JOURNAL OF GENERAL VIROLOGY

RESEARCH ARTICLE

Cook *et al.*, *Journal of General Virology* 2024;105:001997
DOI 10.1099/jgv.0.001997

MICROBIOLOGY SOCIETY

OPEN ACCESS

Decoding huge phage diversity: a taxonomic classification of Lak megaphages

Ryan Cook[1,†], Marco A. Crisci[2,†], Hannah V. Pye[1,†], Andrea Telatin[1], Evelien M. Adriaenssens[1,*] and Joanne M. Santini[2,*]

Abstract

High-throughput sequencing for uncultivated viruses has accelerated the understanding of global viral diversity and uncovered viral genomes substantially larger than any that have so far been cultured. Notably, the Lak phages are an enigmatic group of viruses that present some of the largest known phage genomes identified in human and animal microbiomes, and are dissimilar to any cultivated viruses. Despite the wealth of viral diversity that exists within sequencing datasets, uncultivated viruses have rarely been used for taxonomic classification. We investigated the evolutionary relationships of 23 Lak phages and propose a taxonomy for their classification. Predicted protein analysis revealed the Lak phages formed a deeply branching monophyletic clade within the class *Caudoviricetes* which contained no other phage genomes. One of the interesting features of this clade is that all current members are characterised by an alternative genetic code. We propose the Lak phages belong to a new order, the 'Grandevirales'. Protein and nucleotide-based analyses support the creation of two families, three sub-families, and four genera within the order 'Grandevirales'. We anticipate that the proposed taxonomy of Lak megaphages will simplify the future classification of related viral genomes as they are uncovered. Continued efforts to classify divergent viruses are crucial to aid common analyses of viral genomes and metagenomes.

INTRODUCTION

Advancements in metagenomic sequencing have uncovered phage genomes greater than 200 kb (designated jumbo phages), and megaphages with genomes between ~500 kb and ~735 kb [1]. The colloquially named Lak phages are a group of large dsDNA phages which were first identified in human gut metagenomes from the Laksam Upazila, Bangladesh [2]. Genomes related to these megaphages have since been detected in geographically distinct gut metagenomes from humans and various animals including pigs, non-human primates, dogs, horses and tortoises [2, 3]. To date, 23 phylogenetically related Lak-like genomes (~476–660 kb) have been resolved to completion [2, 3]. However, the Lak phages remain uncultured. CRISPR-spacer targeting has indicated that Lak phages infect bacteria in the genus *Prevotella*, including the since reclassified *Segatella copri* (formerly *Prevotella copri* [4]), which are highly abundant in the gut microbiome of humans and animals that consume high-fibre and low-fat diets [2, 5]. The absence of integrases in the Lak phage genomes and no evidence of prophages in metagenomes suggests that these are virulent phages, rather than temperate phages that integrate into the bacterial host genome [3].

Although Lak phages are associated with gut microbiomes, some of the largest known complete phage genomes (~630–735 kb) were assembled from aquatic environment samples [1, 6]. For example, Mar_Mega_1 (650 kb) was recently assembled and found to be widespread across global marine samples [6]. Phylogenetic analyses of the Lak-like megaphages indicates that they form a single clade when compared to other dsDNA phages, with the marine megaphages belonging to a sister clade of the Lak phages. Mar_Mega_1 was proposed to represent a novel family, and forms a clade with LR756502 (642 kb) and LR745206 (635 kb), which were identified in freshwater metagenome samples from France and Japan, respectively [6].

Received 01 February 2024; Accepted 21 May 2024; Published 30 May 2024
Author affiliations: [1]Quadram Institute Bioscience, Norwich Research Park, Norwich, UK; [2]Department of Structural and Molecular Biology, Division of Biosciences, UCL, London, UK.
***Correspondence:** Evelien M. Adriaenssens, evelien.adriaenssens@quadram.ac.uk; Joanne M. Santini, j.santini@ucl.ac.uk
Keywords: lak; megaphage; phage taxonomy; phage genomics.
Abbreviations: ANI, average nucleotide identity; ICTV, International Committee on Taxonomy of Viruses; ORF, open reading frame; RVC, Royal Veterinary College; VMR, virus metadata resource.
†These authors contributed equally to this work
Eight supplementary tables are available with the online version of this article.

Lak megaphages have previously been predicted to use an alternative genetic code, which differs to that commonly used by bacterial species and most other known phages (genetic code 11), whereby the TAG stop codon is repurposed to encode glutamine (genetic code 15) [7]. However, phage recoding cannot be validated without evidence of functional protein translation [3, 8, 9]. Moreover, the phenomenon of stop-codon repurposing has been reported for ~5% of phage genomes recovered from human and animal gut metagenomes, and is suggested to prevent premature production of late-stage proteins and possibly regulate lysis [7]. Notably, some members of the *Crassvirales* are also thought to repurpose their stop codons [7, 10, 11]. However, this feature is not a characteristic of all *Crassvirales,* and only seems to occur within specific families such as the *Suoliviridae* [11]. Whether alternative code use is a conserved feature of all megaphages remains unknown.

Beyond their shared re-purposing of TAG codons, there are several parallels between the Lak megaphages and *Crassvirales*. First identified from human metagenomic data in 2014, the original crAssphage was thought to be a ubiquitous resident of the human gut [12]. Subsequently, an increasing number of crAss-like phages were identified from metagenomic data and shown to be highly abundant in the GI tract of adults that consume low-fibre and high-fat diets; promoting *Bacteroides*-rich enterotypes instead of *Prevotella*/*Segatella*-rich enterotypes [13, 14]. Despite their widespread abundance, isolation of the first crAss-like phage from *Bacteroides intestinalis* (ΦcrAss001) occurred in 2018 [15]. Since then, the order *Crassvirales* has been ratified by the International Committee on Taxonomy of Viruses (ICTV), an additional member has been cultured (ΦcrAss002), and the particle structure of ΦcrAss001 has recently been resolved [16, 17]. *Crassvirales* therefore offer an example of how analysis of uncultured viral genomes can lead to novel biological insights of ecologically significant viruses.

While the classification of prokaryotic viruses traditionally followed conservation of their morphology, this has been superseded by genomic-based approaches [18]. The introduction of a 15-rank virus hierarchy (species to realm) has most notably led to the abolition of the order Caudovirales (tailed phages), with all members reassigned to the class *Caudoviricetes* [18, 19]. This restructuring now enables megaphages to be assigned to higher taxonomic ranks, following ICTV guidelines alongside the four principles of establishing a universal viral taxonomy [20]. Therefore, the aim of this study was to investigate the evolutionary history of Lak phages and define taxa for the 23 Lak megaphages using genomic-based methods.

METHODS

Genomes

Complete phage genomes that had been previously resolved from metagenomic analysis of the gut microbiomes of humans [2, 3], baboons [2, 21], horses [3], dogs [22, 23] and pigs [3] were downloaded from ggKbase (University of California, Berkley) (Table S1, available in the online version of this article). These genomes were resolved to completion in their respective studies through manual curation and detection of termini by read mapping. Additional large unclassified phage genomes were extracted from the INPHARED database (February, 2023; Table S2) [24]. Emboss v6.6.0 was used to determine the length of each phage genome and the GC content (Tables S1 and S2) [25].

Predicted proteome analysis

A set of publicly available phage genomes (n=3539) belonging to the realm *Duplodnaviria*, which comprises all tailed bacteriophages, was extracted from the INPHARED database (February, 2023) [24], and the up-to-date taxonomy for each genome was extracted from the Virus Metadata Resource (VMR; https://ictv.global/vmr). The classified members of the *Duplodnaviria* were combined with the genomes used in this study for input with the standalone version of ViPTree v1.1.2 [26], and the output was visualised with IToL [27]. To determine the potential effect of alternative TAG codon usage on coding capacity, open reading frames (ORFs) were predicted on all genomes using Prodigal v2.6.3 with both translation tables 11 and 15 used separately on all genomes [28]. Coding capacity was calculated as the sum length of predicted ORFs as a percentage of total genome length.

Core genome analysis

Open reading frames (ORFs) were predicted using Prodigal-gv v2.11.0-gv, a modified version of Prodigal that predicts alternative codon usage and has been optimised for gene identification on virus genomes (https://github.com/apcamargo/prodigal-gv) [28–30]. The translated ORFs were clustered using MMseqs2 v13.45111 at 70% identity with the --cluster-mode 1 flag [31]. Sequences from each cluster were compared to publicly available MMseqs2 profiles of the PHROGs database (https://phrogs.lmge.uca.fr/downloads_from_website/phrogs_mmseqs_db.tar.gz) using mmseqs search with -e 1E-05 and the best hit per ORF was retained [31, 32]. The most frequent PHROG hit was used to infer the function of the cluster. Presence/absence of proteins clusters were plotted using seaborn v0.12.2 [33].

Alignments were produced for the 72 protein clusters found on the 22 genomes belonging to the proposed family 'Lakviridae' using MAFFT v7.520 [34], and the alignments were used as input for IQ-TREE v2.2.2.3 with -B 1000 and -m TEST [35]. As the IQ-TREE '-m TEST' flag optimises models for individual alignments before combining into a final model, multiple models were used in the final tree (Table S3). The resulting concatenated protein maximum likelihood phylogenetic tree was visualised using IToL [27].

Intergenomic similarity of Lak phages

The intergenomic similarity of the 23 Lak phage genomes was deduced by multiplying the average nucleotide identity (ANI) with the aligned fraction of the genome for each pair of genomes. ANI was determined using fastANI v1.33 [36] for all genomes using the many-to-many parameter, for multiple reference and query genomes. The aligned fraction was calculated by dividing the number of fragments that were aligned as orthologous matches between each genome by the total number of sequence fragments for each genome, both of which were computed within the fastANI output. The ANI value was then multiplied by the aligned fraction to compute the intergenomic similarity. The average intergenomic similarity between genomic pairs was visualised in R v4.2.2 using pheatmap v1.0.12 [37].

RESULTS

This study included 56 phage genomes >200 kb, including 23 so-called Lak phages that are not currently classified to any taxonomic rank (Table S1). For the purposes of this study, only complete genomes were included. The 56 genomes ranged from 203 to 735 kb in length and 24–55 molGC (%), with the 23 Lak phages ranging from 476 to 660 kb and 26–31 molGC (%).

The 56 genomes included in this analysis all belonged to prokaryotic viruses with dsDNA genomes which were predicted to encode the set of proteins that are characteristic of the newly ratified class *Caudoviricetes*. These features include putative tail proteins, a major capsid protein with the HK97 fold, a portal protein, and the terminase large subunit. Based on the presence of these proteins, all 56 phages are automatically assigned to the class *Caudoviricetes* and this analysis therefore sought to further classify these phages from the rank of order to species. To infer the order and family rank, a proteomic approach was used. The 56 genomes were processed using VipTree alongside all currently classified members of *Caudoviricetes* (n = 3539; February 2023), producing a hierarchically clustered tree based on pairwise tBLASTx scores (Fig. 1). The 23 Lak phages formed a deeply branching monophyletic clade that contained no other genomes, with Mar_mega_1 being the only genome in its nearest sister clade (Fig. 1a). Of the remaining 32 large phage genomes, 14 were interspersed with known genomes belonging to sister clades of the Lak phages, and the remaining 18 were very distant on the tree (Fig. 1a).

Lak phages are thought to re-purpose TAG codons to encode glutamine, rather than a stop codon, as this is observed elsewhere in biology [38]. To determine if this feature is unique to and conserved within Lak phages, we predicted ORFs on the 56 genomes

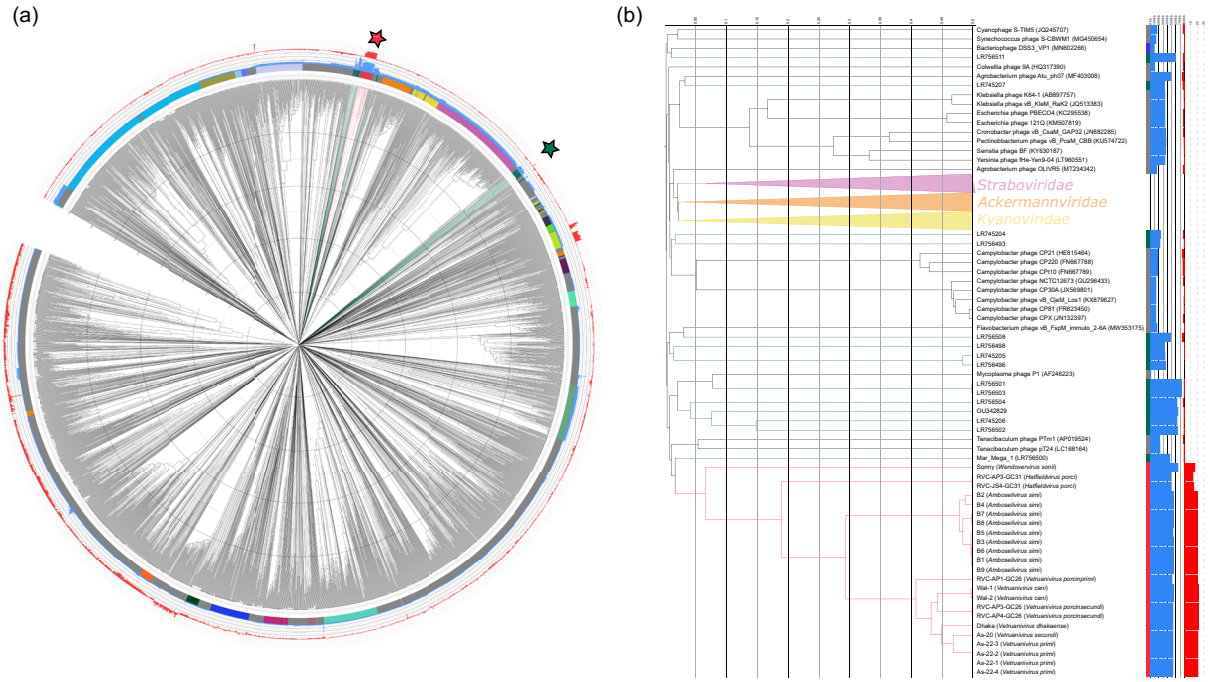

Fig. 1. Proteomic tree of megaphages amongst currently classified *Duplodnaviria*. (a) ViPTree proteomic tree of 'megaphages' and *Duplodnaviria* with viral family shown in the coloured ring. Blue bar chart (inner) represents genome length and red bar chart (outer) shows difference in coding capacity when using translation table 15 rather than 11 (i.e. coding capacity of 90% using 15 and 70% using 11 would lead to a difference of 20). The 'Grandevirales' members are shown with red branches and other megaphages are shown with green and highlighted by a star with corresponding colour. (b) A pruned tree showing the 'Grandevirales' with nearest sister clades only. The distances shown in ViPTree were calculated from genomic distances based upon normalised tBLASTx scores and the tree was rooted at the mid-point.

used in this analysis and the 3539 representative members of the realm *Duplodnaviria* using both translation table 11 (standard bacterial) and table 15 (re-purposed TAG). The mean coding capacity of the 23 Lak phages with translation table 11 was 69% (SD ±1.9) and increased to 89% (SD ±1.2) when using translation table 15 (Figure 1B; Table S4). This feature was conserved among all Lak phages but was not observed in Mar_Mega_1 (table 11 94%, table 15 93%). Considering the distance to other phages, and conserved alternative codon use amongst this clade, we propose the Lak phages represent a new order of phages and suggest the name 'Grandevirales'. As the 23 proposed members of 'Grandevirales' were highly divergent from other phages used in this study, only these 23 were carried forward for taxonomic classification.

To infer taxonomy at and below the rank of family, we performed a core protein analysis on the proposed members of 'Grandevirales'. The 23 members of 'Grandevirales' shared no core proteins across all genomes at 70% amino acid sequence identity (Fig. 2). The genome of phage Sonny (OR769223) was a clear outlier with the predicted proteome vastly different from the remaining 22 phages. Only when the threshold for amino acid identity was lowered to 25% did this phage genome share one core gene with the 22 others. We therefore suggest that this phage represents a new family and propose the name 'Epsomviridae'. The other 22 'Grandevirales' shared 72 core proteins at 70% amino acid identity, the vast majority of which have unknown functions (Fig. 2; Tables S5, S6 and S7). Of 1623 unique protein clusters within this group, there was a median of 583 proteins per genome and the 72 core proteins represent a mean average of 12.5% (SD ±0.8%). We propose that these phages belong to a second new family, 'Lakviridae'. The proposed 'Lakviridae' consist of two clear groupings that are distinct from one another at a higher level than genus (Fig. 2). Therefore, we suggest the 'Lakviridae' is divided into two subfamilies, 'Quadringentisvirinae' and 'Quingentivirinae', with 479 and 218 core proteins in each respectively (Table S6). Encouragingly, this pattern is congruent with that of the VipTree analysis performed for higher level classification (Fig. 1).

The 72 proteins core to the proposed 'Lakviridae' were aligned and used as input for a concatenated protein phylogeny. This analysis revealed two deeply branching clades that mirror the results of the core genome analysis, lending further support to the creation of subfamilies 'Quadringentisvirinae' and 'Quingentivirinae' (Fig. 3). The proposed 'Quingentivirinae' form two distinct clades that we suggest represent two separate genera (Fig. 3). One clade, the 'Amboselivirus', consists of the nine highly similar 'Lak B' phages that were identified from baboon samples (Fig. 3). The other clade, 'Vetruanivirus', consists of 11 genomes identified from separate studies of human and pig samples (Fig. 3). The proposed genera were comprised of large core proteomes when examining the protein clusters used in core proteome analyses, with sizes of 546, 479 and 388 core proteins for 'Amboselivirus', 'Hatfieldvirus' and 'Vetruanivirus' respectively (Fig. 2; Table S6).

To elucidate species level phylogenetic relationships, we investigated the intergenomic similarity of the proposed 'Grandevirales'. The 23 'Grandevirales' phages formed four distinct clusters with varying intergenomic similarity (Fig. 4). Three of the genomes (As-22–2, As-22–1 and As-22–4) shared almost 100% identity (Fig. 4; Table S8). Four phages were classified into the species 'Vetruanivirus primi', due to an intergenomic similarity score of >95%. According to the same criteria, two phages were assigned to each species of 'Vetruanivirus porcinsecundi' and 'Vetruanivirus cani'. Eleven phages were classified as the genus 'Vetruanivirus' as they were all within the typical genus demarcation of 70%, as defined by the ICTV. Another set of nine phages, all isolated from the Baboon gut (B1-B9), made up the sole species ('Amboselivirus simi') of the currently proposed

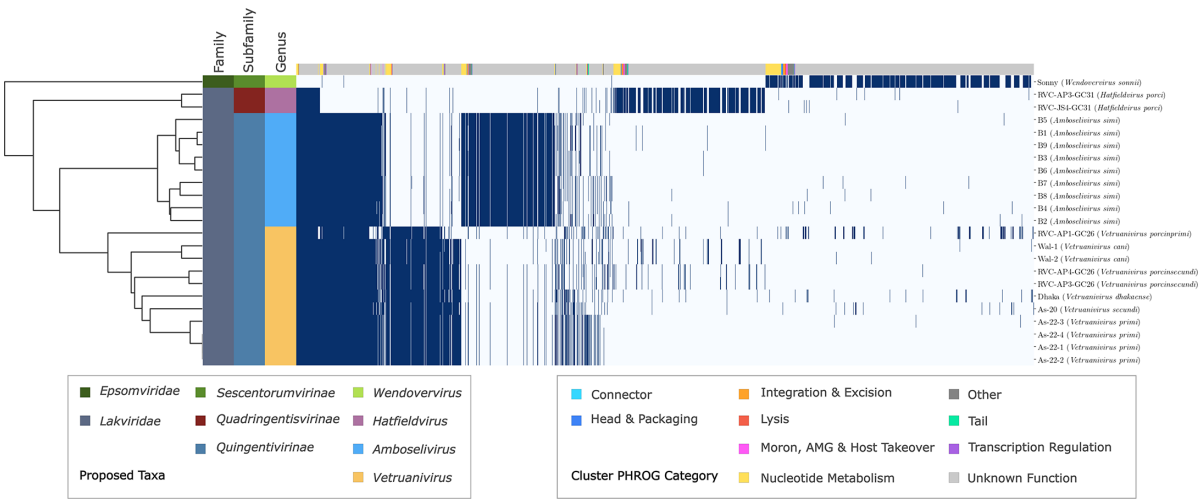

Fig. 2. Shared protein clusters for phages of 'Grandevirales'. Heatmap showing presence/absence of proteins clustered at 70% identity with the dendrogram showing hierarchical clustering. Colour strips on the y-axis show proposed taxonomy, and colour strips on the x-axis show predicted function of the protein cluster derived from PHROGs. X-axis labels show strain name with proposed species name in brackets.

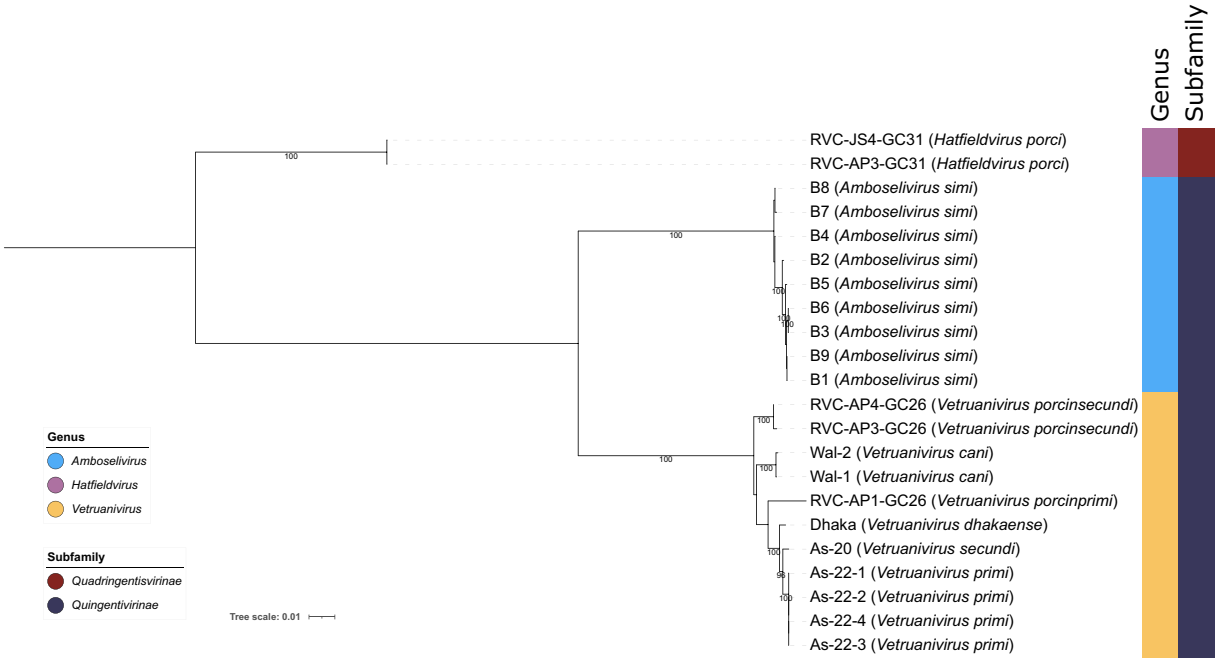

Fig. 3. Core genome phylogeny of 'Lakviridae' phages. A concatenated protein phylogeny of translated sequences of 72 'core' genes present on the 22 members of the proposed family 'Lakviridae'. Proposed species names for each strain are indicated in brackets. Alignments were performed using MAFFT, and the tree was produced using IQ-Tree with 1000 rapid bootstraps and -m TEST to optimize model fits for each alignment. Tree is rooted at the midpoint and bootstraps ≥95% are shown. The coloured strips indicate proposed genera and subfamilies. Node labels are based on strain names with proposed species names shown in brackets.

'Amboselivirus' genus (Fig. 4; Table S8). Almost all of these Baboon-associated phage genomes had a pairwise identity of >95%, and the three genomes with marginally less than 95% pairwise identity (B4 compared to B1, and B4 compared to B9) still cluster distinctly with other members of the 'Amboselivirus' genus within 95% identity (Fig. 4), hence why we chose to classify them into the same species.

Only two phages formed the 'Hatfieldvirus porci' species (~99.36% intergenomic similarity) within the 'Hatfieldvirus' genus. Interestingly, members of this genus share little similarity (< 20%) with members of the 'Amboselivirus' and 'Vetruanivirus' genera (Fig. 4; Table S8). The sole member of the 'Wendovervirus sonii' species shares no detectable similarity with any other phages included in the analysis, and hence further supports our reasoning to designate this phage as a member of a separate family ('Epsomviridae') (Fig. 4; Table S8). All strain names used in the current study have remained consistent with the nomenclature used in their respective publications [2, 3].

The computational analysis of the previously published 23 Lak-like megaphage genomes has led to the proposed formation of the order 'Grandevirales', encompassing two families, 'Lakviridae' and 'Epsomviridae', three sub-families and four separate genera (Table 1).

Rationale and justification of taxonomic names

All taxonomic names proposed in this manuscript follow the guidance suggested by Postler *et al.* [39], with all proposed species using a latinised binomial name [39].

Order 'Grandevirales'

The name 'Grandevirales' is proposed for the order, as Grande means large (or great) in Latin and in many other European languages. Based on the present analyses, some related phages in this order have genomes under 500 kb, therefore it would be unsuitable to denote the order as 'Mega', given that 'megaphage' has historically been used to describe phages with genomes over 500 kb [2]. Furthermore, the order *Megavirales* has been used to describe large eukaryotic viruses (>100 kb) [40]. Given that all phages within this order are 'huge', i.e. >200 kb, and that the order encompasses the largest known complete phage genomes, the order name 'Grandevirales' was chosen.

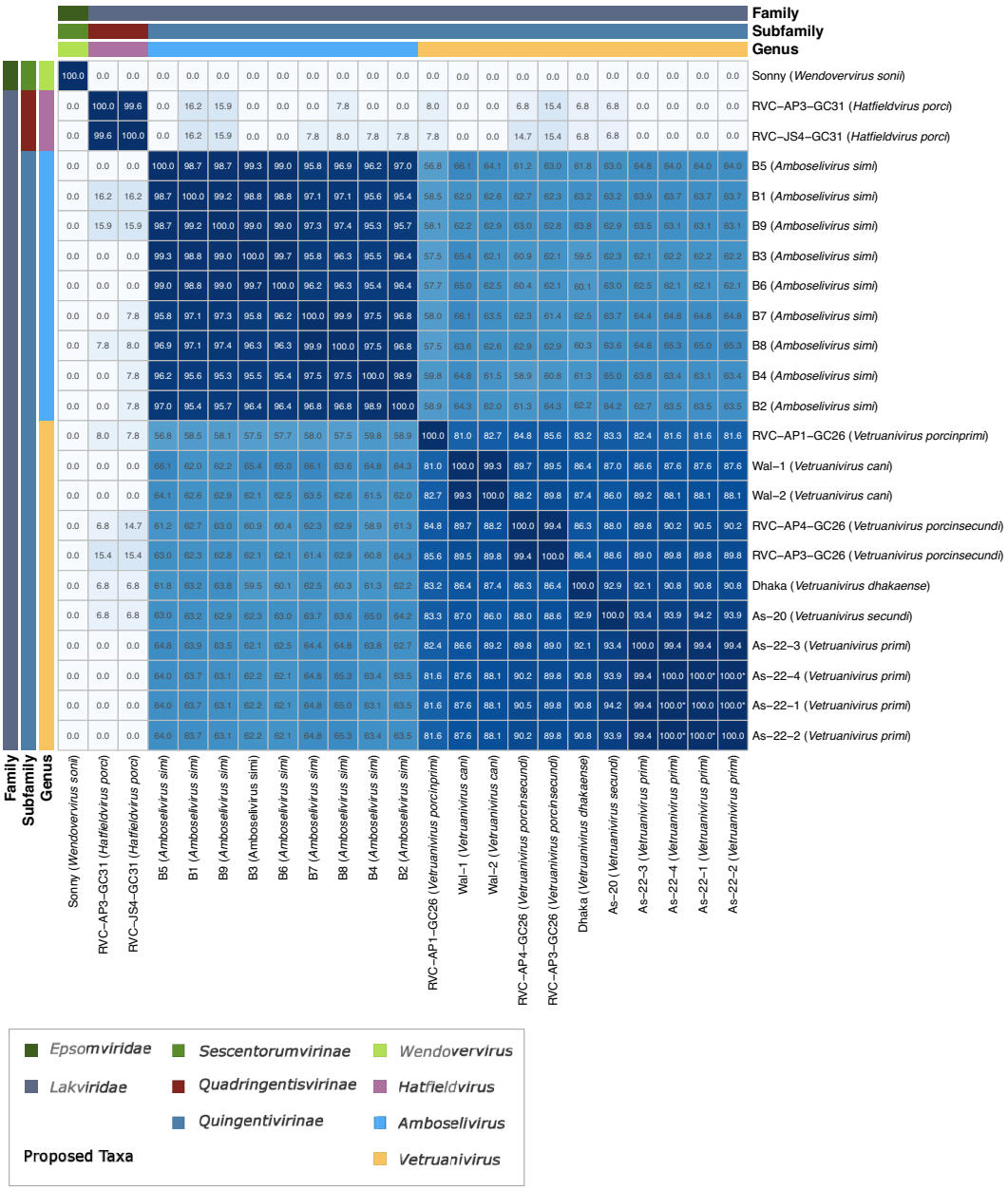

Fig. 4. Intergenomic similarity of 'Grandevirales' phages. Heatmap showing intergenomic distance (ANI multiplied by aligned fraction) of 'Grandevirales' phages with proposed taxonomy shown in coloured strips. Those shown as having a similarity of 100.0* are highly similar but not identical with a low number of single nucleotide variations between them. Axis labels are based on strain names with proposed species names shown in brackets.

Family 'Lakviridae'

The family 'Lakvirdae' was selected as the founding members of this family (A1 and A2) were all identified from the gut microbiomes of people consuming arsenic-contaminated water in Laksam Upazila, Bangladesh [2].

Family 'Epsomviridae'

The second family proposed in this study is 'Epsomviridae', named after Epsom, Surrey (UK). Epsom is a town famous for horse racing and is the location where the racehorse 'Sonny' was stabled and trained. In the current proposal, only one megaphage belongs to this family, which was identified from the horse gut microbiome [3]. Sonny was therefore used for the species and strain name.

Table 1. Proposed taxonomic classification of Lak-like megaphages

Order	*Grandevirales*								
Family	*Lakviridae*								*Epsomviridae*
Subfamily	*Quingentivirinae*							*Quadringentisvirinae*	*Sescentorumvirinae*
Genus	*Vetruanivirus*						*Amboselivirus*	*Hatfieldvirus*	*Wendovervirus*
Species	*Vetruanivirus primi*	*Vetruanivirus secundi*	*Vetruanivirus dhakaense*	*Vetruanivirus porcinprimi*	*Vetruanivirus porcinsecundi*	*Vetruanivirus cani*	*Amboselivirus simi*	*Hatfieldvirus porci*	*Wendovervirus sonnii*
Strains	As-22-1 As-22-2 As-22-3 As-22-4	As-20	Dhaka	RVC-AP1-GC26	RVC-AP4-GC26 RVC-AP3-GC26	Wal-1 Wal-2	B1-B9	RVC-AP3-GC31 RVC-JS4-GC31	Sonny

Subfamilies 'Quingenti-', 'Quadringenti-' and 'Sescentorum-' -virinae

Subfamily names were chosen based on genome sizes of the founding members of each subfamily. 'Quingenti', 'Quadringentis', and 'Sescentorum' are Latin for 400, 500 and 600, respectively. For 'Quingentivirinae', founding members have genome lengths within the 500 kb range. For 'Quadringentisvirinae', current members have ~476 kb genomes, and one phage has currently been classified into the subfamily 'Sescentorumvirinae', with a ~660 kb genome. Although the genome lengths of founding members have been used as the basis of subfamily names, this is not a criterion for taxonomic classification and members of each subfamily form distinct clusters within 'Lakviridae' and 'Epsomviridae'.

Genus 'Vetruanivirus'

The genus name 'Vetruanivirus' is an amalgamation of the words Veterinary and Eruani, which encompasses the isolation source of current phages in this genus. Some phages were isolated from pigs at the Royal Veterinary College (RVC), hence "*Vet-* ", whilst other strains were isolated from individuals living in a village called Eruani in Laksam, Bangladesh, hence 'ruani-" which forms 'Vetruanivrus'.

Genus 'Amboselivirus'

Phage genomes belonging to the proposed 'Amboselivirus' genus (B1-B9) were resolved from faecal samples collected from Kenyan yellow baboons living in the Amboseli national park. The current sole species has been given the name 'Amboselivirus simi' ('simia' is Latin for primate/monkey, with simi in the genitive form). Strains within the proposed genus 'Amboselivirus' (B1-B9) are maintained as described in the original publication [2].

Genus 'Hatfieldvirus'

Two of the phage genomes identified from pig samples formed the 'Hatfieldvirus' genus, which has been named according to the sampling location. These phages were identified from faecal samples from pigs reared and cared for at the RVC in Hatfield, Hertfordshire (UK). The sole species name has been given as, 'Hatfieldvirus porci', as porci is Latin for pig. Strains of this proposed genus (and those predicted to be placed in it), have so far only been found in pig gastrointestinal tracts.

Genus 'Wendovervirus'

The sole genome belonging to this genus, Sonny, was assembled from sequencing data of a microbiome sample from a horse stabled at the Wendover stables in Epsom [3].

DISCUSSION

An increase in the number of high-quality curated phage genomes has transformed phage taxonomy, and morphology-based classification has been superseded with robust genomic frameworks [18, 41, 42]. In this study, 23 Lak phages were taxonomically classified via comprehensive pangenome analysis, concatenated protein phylogeny, and analysis of their intergenomic similarity. The taxonomy and analysis of these Lak phages – which we can now call grandeviruses – posed several challenges related to their origin, alternative codon usage and large genome sizes.

The megaphage genomes classified in this study were resolved from metagenomes and the phages themselves remain uncultured, due to the difficulty in isolating these large phages from biological samples, as also described with *Crassvirales* [16, 17]. Phage cultivability is dependent on identification of the bacterial host and optimal growth conditions, neither of which can be easily determined from metagenomes despite a plethora of tools developed to aid in the identification of these phage-host pairings, such as iPHoP (which uses RaFAH, WIsH, oligonucleotide frequencies, PHP, and BLAST), HostPhinder and PHERI [43–51]. Alternative methods to isolate phages with large genomes have emerged, and suggest the use of filters with pore sizes >0.2 µm, and decreasing the concentration of the overlay agar used [52].

Pangenomes are often used to classify bacterial taxa according to the presence of core and accessory genes in bacterial strains. During the current study, multiple pangenome construction tools, including roary [53], Panaroo [54] and ggcaller [55], were used to generate a pangenome of the 23 megaphages and related huge phages. However, we found that all of these tools were inappropriate for this particular group of sequences. Many published tools do not support translation table 15, the stop codon reassignment that is suggested in these phages. Furthermore, we found no available tools that could construct the pangenome of sequences that use different translation tables to one another (i.e. determining the pangenome of megaphages that use translation table 15, alongside their nearest relatives that use translation table 11). Therefore, a manual method for elucidating megaphage pangenomes was devised that used prodigal-gv to account for the alternative codon usage [29]. Circumventing the need for annotation, we also used tBLASTx-based pairwise comparisons of the genomes as implemented in the Viral Proteomic Tree software VipTree [56], which was not sensitive to the stop codon change.

While the pangenome tools developed for bacteria were not compatible with the alternative codon use of the 'Grandevirales' phages, the virus tools struggled with their genome sizes. The sheer size of the megaphage genomes in question (476–660 kb) made it difficult to use phage and virus-specific published tools to determine viral intergenomic similarity, such as VIRIDIC [57] and VirClust [58]. We therefore manually implemented the general methodology used by VIRIDIC to compute intergenomic similarity, ensuring that the nucleotide identity was normalised according to the aligned fraction of the genome to avoid exaggerated intergenomic similarity scores [57]. Interestingly, there was a low percentage (6.8–16.2%) of intergenomic similarity observed between phages belonging to different genera. This warrants further investigation using comparative genomics to elucidate the origin of these stretches of sequence identity and determine whether they are the result of recombination or horizontal gene transfer between phages or the bacterial host.

We were able to make a few observations with potential biological relevance that could be topics of further investigation. All grandeviruses were identified from gut microbiomes of humans and mammals associated with humans (evolutionarily or through close contact). Phages belonging to each species were isolated from the same source, for example all nine strains of 'Amboselivirus simi' were discovered in baboon gut microbiomes, reflected in their chosen species name. In contrast, at the genus rank, multiple microbiome hosts were observed for the genus 'Vetruanivirus'. Further identification and classification of these phages in microbiomes and, hopefully, isolation in culture will uncover their function in the gut and their niche-adaptation in different animals.

CONCLUSION

The results presented here, combined with previously published work, provide robust evidence that the Lak phage clade is monophyletic compared to other known phage genomes and justifies the creation of a new viral order, 'Grandevirales'. The order encompasses some of the largest phage genomes ever reported and can be further sub-divided into two new families and three subfamilies, according to concatenated protein phylogeny and intergenomic similarity. Four novel genera have also been proposed, encompassing 23 phage strains. Overall, this study has overcome the challenges associated with the classification of 'megaphages', has successfully classified phages resolved from metagenomes, and provided justification for the classification of huge viral genomes based on a shared core genome, intergenomic similarity and phylogeny.

Funding information
R.C. and E.M.A. are funded through the Biotechnology and Biological Sciences Research Council (BBSRC) grant Bacteriophages in Gut Health BB/W015706/1. M.A.C. was funded by the BBSRC through the London Interdisciplinary Doctoral Programme BB/M009513/1. H.V.P. and E.M.A. are funded through the Medical Research Council grant MR/W031205/1. E.M.A. gratefully acknowledges the support of the BBSRC; this research was funded by the BBSRC Institute Strategic Programme Food Microbiome and Health BB/X011054/1 and its constituent projects BBS/E/F/000PR13631 and BBS/E/F/000PR13633; and by the BBSRC Institute Strategic Programme Microbes and Food Safety BB/X011011/1 and its constituent projects BBS/E/F/000PR13634, BBS/E/F/000PR13635 and BBS/E/F/000PR13636. Bioinformatics analysis was carried out on infrastructure provided by MRC-CLIMB (MR/L015080/1). For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising.

Author contributions
Conceptualisation: M.A.C., E.M.A., J.M.S.; Data curation: R.C., M.A.C., H.V.P.; Formal analysis: R.C., M.A.C., H.V.P., A.T.; Funding acquisition: J.M.S.; Supervision: E.M.A., J.M.S.; Writing – original draft preparation: R.C., M.A.C., H.V.P., A.T., E.M.A., J.M.S.; Writing – review and editing: E.M.A., J.M.S.

Conflicts of interest
The authors have no conflicts to declare.

References

1. Al-Shayeb B, Sachdeva R, Chen L-X, Ward F, Munk P, *et al*. Clades of huge phages from across Earth's ecosystems. *Nature* 2020;578:425–431.
2. Devoto AE, Santini JM, Olm MR, Anantharaman K, Munk P, *et al*. Megaphages infect Prevotella and variants are widespread in gut microbiomes. *Nat Microbiol* 2019;4:693–700.
3. Crisci MA, Chen L-X, Devoto AE, Borges AL, Bordin N, *et al*. Closely related Lak megaphages replicate in the microbiomes of diverse animals. *iScience* 2021;24:102875.
4. Hitch TCA, Bisdorf K, Afrizal A, Riedel T, Overmann J, *et al*. A taxonomic note on the genus *Prevotella*: description of four novel genera and emended description of the genera *Hallella* and *Xylanibacter*. *Syst Appl Microbiol* 2022;45:126354.
5. PrasoodananPKV, Sharma AK, Mahajan S, Dhakan DB, Maji A, *et al*. Western and non-western gut microbiomes reveal new roles of *Prevotella* in carbohydrate metabolism and mouth-gut axis. *NPJ Biofilms Microbiomes* 2021;7:77.
6. Michniewski S, Rihtman B, Cook R, Jones MA, Wilson WH, *et al*. A new family of "megaphages" abundant in the marine environment. *ISME Commun* 2021;1:58.
7. Borges AL, Lou YC, Sachdeva R, Al-Shayeb B, Penev PI, *et al*. Widespread stop-codon recoding in bacteriophages may regulate translation of lytic genes. *Nat Microbiol* 2022;7:918–927.
8. Ivanova NN, Schwientek P, Tripp HJ, Rinke C, Pati A, *et al*. Stop codon reassignments in the wild. *Science* 2014;344:909–913.
9. Peters SL, Borges AL, Giannone RJ, Morowitz MJ, Banfield JF, *et al*. Experimental validation that human microbiome phages use alternative genetic coding. *Nat Commun* 2022;13:5710.
10. Gulyaeva A, Garmaeva S, Ruigrok RAAA, Wang D, Riksen NP, *et al*. Discovery, diversity, and functional associations of crAss-like phages in human gut metagenomes from four Dutch cohorts. *Cell Rep* 2022;38:110204.
11. Yutin N, Benler S, Shmakov SA, Wolf YI, Tolstoy I, *et al*. Analysis of metagenome-assembled viral genomes from the human gut reveals diverse putative CrAss-like phages with unique genomic features. *Nat Commun* 2021;12:1044.
12. Dutilh BE, Cassman N, McNair K, Sanchez SE, Silva GGZ, *et al*. A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. *Nat Commun* 2014;5:4498.

13. Ley RE. Prevotella in the gut: choose carefully. *Nat Rev Gastroenterol Hepatol* 2016;13:69–70.
14. Guerin E, Shkoporov A, Stockdale SR, Clooney AG, Ryan FJ, *et al*. Biology and taxonomy of crAss-like bacteriophages, the most abundant virus in the human gut. *Cell Host Microbe* 2018;24:653–664.
15. Shkoporov AN, Khokhlova EV, Fitzgerald CB, Stockdale SR, Draper LA, *et al*. ΦCrAss001 represents the most abundant bacteriophage family in the human gut and infects *Bacteroides intestinalis*. *Nat Commun* 2018;9:1–8.
16. Bayfield OW, Shkoporov AN, Yutin N, Khokhlova EV, Smith JLR, *et al*. Structural atlas of a human gut crassvirus. *Nature* 2023;617:409–416.
17. Guerin E, Shkoporov AN, Stockdale SR, Comas JC, Khokhlova EV, *et al*. Isolation and characterisation of ΦcrAss002, a crAss-like phage from the human gut that infects *Bacteroides xylanisolvens*. *Microbiome* 2021;9:89.
18. Turner D, Shkoporov AN, Lood C, Millard AD, Dutilh BE, *et al*. Abolishment of morphology-based taxa and change to binomial species names: 2022 taxonomy update of the ICTV bacterial viruses subcommittee. *Arch Virol* 2023;168:74.
19. Adriaenssens EM. Phage diversity in the human gut microbiome: a taxonomist's perspective. *mSystems* 2021;6:e0079921.
20. Walker PJ, Siddell SG, Lefkowitz EJ, Mushegian AR, Adriaenssens EM, *et al*. Recent changes to virus taxonomy ratified by the International Committee on Taxonomy of Viruses (2022). *Arch Virol* 2022;167:2429–2440.
21. Tung J, Barreiro LB, Burns MB, Grenier J-C, Lynch J, *et al*. Social networks predict gut microbiome composition in wild baboons. *Elife* 2015;4:e05224.
22. Allaway D, Haydock R, Lonsdale ZN, Deusch OD, O'Flynn C, *et al*. Rapid reconstitution of the fecal microbiome after extended diet-induced changes indicates a stable gut microbiome in healthy adult dogs. *Appl Environ Microbiol* 2020;86:e00562-20.
23. Edgar RC, Taylor J, Lin V, Altman T, Barbera P, *et al*. Petabase-scale sequence alignment catalyses viral discovery. *Nature* 2022;602:142–147.
24. Cook R, Brown N, Redgwell T, Rihtman B, Barnes M, *et al*. INfrastructure for a PHAge REference database: identification of large-scale biases in the current collection of cultured phage genomes. *Phage*2021;2:214–223.
25. Rice P, Longden I, Bleasby A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* 2000;16:276–277.
26. Nishimura Y, Yoshida T, Kuronishi M, Uehara H, Ogata H, *et al*. ViPTree: the viral proteomic tree server. *Bioinformatics* 2017;33:2379–2380.
27. Letunic I, Bork P. Interactive tree of life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res* 2019;47:W256–W259.
28. Hyatt D, Chen G-L, Locascio PF, Land ML, Larimer FW, *et al*. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 2010;11:1–11.
29. Camargo AP, Roux S, Schulz F, Babinski M, Xu Y, *et al*. Identification of mobile genetic elements with geNomad. *Nat Biotechnol* 2023.
30. Cook R, Telatin A, Bouras G, Camargo AP, Larralde M, *et al*. Predicting stop codon reassignment improves functional annotation of bacteriophages. *bioRxiv* 2023.
31. Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol* 2017;35:1026–1028.
32. Terzian P, Olo Ndela E, Galiez C, Lossouarn J, Pérez Bucio RE, *et al*. PHROG: families of prokaryotic virus proteins clustered using remote homology. *NAR Genom Bioinform* 2021;3.
33. Waskom ML. seaborn: statistical data visualization. *JOSS* 2021;60.
34. Nakamura T, Yamada KD, Tomii K, Katoh K. Parallelization of MAFFT for large-scale multiple sequence alignments. *Bioinformatics* 2018;34:2490–2492.
35. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* 2015;32:268–274.
36. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat Commun* 2018;9:5114.
37. Kolde R. Pheatmap: pretty heatmaps. *R package version* 2012;1:726.
38. Keeling PJ, Doolittle WF. Widespread and ancient distribution of a noncanonical genetic code in diplomonads. *Mol Biol Evol* 1997;14:895–901.
39. Postler TS, Rubino L, Adriaenssens EM, Dutilh BE, Harrach B, *et al*. Guidance for creating individual and batch latinized binomial virus species names. *J Gen Virol* 2022;103.
40. Colson P, de Lamballerie X, Fournous G, Raoult D. Reclassification of giant viruses composing a fourth domain of life in the new order megavirales. *Intervirology* 2012;55:321–332.
41. Turner D, Kropinski AM, Adriaenssens EM. A roadmap for genome-based phage taxonomy. *Viruses* 2021;13:506.
42. Simmonds P, Adriaenssens EM, Zerbini FM, Abrescia NGA, Aiewsakun P, *et al*. Four principles to establish a universal virus taxonomy. *PLOS Biol* 2023;21:e3001922.
43. Villarroel J, Kleinheinz KA, Jurtz VI, Zschach H, Lund O, *et al*. HostPhinder: a phage host prediction tool. *Viruses* 2016;8:116.
44. Baláž A, Kajsik M, Budiš J, Szemes T, Turňa J. PHERI-phage host exploRation pipeline. *Microorganisms* 2023;11:1398.
45. Ostenfeld LJ, Munk P, Aarestrup FM, Otani S. Detection of specific Uncultured Bacteriophages by fluorescence in situ Hybridisation in pig Microbiome. *PLoS One* 2023;18:e0283676.
46. Roux S, Camargo AP, Coutinho FH, Dabdoub SM, Dutilh BE, *et al*. iPHoP: an integrated machine learning framework to maximize host prediction for metagenome-derived viruses of archaea and bacteria. *PLoS Biol* 2023;21:e3002083.
47. 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.
48. Coutinho FH, Zaragoza-Solas A, López-Pérez M, Barylski J, Zielezinski A, *et al*. RaFAH: host prediction for viruses of Bacteria and Archaea based on protein content. *Patterns* 2021;2:100274.
49. Ahlgren NA, Ren J, Lu YY, Fuhrman JA, Sun F. Alignment-free d_2^* oligonucleotide frequency dissimilarity measure improves prediction of hosts from metagenomically-derived viral sequences. *Nucleic Acids Res* 2017;45:39–53.
50. Lu C, Zhang Z, Cai Z, Zhu Z, Qiu Y, *et al*. Prokaryotic virus host predictor: a Gaussian model for host prediction of prokaryotic viruses in metagenomics. *BMC Biol* 2021;19:5.
51. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol* 1990;215:403–410.
52. Yuan Y, Gao M. Jumbo bacteriophages: an overview. *Front Microbiol* 2017;8:403.
53. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, *et al*. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* 2015;31:3691–3693.
54. Tonkin-Hill G, MacAlasdair N, Ruis C, Weimann A, Horesh G, *et al*. Producing polished prokaryotic pangenomes with the Panaroo pipeline. *Genome Biol* 2020;21:180.
55. Horsfield ST, Tonkin-Hill G, Croucher NJ, Lees JA. Accurate and fast graph-based pangenome annotation and clustering with ggCaller. *Genome Res* 2023;33:1622–1637.
56. Nishimura Y, Yoshida T, Kuronishi M, Uehara H, Ogata H, *et al*. ViPTree: the viral proteomic tree server. *Bioinformatics* 2017;33:2379–2380.
57. Moraru C, Varsani A, Kropinski AM. VIRIDIC-A novel tool to calculate the intergenomic similarities of prokaryote-infecting viruses. *Viruses* 2020;12:1268.
58. Moraru C. VirClust-a tool for hierarchical clustering, core protein detection and annotation of (*Prokaryotic*) viruses. *Viruses* 2023;15:1007.