlbr.00004-21.
 7. Andrews S.: FastQC: A quality control tool for high throughput sequence data. Babraham Bioinformatics, Babraham Institute, Babraham, UK, 2010, <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>.
 8. Antipov D., Raiko M., Lapidus A., Pevzner P.A.: MetaviralSPAdes: assembly of viruses from metagenomic data. *Bioinformatics* 2020, 36, 4126–4129, doi: 10.1093/bioinformatics/btaa490.
 9. Antipov D., Rayko M., Kolmogorov M., Pevzner P.A.: viralFlye: assembling viruses and identifying their hosts from long-read metagenomics data. *Genome Biol* 2022, 23, 57, doi: 10.1186/s13059-021-02566-x.
 10. Armstrong G., Rahman G., Martino C., McDonald D., Gonzalez A., Mishne G., Knight R.: Applications and comparison of dimensionality reduction methods for microbiome data. *Front Bioinform* 2022, 2, 821861, doi: 10.3389/fbinf.2022.821861.
 11. Auslander N., Gussow A.B., Benler S., Wolf Y.I., Koonin E.V.: Seeker: alignment-free identification of bacteriophage genomes by deep learning. *Nucleic Acids Res* 2020, 48, e121, doi: 10.1093/nar/gkaa856.
 12. Avellaneda-Franco L., Dahlman S., Barr J.J.: The gut virome and the relevance of temperate phages in human health. *Front Cell Infect Microbiol* 2023, 13, 1241058, doi: 10.3389/fcimb.2023.1241058.
 13. Azeez S.S., Hamad R.S., Hamad B.K., Shekha M.S., Bergsten P.: Advances in CRISPR-Cas technology and its applications: revolutionising precision medicine. *Front Genome Ed* 2024, 6, 1509924, doi: 10.3389/fgeed.2024.1509924.
 14. Balcha F.B., Neja S.A.: CRISPR-Cas9 mediated phage therapy as an alternative to antibiotics. *Anim Dis* 2023, 3, 4, doi: 10.1186/s44149-023-00065-z.
 15. Black E.J., Powell C.S., Dempsey D.M., Hendrickson R.C., Mims L.R., Lefkowitz E.J.: Virus taxonomy: the database of the International Committee on Taxonomy of Viruses. *Nucleic Acids Res* 2025, 54, D776–D789, doi: 10.1093/nar/gkaf1159.
 16. Blum M., Chang H.Y., Chuguransky S., Grego T., Kandasamy S., Mitchell A., Nuka G., Paysan-Lafosse T., Qureshi M., Raj S., Richardson L., Salazar G.A., Williams L., Bork P., Bridge A., Gough J., Haft D.H., Letunic I., Marchler-Bauer A., Mi H., Natale D.A., Necci M., Orengo C.A., Pandurangan A.P., Rivoire C., Sigrist C.J.A., Sillitoe I., Thanki N., Thomas P.D., Tosatto S.C.E., Wu C.H., Bateman A., Finn R.D.: The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res* 2021, 49, D344–D354, doi: 10.1093/nar/gkaa977.
 17. Bodaka S., Kolliputi N.: CoMPHI: a novel composite machine learning approach utilizing multiple feature representation to predict hosts of bacteriophages. *Front Bioinform* 2025, 5, 1622931, doi: 10.3389/fbinf.2025.1622931.
 18. Bouchemla F., Akchurin S.V., Akchurina I.V., Dyulger G.P., Latynina E.S., Grecheneva A.V.: Artificial intelligence feasibility in veterinary medicine: A systematic review. *Vet World* 2023, 16, 2143–2149, doi: 10.14202/vetworld.2023.2143-2149.
 19. Bouras G., Nepal R., Houtak G., Psaltis A.J., Wormald P.J., Vreugde S.: Pharokka: a fast scalable bacteriophage annotation tool. *Bioinformatics* 2023, 39, btac776, doi: 10.1093/bioinformatics/btac776.
 20. Brunse A., Deng L., Pan X., Hui Y., Castro-Mejía J.L., Kot W., Nguyen D.N., Secher J.B., Nielsen D.S., Thymann T.: Fecal filtrate transplantation protects against necrotizing enterocolitis. *ISME J* 2022, 16, 686–694, doi: 10.1038/s41396-021-01107-5.
 21. Bucher M.J., Czyż D.M.: Phage against the machine: the SIE-ence of superinfection exclusion. *Viruses* 2024, 16, 1348, doi: 10.3390/v16091348.
 22. Buchfink B., Reuter K., Drost H.G.: Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat Methods* 2021, 18, 366–368, doi: 10.1038/s41592-021-01101-x.
 23. Camargo A.P., Roux S., Schulz F., Babinski M., Xu Y., Hu B., Chain P.S.G., Nayfach S., Kyrpides N.C.: Identification of mobile genetic elements with geNomad. *Nat Biotechnol* 2024, 42, 1303–1312, doi: 10.1038/s41587-023-01953-y.
 24. Chen S.: Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp. *iMeta* 2023, 2, e107, doi: 10.1002/imt2.107.
 25. Chen S.: fastp 1.0: An ultra-fast all-round tool for FASTQ data quality control and preprocessing. *iMeta* 2025, 4, e70078, doi: 10.1002/imt2.70078.
 26. De Coster W., D’Hert S., Schultz D.T., Cruts M., Van Broeckhoven C.: NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics* 2018, 34, 2666–2669, doi: 10.1093/bioinformatics/bty149.
 27. Dong W., Ricker N., Holman D.B., Johnson T.A.: Meta-analysis reveals the predictable dynamic development of the gut microbiota in commercial pigs. *Microbiol Spectr* 2023, 11, e01722-23, doi: 10.1128/spectrum.01722-23.
 28. Eck N.J. van, Waltman L.: Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics* 2010, 84, 523–538, doi: 10.1007/s11192-009-0146-3.
 29. Ejigu G.F., Jung J.: Review on the computational genome annotation of sequences obtained by next-generation sequencing. *Biology* 2020, 9, 295, doi: 10.3390/biology9090295.
 30. Fang Z., Tan J., Wu S., Li M., Xu C., Xie Z., Zhu H.: PPR-Meta: a tool for identifying phages and plasmids from metagenomic fragments using deep learning. *Gigascience* 2019, 8, giz066, doi: 10.1093/gigascience/giz066.
 31. Fritz A., Bremges A., Deng Z.-L., Lesker T.R., Götting J., Ganzenmueller T., Sczyrba A., Dilthey A., Klawonn F., McHardy A.C.: Haploflow: strain-resolved de novo assembly of viral genomes. *Genome Biol* 2021, 22, 212, doi: 10.1186/s13059-021-02426-8.
 32. Galiez C., Siebert M., Enault F., Vincent J., Söding J.: WIsH: who is the host? Predicting prokaryotic hosts from metagenomic phage contigs. *Bioinformatics* 2017, 33, 3113–3114, doi: 10.1093/bioinformatics/btx383.
 33. Galloway-Peña J., Hanson B.: Tools for analysis of the microbiome. *Dig Dis Sci* 2020, 65, 674–685, doi: 10.1007/s10620-020-06091-y.
 34. García-López R., Vázquez-Castellanos J.F., Moya A.: Fragmentation and coverage variation in viral metagenome assemblies, and their effect in diversity calculations. *Front Bioeng Biotechnol* 2015, 3, 141, doi: 10.3389/fbioe.2015.00141.
 35. Ghosh T.S., Haque M.M., Komanduri D., Mande S.S.: ProViDE: A software tool for accurate estimation of viral diversity in

- metagenomic samples. *Bioinformatics* 2011, 6, 91–94. doi: 10.6026/97320630006091.
36. Gish W., States D.J.: Identification of protein coding regions by database similarity search. *Nat Genet* 1993, 3, 266–272, doi: 10.1038/ng0393-266.
 37. Guan J., Ji Y., Peng C., Zou W., Tang X., Shang J., Sun Y.: GOPhage: protein function annotation for bacteriophages by integrating the genomic context. *Brief Bioinform* 2024, 26, bbaf014, doi: 10.1093/bib/bbaf014.
 38. Guan J., Peng C., Shang J., Tang X., Sun Y.: PhaGenus: genus-level classification of bacteriophages using a Transformer model. *Brief Bioinform* 2023, 24, bbad 408, doi: 10.1093/bib/bbad408.
 39. Guo J., Bolduc B., Zayed A.A., Varsani A., Dominguez-Huerta G., Delmont T.O., Pratama A.A., Gazitúa M.C., Vik D., Sullivan M.B., Roux S.: VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. *Microbiome* 2021, 9, 37, doi: 10.1186/s40168-020-00990-y.
 40. He L., Zou Q., Wang Y.: metaTP: a meta-transcriptome data analysis pipeline with integrated automated workflows. *BMC Bioinformatics* 2025, 26, 111, doi: 10.1186/s12859-025-06137-w.
 41. Ho S.F.S., Wheeler N.E., Millard A.D., van Schaik W.: Gauge your phage: benchmarking of bacteriophage identification tools in metagenomic sequencing data. *Microbiome* 2023, 11, 84, doi: 10.1186/s40168-023-01533-x.
 42. Hu J., Chen J., Nie Y., Zhou C., Hou Q., Yan X.: Characterizing the gut phageome and phage-borne antimicrobial resistance genes in pigs. *Microbiome* 2024, 12, 102, doi: 10.1186/s40168-024-01818-9.
 43. Huerta-Cepas J., Forslund K., Coelho L.P., Szklarczyk D., Jensen L.J., von Mering C., Bork P.: Fast genome-wide functional annotation through orthology assignment by eggNOG-Mapper. *Mol Biol Evol* 2017, 34, 2115–2122, doi: 10.1093/molbev/msx148.
 44. Hunt M., Gall A., Ong S.H., Brenner J., Ferns B., Goulder P., Nastouli E., Keane J.A., Kellam P., Otto T.D.: IVA: accurate de novo assembly of RNA virus genomes. *Bioinformatics* 2015, 31, 2374–2376, doi: 10.1093/bioinformatics/btv120.
 45. Jang H.B., Bolduc B., Zablocki O., Kuhn J.H., Roux S., Adriaenssens E.M., Brister J.R., Kropinski A.M., Krupovic M., Lavigne R., Turner D., Sullivan M.B.: Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. *Nat Biotechnol* 2019, 37, 632–639, doi: 10.1038/s41587-019-0100-8.
 46. Jankowski W., Mizielińska M., Nawrotek P.: Microbiome and phageome: key factors in host organism function and disease prevention in the context of microbiome transplants. *Appl Sci* 2025, 15, 5330, doi: 10.3390/app15105330.
 47. Jayasundara D., Herath D., Senanayake D., Saeed I., Yang C.Y., Sun Y., Chang B.C., Tang S.L., Halgamuge S.K.: ENViT: inference of ecological characteristics of viruses from metagenomic data. *BMC Bioinformatics* 2019, 19, 377, doi: 10.1186/s12859-018-2398-5.
 48. Jin Song S., Woodhams D.C., Martino C., Allaband C., Mu A., Javorschi-Miller-Montgomery S., Suchodolski J.S., Knight R.: Engineering the microbiome for animal health and conservation. *Exp Biol Med* 2019, 244, 494–504, doi: 10.1177/1535370219830075.
 49. Jochheim A., Jochheim F.A., Kolodyazhnaya A., Morice É., Steinegger M., Söding J.: Strain-resolved de-novo metagenomic assembly of viral genomes and microbial 16S rRNAs. *Microbiome* 2024, 12, 187, doi: 10.1186/s40168-024-01904-y.
 50. Johansen J., Plichta D.R., Nissen J.N., Jespersen M.L., Shah S.A., Deng L., Stokholm J., Bisgaard H., Nielsen D.S., Sørensen S.J., Rasmussen S.: Genome binning of viral entities from bulk metagenomics data. *Nat Commun* 2022, 13, 965, doi: 10.1038/s41467-022-28581-5.
 51. Joshi A., Tiwari H.: An overview of Python libraries for data science. *J Eng Technol Appl Phys* 2023, 5, 85–90, doi: 10.33093/jetap.2023.5.2.10.
 52. Jurtz V.I., Villarreal J., Lund O., Voldby Larsen M., Nielsen M.: MetaPhinder—identifying bacteriophage sequences in metagenomic data sets. *PLoS One* 2016, 11, e0163111, doi: 10.1371/journal.pone.0163111.
 53. Kam S.L., Bhartiya S.K., Pratap A., Saroj S.K., Kumar R., Sahu M., Gangwar M., Nath G.: A randomized, placebo-controlled, double-blind clinical trial of bacteriophage cocktails in chronic wound infections. *Int J Low Extrem Wounds* 2024, 25, 427–437, doi: 10.1177/15347346231226342.
 54. Kieft K., Zhou Z., Anantharaman K.: VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. *Microbiome* 2020, 8, 90, doi: 10.1186/s40168-020-00867-0.
 55. Kleine Bardenhorst S., Berger T., Klawonn F., Vital M., Karch A., Rübsamen N.: Data analysis strategies for microbiome studies in human populations—a systematic review of current practice. *mSystems* 2021, 6, 10.1128/msystems.01154-20, doi: 10.1128/mSystems.01154-20.
 56. Kolmogorov M., Bickhart D.M., Behsaz B., Gurevich A., Rayko M., Shin S.B., Kuhn K., Yuan J., Polevikov E., Smith T.P.L., Pevzner P.A.: metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nat Methods* 2020, 17, 1103–1110, doi: 10.1038/s41592-020-00971-x.
 57. Kolmogorov M., Yuan J., Lin Y., Pevzner P.A.: Assembly of long, error-prone reads using repeat graphs. *Nat Biotechnol* 2019, 37, 540–546, doi: 10.1038/s41587-019-0072-8.
 58. Kosmopoulos J.C., Anantharaman K.: Viral dark matter: illuminating protein function, ecology, and biotechnological promises. *Biochemistry* 2025, 64, 4609–4627, doi: 10.1021/acs.biochem.5c00349.
 59. Lathakumari R.H., Vajravelu L.K., Satheesan A., Ravi S., Thulukanam J.: Antibiotics and the gut microbiome: Understanding the impact on human health. *Med Microecol* 2024, 20, 100106, doi: 10.1016/j.medmic.2024.100106.
 60. 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* 2015, 31, 1674–1676, doi: 10.1093/bioinformatics/btv033.
 61. Li Y., Gordon E., Shean R.C., Idle A., Deng X., Greninger A.L., Delwart E.: CrAssphage and its bacterial host in cat feces. *Sci Rep* 2021, 11, 815, doi: 10.1038/s41598-020-80076-9.
 62. Li Y., Sun C., Zhu J., Geng M., Li M., Zhao X.M., Chen W.H.: Biases and complementarity in gut viromes obtained from bulk and virus-like particle-enriched metagenomic sequencing. *Microbiol Spectr* 2025, 13, e00013-25, doi: 10.1128/spectrum.00013-25.
 63. Lin H., Peddada S.D.: Analysis of microbial compositions: a review of normalization and differential abundance analysis. *NPJ Biofilms Microbiomes* 2020, 6, 60, doi: 10.1038/s41522-020-00160-w.
 64. Lin J., Du F., Long M., Li P.: Limitations of phage therapy and corresponding optimization strategies: a review. *Molecules* 2022, 27, 1857, doi: 10.3390/molecules27061857.
 65. Lopez J.A., McKeithen-Mead S., Shi H., Nguyen T.H., Huang K.C., Good B.H.: Abundance measurements reveal the balance between lysis and lysogeny in the human gut microbiome. *Curr Biol* 2025, 35, 2282–2294, doi: 10.1016/j.cub.2025.03.073.
 66. Lu C., Zhang Z., Cai Z., Zhu Z., Qiu Y., Wu A., Jiang T., Zheng H., Peng Y.: Prokaryotic virus host predictor: a Gaussian model for host prediction of prokaryotic viruses in metagenomics. *BMC Biol* 2021, 19, 5, doi: 10.1186/s12915-020-00938-6.
 67. MacDonald M.L., Polson S.W., Lee K.H.: *k*-mer-based metagenomics tools provide a fast and sensitive approach for the detection of viral contaminants in biopharmaceutical and vaccine manufacturing applications using next-generation sequencing. *mSphere* 2021, 6, 10.1128/mSphere.01336-20, doi: 10.1128/mSphere.01336-20.
 68. Majewska J., Kaźmierczak Z., Lahutta K., Lecion D., Szymczak A., Miernikiewicz P., Drapała J., Harhala M., Marek-Bukowiec K., Jędruchiewicz N., Owczarek B., Górski A., Dąbrowska K.: Induction of phage-specific antibodies by two

- therapeutic staphylococcal bacteriophages administered *per os*. *Front Immunol* 2019, 10, 2607, doi: 10.3389/fimmu.2019.02607.
69. Malla B., Makise K., Nakaya K., Mochizuki T., Yamada T., Haramoto E.: Evaluation of human- and animal-specific viral markers and application of CrAssphage, pepper mild mottle virus, and tobacco mosaic virus as potential fecal pollution markers to river water in Japan. *Food Environ Virol* 2019, 11, 446–452, doi: 10.1007/s12560-019-09398-w.
 70. Mandal S., Van Treuren W., White R.A., Eggesbø M., Knight R., Peddada S.D.: Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microb Ecol Health Dis* 2015, 26, 27663, doi: 10.3402/mehd.v26.27663.
 71. Mi J., Jing X., Ma C., Yang Y., Li Y., Zhang Y., Long R., Zheng H.: Massive expansion of the pig gut virome based on global metagenomic mining. *NPJ Biofilms Microbiomes* 2024, 10, 76, doi: 10.1038/s41522-024-00554-0.
 72. Mirdita M., Steinegger M., Breitwieser F., Söding J., Levy Karin E.: Fast and sensitive taxonomic assignment to metagenomic contigs. *Bioinformatics* 2021, 37, 3029–3031, doi: 10.1093/bioinformatics/btab184.
 73. Mirdita M., Steinegger M., Söding J.: MMseqs2 desktop and local web server app for fast, interactive sequence searches. *Bioinformatics* 2019, 35, 2856–2858, doi: 10.1093/bioinformatics/bty1057.
 74. Nami Y., Imeni N., Panahi B.: Application of machine learning in bacteriophage research. *BMC Microbiol* 2021, 21, 193, doi: 10.1186/s12866-021-02255-5.
 75. Nawrotek P.: Faecal virome transplantation: a promising alternative to antibiotics in livestock production (in Polish). *Życie Wet* 2026, 101, 148–153.
 76. Nayfach S., Camargo A.P., Schulz F., Elie-Fadrosh E., Roux S., Kyrpides N.C.: CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nat Biotechnol* 2021, 39, 578–585, doi: 10.1038/s41587-020-00774-7.
 77. Nishimura Y., Yoshida T., Kuronishi M., Uehara H., Ogata H., Goto S.: ViPTree: the viral proteomic tree server. *Bioinformatics* 2017, 33, 2379–2380, doi: 10.1093/bioinformatics/btx157.
 78. Onsea J., Post V., Buchholz T., Schwegler H., Zeiter S., Wagemans J., Pimay J.P., Merabishvili M., D'Este M., Rotman S.G., Trampuz A., Verhofstad M.H.J., Obremskey W.T., Lavigne R., Richards R.G., Moriarty T.F., Metsemakers W.J.: Bacteriophage therapy for the prevention and treatment of fracture-related infection caused by *Staphylococcus aureus*: a preclinical study. *Microbiol Spectr* 2021, 9, e01736-21, doi: 10.1128/spectrum.01736-21.
 79. Paez-Espino D., Chen I.M.A., Palaniappan K., Ratner A., Chu K., Szeto E., Pillay M., Huang J., Markowitz V.M., Nielsen T., Huntemann M., Reddy T.B.K., Pavlopoulos G.A., Sullivan M.B., Campbell B.J., Chen F., McMahon K., Hallam S.J., Denev V., Cavicchioli R., Caffrey S.M., Streit W.R., Webster J., Handley K.M., Salekdeh G.H., Tesmetzis N., Setubal J.C., Pope P.B., Liu W.T., Rivers A.R., Ivanova N.N., Kyrpides N.C.: IMG/VR: a database of cultured and uncultured DNA viruses and retroviruses. *Nucleic Acids Res* 2017, 45, D457–D465, doi: 10.1093/nar/gkw1030.
 80. Palma M., Qi B.: Advancing phage therapy: a comprehensive review of the safety, efficacy, and future prospects for the targeted treatment of bacterial infections. *Infect Dis Rep* 2024, 16, 1127–1181, doi: 10.3390/idr16060092.
 81. Parras-Moltó M., Rodríguez-Galet A., Suárez-Rodríguez P., López-Bueno A.: Evaluation of bias induced by viral enrichment and random amplification protocols in metagenomic surveys of saliva DNA viruses. *Microbiome* 2018, 6, 119, doi: 10.1186/s40168-018-0507-3.
 82. Pires D., Melo L., Vilas Boas D., Sillankorva S., Azeredo J.: Phage therapy as an alternative or complementary strategy to prevent and control biofilm-related infections. *Curr Opin Microbiol* 2017, 39, 48–56, doi: 10.1016/j.mib.2017.09.004.
 83. Pons J.C., Paez-Espino D., Riera G., Ivanova N., Kyrpides N.C., Llabrés M.: VPF-Class: taxonomic assignment and host prediction of uncultivated viruses based on viral protein families. *Bioinformatics* 2021, 37, 1805–1813, doi: 10.1093/bioinformatics/btab026.
 84. Posit Team: RStudio: Integrated development environment for R. Posit Software, Boston, MA, USA, 2025.
 85. R Core Team: R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2025.
 86. Ramirez J., Guarner F., Bustos Fernandez L., Maruy A., Sdepanian V.L., Cohen H.: Antibiotics as major disruptors of gut microbiota. *Front Cell Infect Microbiol* 2020, 10, 572912, doi: 10.3389/fcimb.2020.572912.
 87. Rasmussen T.S., Koefoed A.K., Jakobsen R.R., Deng L., Castro-Mejía J.L., Brunse A., Neve H., Vogensen F.K., Nielsen D.S.: Bacteriophage-mediated manipulation of the gut microbiome – promises and presents limitations. *FEMS Microbiol Rev* 2020, 44, 507–521, doi: 10.1093/femsre/fuaa020.
 88. Ren J., Song K., Deng C., Ahlgren N.A., Fuhrman J.A., Li Y., Xie X., Poplin R., Sun F.: Identifying viruses from metagenomic data using deep learning. *Quant Biol* 2020, 8, 64–77, doi: 10.1007/s40484-019-0187-4.
 89. Rose R., Constantinides B., Tapinos A., Robertson D.L., Prosperi M.: Challenges in the analysis of viral metagenomes. *Virus Evol* 2016, 2, vew022, doi: 10.1093/ve/vew022.
 90. Roux S., Emerson J.B., Elie-Fadrosh E.A., Sullivan M.B.: Benchmarking viromics: an *in silico* evaluation of metagenome-enabled estimates of viral community composition and diversity. *PeerJ* 2017, 5, e3817, doi: 10.7717/peerj.3817.
 91. Roux S., Enault F., Hurwitz B.L., Sullivan M.B.: VirSorter: mining viral signal from microbial genomic data. *PeerJ* 2015, 3, e985, doi: 10.7717/peerj.985.
 92. Ruby J.G., Bellare P., DeRisi J.L.: PRICE: software for the targeted assembly of components of (meta) genomic sequence data. *G3 (Bethesda)* 2013, 3, 865–880, doi: 10.1534/g3.113.005967.
 93. Santos-Medellin C., Zinke L.A., ter Horst A.M., Gelardi D.L., Parikh S.J., Emerson J.B.: Viromes outperform total metagenomes in revealing the spatiotemporal patterns of agricultural soil viral communities. *ISME J* 2021, 15, 1956–1970, doi: 10.1038/s41396-021-00897-y.
 94. Sausset R., Petit M.A., Gaboriau-Routhiau V., De Paepe M.: New insights into intestinal phages. *Mucosal Immunol* 2020, 13, 205–215, doi: 10.1038/s41385-019-0250-5.
 95. Seemann T.: Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 2014, 30, 2068–2069, doi: 10.1093/bioinformatics/btu153.
 96. Shaffer M., Borton M.A., McGivern B.B., Zayed A.A., La Rosa S.L., Solden L.M., Liu P., Narrowe A.B., Rodríguez-Ramos J., Bolduc B., Gazitúa M.C., Daly R.A., Smith G.J., Vik D.R., Pope P.B., Sullivan M.B., Roux S., Wrighton K.C.: DRAM for distilling microbial metabolism to automate the curation of microbiome function. *Nucleic Acids Res* 2020, 48, 8883–8900, doi: 10.1093/nar/gkaa621.
 97. Shang J., Peng C., Liao H., Tang X., Sun Y.: PhaBOX: a web server for identifying and characterizing phage contigs in metagenomic data. *Bioinform Adv* 2023, 3, vbad101, doi: 10.1093/bioadv/vbad101.
 98. Shang J., Sun Y.: Predicting the hosts of prokaryotic viruses using GCN-based semi-supervised learning. *BMC Biol* 2021, 19, 250, doi: 10.1186/s12915-021-01180-4.
 99. Shang J., Tang X., Guo R., Sun Y.: Accurate identification of bacteriophages from metagenomic data using Transformer. *Brief Bioinform* 2022, 23, bbac258, doi: 10.1093/bib/bbac258.
 100. Shkoporov A.N., Stockdale S.R., Lavelle A., Kondova I., Houston C., Upadrasa A., Khokhlova E.V., van der Kamp I., Ouwerling B., Draper L.A., Langermans J.A.M., Paul Ross R., Hill C.: Viral biogeography of the mammalian gut and parenchymal organs. *Nat Microbiol* 2022, 7, 1301–1311, doi: 10.1038/s41564-022-01178-w.
 101. Sirén K., Millard A., Petersen B., Gilbert M.T.P., Clokie M.R.J., Sicheritz-Pontén T.: Rapid discovery of novel prophages using

- biological feature engineering and machine learning. *NAR Genom Bioinform* 2021, 3, lqaa109, doi: 10.1093/nargab/lqaa109.
102. Spiegelhauer M.R., Offersen S.M., Krych L., Nielsen D.S., Gambino M., Brunse A., Nguyen D.N.: Characterization of a virulent bacteriophage consortium targeting Enterobacteriaceae from inflamed preterm gut mucosa. *Future Virol* 2025, 20, 205–218, doi: 10.1080/17460794.2025.2497647.
 103. Starikova E.V., Tikhonova P.O., Prianchnikov N.A., Rands C.M., Zdobnov E.M., Ilina E.N., Govorun V.M.: Phigaro: high-throughput prophage sequence annotation. *Bioinformatics* 2020, 36, 3882–3884, doi: 10.1093/bioinformatics/btaa250.
 104. Steinegger M., Mirdita M., Söding J.: Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold. *Nat Methods* 2019, 16, 603–606, doi: 10.1038/s41592-019-0437-4.
 105. Tanaka T., Sugiyama R., Sato Y., Kawaguchi M., Honda K., Iwaki H., Okano K.: Precise microbiome engineering using natural and synthetic bacteriophages targeting an artificial bacterial consortium. *Front Microbiol* 2024, 15, 1403903, doi: 10.3389/fmicb.2024.1403903.
 106. Tancredi A., Matthijs T., Cox E., Van Immerseel F., Goossens E.: From mother to piglet: the lasting influence of the maternal microbiome. *Anim Microbiome* 2025, 7, 52, doi: 10.1186/s42523-025-00420-0.
 107. Tetz G., Brown S.M., Hao Y., Tetz V.: Parkinson's disease and bacteriophages as its overlooked contributors. *Sci Rep* 2018, 8, 10812, doi: 10.1038/s41598-018-29173-4.
 108. Tian F., Wainaina J.M., Howard-Varona C., Domínguez-Huerta G., Bolduc B., Gazitúa M.C., Smith G., Gittrich M.R., Zablocki O., Cronin D.R., Eveillard D., Hallam S.J., Sullivan M.B.: Prokaryotic-virus-encoded auxiliary metabolic genes throughout the global oceans. *Microbiome* 2024, 12, 159, doi: 10.1186/s40168-024-01876-z.
 109. Van Belleghem J.D., Dąbrowska K., Vaneechoutte M., Barr J.J., Bollyky P.L.: Interactions between bacteriophage, bacteria, and the mammalian immune system. *Viruses* 2019, 11, 10, doi: 10.3390/v11010010.
 110. Vaser R., Šikić M.: Time- and memory-efficient genome assembly with Raven. *Nat Comput Sci* 2021, 1, 332–336, doi: 10.1038/s43588-021-00073-4.
 111. Venturini C., Petrovic Fabijan A., Fajardo Lubian A., Barbirz S., Iredell J.: Biological foundations of successful bacteriophage therapy. *EMBO Mol Med* 2022, 14, EMMM202012435, doi: 10.15252/emmm.202012435.
 112. von Meijenfildt F.A.B., Arkhipova K., Cambuy D.D., Coutinho F.H., Dutilh B.E.: Robust taxonomic classification of uncharted microbial sequences and bins with CAT and BAT. *Genome Biol* 2019, 20, 217, doi: 10.1186/s13059-019-1817-x.
 113. Wang B., Luan Y.: Evaluation of normalization methods for predicting quantitative phenotypes in metagenomic data analysis. *Front Genet* 2024, 15, 1369628, doi: 10.3389/fgene.2024.1369628.
 114. Wang W., Hu H., Zijlstra R.T., Zheng J., Gänzle M.G.: Metagenomic reconstructions of gut microbial metabolism in weanling pigs. *Microbiome* 2019, 7, 48, doi: 10.1186/s40168-019-0662-1.
 115. Wang W., Ren J., Tang K., Dart E., Ignacio-Espinoza J.C., Fuhrman J.A., Braun J., Sun F., Ahlgren N.A.: A network-based integrated framework for predicting virus–prokaryote interactions. *NAR Genom Bioinform* 2020, 2, lqaa044, doi: 10.1093/nargab/lqaa044.
 116. Wang Y., Wei C., Chen Z., Zhou M., Huang L., Chen C.: Characterization of the diversity, genomic features, host bacteria, and distribution of crAss-like phages in the pig gut microbiome. *Front Vet Sci* 2025, 12, 1582122, doi: 10.3389/fvets.2025.1582122.
 117. Wei C., Chen Z.: Comprehensive analysis of phage genomes from diverse environments reveals their diversity, potential applications, and interactions with hosts and other phages. *Front Microbiol* 2025, 16, 1686402, doi: 10.3389/fmicb.2025.1686402.
 118. Wei C., Wang Y., Chen Z.: Comprehensive discovery and functional characterization of diverse prophages in the pig gut microbiome. *Front Microbiol* 2025, 16, 1662087, doi: 10.3389/fmicb.2025.1662087.
 119. Wei Y., Zhou C.: Bacteriophages: a double-edged sword in the gastrointestinal tract. *Front Microbiomes* 2024, 3, 1450523, doi: 10.3389/fmicb.2024.1450523.
 120. Wen T., Niu G., Chen T., Shen Q., Yuan J., Liu Y.X.: The best practice for microbiome analysis using R. *Protein Cell* 2023, 14, 713–725, doi: 10.1093/procel/pwad024.
 121. Wendling C.C., Vasse M., Wielgoss S.: Phage quest: a beginner's guide to explore viral diversity in the prokaryotic world. *Brief Bioinform* 2025, 26, bbaf449, doi: 10.1093/bib/bbaf449.
 122. Wernicki A., Nowaczek A., Urban-Chmiel R.: Bacteriophage therapy to combat bacterial infections in poultry. *Virol J* 2017, 14, 179, doi: 10.1186/s12985-017-0849-7.
 123. Wickham H.: *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, NY, USA, 2016, <https://ggplot2.tidyverse.org>.
 124. Wickham H., François R., Henry L., Müller K., Vaughan D.: *dplyr: a grammar of data manipulation*. R package version 1.1.4, 2023, Posit Software, Boston, MA, USA, 2023, <https://github.com/tidyverse/dplyr>, <https://dplyr.tidyverse.org>.
 125. Wood D.E., Lu J., Langmead B.: Improved metagenomic analysis with Kraken 2. *Genome Biol* 2019, 20, 257, doi: 10.1186/s13059-019-1891-0.
 126. Wu D., Liang S., Du X., Xiao J., Feng H., Ren Z., Yang X., Yang X.: Effects of fecal microbiota transplantation and fecal virome transplantation on LPS-induced intestinal injury in broilers. *Poult Sci* 2024, 103, 103316, doi: 10.1016/j.psj.2023.103316.
 127. Wu L.Y., Wijesekera Y., Piedade G.J., Pappas N., Brussaard C.P.D., Dutilh B.E.: Benchmarking bioinformatic virus identification tools using real-world metagenomic data across biomes. *Genome Biol* 2024, 25, 97, doi: 10.1186/s13059-024-03236-4.
 128. Yang X., Charlebois P., Gnerre S., Coole M.G., Lennon N.J., Levin J.Z., Qu J., Ryan E.M., Zody M.C., Henn M.R.: De novo assembly of highly diverse viral populations. *BMC Genomics* 2012, 13, 475, doi: 10.1186/1471-2164-13-475.
 129. Yin Y., Fischer D.: Identification and investigation of ORFans in the viral world. *BMC Genomics* 2008, 9, 24, doi: 10.1186/1471-2164-9-24.
 130. Young M.J., Hall L.M.L., Merabishvili M., Pirnay J.P., Clark J.R., Jones J.D.: Phage therapy for diabetic foot infection: a case series. *Clin Ther* 2023, 45, 797–801, doi: 10.1016/j.clinthera.2023.06.009.
 131. Yu M., Chu Y., Wang Y., Mo L., Tan X., Guo S., Yuan S., Ma Y.: Metagenomic analysis reveals gut phage diversity across three mammalian models. *Microbiome* 2025, 13, 146, doi: 10.1186/s40168-025-02144-4.
 132. Yüping L., Guan L., Becher I., Makarova K.S., Cao X., Hareendranath S., Guan J., Stein F., Yang S., Boergel A., Lapouge K., Remans K., Agard D., Savitski M., Typas A., Koonin E.V., Feng Y., Bondy-Denomy J.: Jumbo phage killer immune system targets early infection of nucleus-forming phages. *Cell* 2025, 188, 2127–2140, doi: 10.1016/j.cell.2025.02.016.
 133. Zhang H., Yohe T., Huang L., Entwistle S., Wu P., Yang Z., Busk P.K., Xu Y., Yin Y.: dbCAN2: a meta server for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res* 2018, 46, W95–W101, doi: 10.1093/nar/gky418.
 134. Zhang R., Mirdita M., Levy Karin E., Norroy C., Galiez C., Söding J.: SpacePHARER: sensitive identification of phages from CRISPR spacers in prokaryotic hosts. *Bioinformatics* 2021, 37, 3364–3366, doi: 10.1093/bioinformatics/btab222.
 135. Zheng T., Li J., Ni Y., Kang K., Misiakou M.A., Imamovic L., Chow B.K.C., Rode A.A., Bytzer P., Sommer M., Panagiotou G.: Mining, analyzing, and integrating viral signals from metagenomic data. *Microbiome* 2019, 7, 42, doi: 10.1186/s40168-019-0657-y.
 136. Zhu Y., Shang J., Peng C., Sun Y.: Phage family classification under *Caudoviricetes*: a review of current tools using the latest

- ICTV classification framework. *Front Microbiol* 2022, 13, 1032186, doi: 10.3389/fmicb.2022.1032186.
137. Zielezinski A., Barylski J., Karłowski W.M.: Taxonomy-aware, sequence similarity ranking reliably predicts phage–host relationships. *BMC Biol* 2021, 19, 223, doi: 10.1186/s12915-021-01146-6.
138. Zielezinski A., Deorowicz S., Gudyś A.: PHIST: fast and accurate prediction of prokaryotic hosts from metagenomic viral sequences. *Bioinformatics* 2022, 38, 1447–1449, doi: 10.1093/bioinformatics/btab837.