A hybrid and poly-polish workflow for the complete and accurate assembly of phage genomes

Elek, Claire K.A.; Brown, Teagan L.; Viet, Thanh Le; Evans, Rhiannon; Baker, David J.; Telatin, Andrea; Tiwari, Sumeet K.; Al-Khanaq, Haider; Thilliez, Gaëtan; Kingsley, Robert A.; Hall, Lindsay J.; Webber, Mark A.; Adriaenssens, Evelien M.

DOI:
10.1099/mgen.0.001065

MICROBIAL GENOMICS

RESEARCH ARTICLE
Elek *et al.*, *Microbial Genomics* 2023;9:001065
DOI 10.1099/mgen.0.001065

A hybrid and poly-polish workflow for the complete and accurate assembly of phage genomes: a case study of ten przondoviruses

Claire K. A. Elek[1,2], Teagan L. Brown[1], Thanh Le Viet[1], Rhiannon Evans[1], David J. Baker[1], Andrea Telatin[1], Sumeet K. Tiwari[1], Haider Al-Khanaq[1], Gaëtan Thilliez[1], Robert A. Kingsley[1,2], Lindsay J. Hall[1,2,3], Mark A. Webber[1,2] and Evelien M. Adriaenssens[1,*]

Abstract

Bacteriophages (phages) within the genus *Przondovirus* are T7-like podoviruses belonging to the subfamily *Studiervirinae*, within the family *Autographiviridae,* and have a highly conserved genome organisation. The genomes of these phages range from 37 to 42 kb in size, encode 50–60 genes and are characterised by the presence of direct terminal repeats (DTRs) flanking the linear chromosome. These DTRs are often deleted during short-read-only and hybrid assemblies. Moreover, long-read-only assemblies are often littered with sequencing and/or assembly errors and require additional curation. Here, we present the isolation and characterisation of ten novel przondoviruses targeting *Klebsiella* spp. We describe HYPPA, a HYbrid and Poly-polish Phage Assembly workflow, which utilises long-read assemblies in combination with short-read sequencing to resolve phage DTRs and correcting errors, negating the need for laborious primer walking and Sanger sequencing validation. Our assembly workflow utilised Oxford Nanopore Technologies for long-read sequencing for its accessibility, making it the more relevant long-read sequencing technology at this time, and Illumina DNA Prep for short-read sequencing, representing the most commonly used technologies globally. Our data demonstrate the importance of careful curation of phage assemblies before publication, and prior to using them for comparative genomics.

DATA SUMMARY

Phage raw reads are available from the National Center for Biotechnology Information Sequence Read Archive (NCBI-SRA) under the BioProject number PRJNA914245. Phage annotated genomes have been deposited at GenBank under the accessions OQ579023–OQ579032 (Table 1). Bacterial WGS data for clinical preterm infant samples have been deposited at GenBank under BioProject accession PRJNA471164 (Table S1, available in the online version of this article). Bacterial raw reads for food samples are available from NCBI-SRA with individual accessions (SAMN33593347–SAMN33593351), and can be found under the BioProject number PRJNA941224 (Table S1). Strain-specific details for bacteria and publicly available phages used in these analyses, along with accessions for the latter, can be found in Tables S1 and S6, respectively. The CL1–CL8 clinical *Klebsiella* strains (Table S1) were under a Materials Transfer Agreement, for which sequencing data and strain information is not available.

INTRODUCTION

Double-stranded (ds) DNA bacteriophages with the characteristic head–tail morphology, also known as tailed phages, are a diverse group of viruses spanning 47 families, 98 subfamilies and 1197 genera, with many more being unclassified [1–4]. Phages within the

Received 30 March 2023; Accepted 17 June 2023; Published 18 July 2023
Author affiliations: [1]Quadram Institute Bioscience, Rosalind Franklin Road, Norwich Research Park, Norwich, UK; [2]University of East Anglia, Norwich Research Park, Norwich, UK; [3]Chair of Intestinal Microbiome, ZIEL—Institute for Food and Health, School of Life Sciences, Technical University of Munich, Freising, Germany.
***Correspondence:** Evelien M. Adriaenssens, evelien.adriaenssens@quadram.ac.uk
Keywords: assembly; bacteriophage; *Klebsiella*; phage; *Przondovirus*; sequencing.
Abbreviations: BHI, brain heart infusion; CPS, capsular polysaccharide; DNAP, DNA polymerase; DTR, direct terminal repeat; HYPPA, hybrid and poly-polish phage assembly; ICTV, International Committee on Taxonomy of Viruses; LPS, lipopolysaccharide; MLST, multilocus sequence typing; NCBI, National Center for Biotechnology Information; ONT, Oxford Nanopore Technologies; QC, quality control; QIB, Quadram Institute Bioscience; RNAP, RNA polymerase.
Data statement: All supporting data, code and protocols have been provided within the article or through supplementary data files. Seven supplementary tables and four supplementary figures are available with the online version of this article.

Impact Statement

The current workflows employed for phage genome assembly are often error-prone and can lead to many incomplete phage genomes being deposited within databases. This can create challenges when performing comparative genomics, and may also lead to incorrect taxonomic assignment. To overcome these challenges we proposed HYPPA, a workflow that can produce complete and high-quality phage genomes without the need for laborious lab-based validation.

genus *Przondovirus* are T7-like podoviruses, meaning they have a short tail morphotype, belonging to the subfamily *Studiervirinae*, within the family *Autographiviridae* [5]. T7-like phages are renowned for following a strictly lytic life cycle, with the eponymous *Escherichia coli* phage T7 often used as the type isolate to represent the family *Autographiviridae* [6, 7].

Autographiviridae phages typically have genomes ranging from 37 to 42 kb in size and encode 50–60 genes, with the DNA-directed RNA polymerase (RNAP) being a hallmark of the family [5, 6, 8]. The genome organisation of genera within the subfamily *Studiervirinae* is highly conserved: all but a few genes are unidirectional and show a high degree of synteny [2, 5–7].

Tailed phages employ a remarkably diverse array of packaging methods that generate distinct termini [9, 10]. The termini of T7-like phages consist of direct terminal repeats (DTRs) of varying lengths that flank the genome [6]. The DNA of T7-like phages is concatemeric when generated within the bacterial cell and requires the assistance of terminases to cut at specific sites to package the DNA into the procapsid [9–11]. Whilst each concatemer contains a single copy of the repeat, a second repeat is synthesised at the other end of the genome to prevent loss of genetic material [9, 12]. Additionally, the DTRs are thought to prevent host-associated digestion *in vivo* and assist in DNA replication during phage infection [10, 13].

Many phage genomes deposited within public sequence databases are incomplete, often with DTR sequences missing or simply not annotated. Thus, our relatively limited understanding of phage biology is exacerbated by incomplete data and can make classification and comparative genomics more challenging [14]. Indeed, high-quality genomic data will help identify relationships between taxonomic classification, infection kinetics and phage–host interactions that are essential to the use of phages as therapeutics [14].

The genus *Klebsiella* comprises a heterogeneous group of Gram-negative bacteria in the order Enterobacterales [15]. *Klebsiella* spp. are common commensals of human mucosae, presenting a major risk factor for developing invasive disease and are therefore important opportunistic pathogens [15, 16]. Antibiotic resistance among *Klebsiella* spp. represents a major threat to human health, with many isolates now being multidrug-resistant [15, 16]. Therefore, conventional treatment using currently available antibiotics is becoming increasingly ineffective, and combined with no new antibiotics in the drug development pipeline, we are entering a post-antibiotic era [17, 18]. Treatment of recalcitrant infections with bacterial viruses, bacteriophage therapy, has seen a resurgence in recent years as an alternative or adjunctive to current antibiotic therapy [19, 20].

Phage isolation involves monomicrobial or polymicrobial enrichment that often selects for the fittest phages [14, 21–23]. Indeed, the rapid infection cycle of T7-like phages means that they are often overrepresented following traditional isolation methods [14, 21, 23]. Here, ten novel T7-like phages belonging to the genus *Przondovirus* in the family *Autographiviridae* have been isolated against four *Klebsiella* strains belonging to different species, and characterised. Hybrid poly-polish assembly methods have recently been described for assembling bacterial genomes [24]. We developed and validated a similar approach to ensure accurate and complete phage genome assembly, in a new worklow HYPPA, a HYbrid and Poly-polish Phage Assembly, which was tested and validated for these new phages. The workflow utilises long-read assemblies in combination with short-read sequencing to resolve phage DTRs and correct sequencing and/or assembly errors, which negates the need for laborious primer walking and Sanger sequencing validation.

METHODS

Bacterial strains and growth conditions

Where specified, *Klebsiella* spp. used here were derived from previous studies [25–29] and are listed in Table S1. All *Klebsiella* strains were cultured overnight on brain heart infusion (BHI) agar (Oxoid) at 37 °C. Liquid cultures were prepared by inoculation of 10 ml BHI broth with each bacterial strain and incubated at 37 °C with shaking at 200 r.p.m. for 3 h. Single colony variants were identified on solid media by changes in colony morphology and were purified by selecting a single colony for three successive rounds of purification on MacConkey no. 3 agar (Oxoid), and incubated overnight at 37 °C.

Preparation of bacterial DNA and sequencing

Genomic DNA for each *Klebsiella* strain was extracted using the AllPrep Bacterial DNA/RNA/Protein kit (Qiagen) according to the manufacturer's instructions. DNA was quantified by a Qubit 3.0 fluorometer using the broad range dsDNA kit (Invitrogen) and normalised to 5 ng μl^{-1}.

DNA was prepared using the Illumina DNA Prep library preparation kit and was whole-genome sequenced on the Illumina NextSeq500 platform generating 2×150 bp paired-end reads by QIB Sequencing Core Services.

Additionally, *Klebsiella michiganensis* M7 21 2 #35, *K. pneumoniae* M26 18 1, *K. pneumoniae* M26 18 2 #21 KpnN, *K. pneumoniae* M26 18 2 #21 KpnA and *K. pneumoniae* ST38 01 were prepared for enhanced sequencing by MicrobesNG whole genome sequencing services (www.microbesng.com), which is supported by the Biotechnology and Biological Sciences Research Council (BBSRC). Bacterial samples were prepared according to the sequence facilities instructions and genomes were received from MicrobesNG assembled and annotated.

Bacterial genomics

Short-read data provided without pre-processing by QIB Sequencing Core Services was QC filtered, trimmed, assembled, annotated and analysed using the ASA^{3}P v1.2.2 [30] or Bactopia v1.6.4 [31] pipelines. Preliminary strain designations were determined by ribosomal multilocus sequence typing (rMLST) (https://pubmlst.org/species-id) [32]. The PubMLST database (https://pubmlst.org/) [33] was used to determine sequence types (ST) for the *K. oxytoca* species complex and *K. aerogenes*, while the Institute Pasteur MLST database (https://bigsdb.pasteur.fr/) was used to determine STs of the *K. pneumoniae* species complex. The capsular type for each strain was predicted using Kleborate [34] and Kaptive [35] on the QIB Galaxy platform, and those with a match confidence of good or higher were included.

Isolation and single-plaque purification of phages

Samples from various UK wastewater treatment plants were screened for *Klebsiella*-specific phages using a range of *Klebsiella* strains as hosts for enrichment, adapted from Van Twest *et al.* [36]. Briefly, 300 µl filtered wastewater was mixed with 60 µl exponential growth bacterial culture and used to inoculate 5 ml BHI broth. Enrichments were incubated overnight at 37 °C with shaking at 200 r.p.m. Enrichments were then centrifuged (4000 ***g*** for 15 min) and passed through a 0.45 µm filter before spot testing by a double agar overlay plaque assay, as previously described [37]. All incubations for the overlay method were performed over 4–17 h at 37 °C. Single plaque purifications were made by extracting single plaques from the soft agar layer using sterile toothpicks and suspended in approximately 300 µl BHI broth. Suspensions were centrifuged (13 000 ***g*** for 5 min) and supernatant collected. Ten-fold serial dilutions of the supernatant were performed in phage buffer (75 mM NaCl; 10 mM $MgSO_4$; 10 mM Tris, pH 7.5; 0.1 mM $CaCl_2$) and 10 µl of each dilution was plated onto double agar overlay and incubated as described above. This process was repeated at least three times to create phage stocks.

Phage amplification was performed as for single plaque purification in BHI broth. Once the supernatant was collected, approximately 100 µl of phage suspension was spread on to three double agar overlay plates and incubated as before. Phage stocks were prepared by extraction of phage clearance zones. This was achieved by removal of the soft agar layer, which was resuspended in phage buffer, and centrifuged (4000 ***g*** for 15 min). Phage supernatant was passed through a 0.45 µm filter into a sterile glass vial and stored at 4 °C.

Phage host range

Phage host range was tested by a plaque assay as described above on a range of clinical, wastewater, food and type strain *Klebsiella* spp. as described previously [38]. Only assays where individual plaques were identified were recorded as positive.

Phage DNA extraction and whole-genome sequencing

Phage virions were concentrated by PEG 8000 (Thermo Fisher) precipitation for DNA extraction. Briefly, phage stock was treated with 1 µl DNase I (10 U µl^{-1}) (Merck) and 1 µl RNase A (10 U µl^{-1}) (Merck) per millilitre of stock and incubated at 37 °C for 30 min. PEG precipitation was performed with PEG 8000 (10%, w/v) and 1 M NaCl and incubated overnight at 4 °C. The precipitate was centrifuged (17 000 ***g*** for 10 min) and resuspended in 200 µl nuclease-free water. Resuspended phage pellets were treated with proteinase K (50 µg ml^{-1}) (Merck), EDTA (final concentration 20 mM) and 10% SDS (final concentration 0.5%, v/v) and incubated at 55 °C for 1 h.

DNA was extracted using the Maxwell RSC Viral Total Nucleic Acid Purification kit (Promega), as per the manufacturer's instructions, into nuclease-free water. Phage DNA was quantified by a Qubit 3.0 fluorometer using the high-sensitivity dsDNA kit (Invitrogen). DNA was prepared using an Illumina DNA Prep (formerly Nextera Flex) library preparation kit and was whole-genome sequenced on the Illumina NextSeq500 platform generating 2×150 bp paired-end reads by QIB Sequencing Core Services. MinION libraries (Oxford Nanopore Technologies, ONT) were constructed without shearing using the short fragment buffer and loaded onto the R9.4.1 flow cell according to the manufacturer's instructions by QIB Sequencing Core Services.

Both long-read and short-read raw data for all ten przondoviruses were deposited in NCBI under BioProject number PRJNA914245.

Phage genomics

Assembly and annotation

All quality control, pre-processing, assembly and annotation of phage genomes were performed on the QIB Galaxy platform.

We checked short-read data for quality using fastQC v0.11.8 [39]. Based on this fastQC analysis, reads were pre-processed with fastp v0.19.5 [40], using a hard trim of between 4 and 10 bases on both the front and tail to retain at least a per-base quality of 28.

Long-read data were demultiplexed following sequencing and quality checked with NanoStat v0.1.0 [41]. Pre-processing was performed as part of the assembly, and assembled using Flye v2.9 [42] with default settings, which included correction and trimming of reads. Flye was used in the first instance as previously published work has determined it is the most accurate and reliable assembler [43–45]. Where Flye was unable to generate a high-quality assembly, Canu v2.2 [46] was used as an alternative. Error correction and trimming were performed as part of the default settings when assembling using Flye or Canu. Flye additionally performed one iteration of long-read polishing by default. We assembled all phages with and without trimming adapter/barcode sequences for long reads. Trimming was performed with Porechop v0.2.3 (https://github.com/rrwick/Porechop) [47] with default settings.

We performed several iterations of long-read and short-read polishing on long-read-only assemblies in a specific order. First, two iterations of long-read polishing were performed using Medaka [48] with default settings, using the previous polished data as the input for the next round of polishing. Second, one iteration of short-read polishing was performed using Polypolish [49] with default settings. Finally, a second iteration of short-read polishing was performed using POLCA [50] with default settings. We used raw reads for each iteration of long-read polishing and pre-processed reads for each iteration of short-read polishing.

Prior to development of the current phage assembly workflow, we had adopted a few other methodologies for resolving the genomes. One method was short-read-only assembly, where phages were assembled *de novo* using Shovill v1.0.4 (https://github.com/tseemann/shovill) with default settings [51, 52]. Briefly, trimming was disabled by default and manual trimming was performed as part of the pre-processing step prior to assembly. Additionally, SPAdes was used as the default assembler within the Shovill pipeline. We attempted short-read polishing of long-read-only data using Pilon v1.20.1 [53] with default settings. Where specified, we also performed hybrid assembly using raw long-read and pre-processed short-read data, as previously described using Unicycler v0.4.8.0 [54] with default settings. Porechop v0.2.3 (https://github.com/rrwick/Porechop) [47] was used for *Klebsiella* phage Oda only. All assembly details are given in Tables S3–S5.

Following assembly, the contigs were manually checked for DTRs flanking the genome, as well as with PhageTerm [55] which was unable to identify the DTRs since it does not work well for Nextera-based sequence libraries. Where we could not determine the length and sequence of the DTRs, we performed primer walking. Outward-facing primers were designed to 'walk' the genome termini using Sanger sequencing [56]. Phage DNA was extracted, and for each phage at least two primers were designed for the reverse strand to walk the beginning of the genome and identify the left terminal repeat, and at least two primers were designed for the forward strand to walk the end of the genome to identify the right terminal repeat. The phage DNA and each primer were then sent for Sanger sequencing separately (Eurofins). Sanger sequences were visualised in FinchTV v1.5.0 (https://digitalworldbiology.com/FinchTV) and compared to the reference phage genome, and DTRs were annotated using the Molecular Biology suite on the Benchling platform (https://www.benchling.com/).

Assemblies generating multiple contigs were checked for contamination using Kraken 2 v2.1.1 [57].

Verification of the DTRs and assessment of assembly quality was performed by mapping the raw reads back to the assembled genome using Bowtie 2 v2.3.4.3 [58] and visualised using IGV v2.7.2 [59], and variant calling was performed using iVar v1.0.1 [60]. Additionally, BWA-MEM v0.7.17.1 (https://github.com/lh3/bwa) was used to map long reads back to the reference using default settings optimised for ONT reads [61, 62].

Assemblies in the reverse orientation were reorientated by reverse complementation of the genome in UGENE v38.0 [63] and uploaded to Benchling. Contigs were then reoriented to begin at the same start point, based on well-curated reference phages and analysis of the DTRs.

Genome annotation was performed using Pharokka v1.2.1 with default settings (https://github.com/gbouras13/pharokka) [64]. Specifically, coding sequences were predicted with PHANOTATE [65].

Comparative genomics

Where specified, publicly available phage genomes used for comparative genomics were derived from these studies [20, 66–78], listed in Table S6, and downloaded from the GenBank database.

The closest relative for each phage was determined as as the top hit according to the maximum score identified by nucleotide BLAST (BLASTn) (https://blast.ncbi.nlm.nih.gov/Blast.cgi) and optimised for somewhat similar sequences [79]. Genes associated with specific phage families, such as the DNA-directed RNAP for *Autographiviridae*, were identified and used for

Table 1. Przondoviruses within the collection to date and data relating to the closest database relative

Phage name	Source	Isolation host	Genome size (bp)	GC content (%)	DTR size (bp)	No. of CDS	Accession	Closest database relative according to BLASTn		
								Name	Coverage (%)	ID (%)
Oda	River water	K.mi.	41642	52.64	181	58	OQ579023	*Klebsiella* phage SH-KP152226	92.0	94.62
Toyotomi	Wastewater	K.mi.	41268	52.64	180	55	OQ579024	*Klebsiella* phage SH-KP152226	92.0	94.71
Mera	Wastewater	K.mi.	41400	52.58	180	56	OQ579025	*Klebsiella* phage SH-KP152226	92.0	94.34
Speegle	Wastewater	K.mi.	41395	52.64	180	58	OQ579026	*Klebsiella* phage SH-KP152226	93.0	94.70
Cornelius	Wastewater	K.mi.	40437	52.72	180	55	OQ579027	*Klebsiella* phage SH-KP152226	94.0	94.84
Tokugawa	Wastewater	K.mi.	41414	52.64	181	56	OQ579028	*Klebsiella* phage SH-KP152226	92.0	94.70
Saitama	Wastewater	K.qp.	40741	53.06	181	51	OQ579029	*Klebsiella* phage K11	96.0	95.71
Emom	Wastewater	K.ox.	40788	52.56	183	53	OQ579030	*Klebsiella* phage KP32	94.0	93.14
Amrap	Wastewater	K.ox.	41209	52.47	182	57	OQ579031	*Klebsiella* phage KPN3	85.0	95.05
Whistle	Wastewater	K.va.	40735	52.40	181	54	OQ579032	*Klebsiella* phage IME264	94.0	94.78

K.mi., *K. michiganensis* M7 21 2 #21; K.ox., *K. oxytoca* M59 22 8; K.qp., *K. quasipneumoniae* P057K W; K.va., *K. variicola* DSM 15968. CDS, coding sequences. Bacterial host species accessions are given in Table S2.

preliminary taxonomic assignment [5, 6, 8]. Alignments were performed using Mauve v20150226 [80] between the closest relative and phages from the same genera. The intergenomic similarity between przondoviruses in the collection and a selection of publicly available related phages was calculated using VIRIDIC on the web server (http://rhea.icbm.uni-oldenburg.de/VIRIDIC/) [81].

Phylogenetic analyses were performed using the hallmark DNA-directed RNAP amino acid sequence for all phages and a selection of publicly available phylogenetically related phages downloaded from the NCBI protein database (https://www.ncbi.nlm.nih.gov/). Multisequence alignment of the RNAP amino acid sequences was performed using the MUSCLE algorithm in MEGA X v10.0.5 [82] with default settings. A maximum-likelihood tree was generated with 500 boostraps using the default Jones–Taylor–Thornton model. Phylogenetic analysis was performed using 35 amino acid sequences, with a total of 684 positions in the final analysis. Tree image rendering was performed using iTOL v6.1.1 (https://itol.embl.de/) [83].

Linear mapping of coding sequences for phage final assemblies was performed using Clinker v0.0.23 [84].

RESULTS AND DISCUSSION

Phage isolation and host range determination

In this study, we isolated ten lytic T7-like phages from a variety of river water and wastewater samples, using four different *Klebsiella* spp. as isolation hosts (Table 1). To examine the host range, we tested the ten phages against a collection of *Klebsiella* spp. from different sources, representing a range of capsule and sequence types. All phages had a narrow host range, with eight being able to infect only a single *Klebsiella* strain within our collection (Fig. 1).

Three of the ten przondoviruses were used to test and validate the HYPPA workflow: Oda, Toyotomi and Tokugawa. As the three unifiers of the HYPPA workflow, these were named after the three unifiers of Japan (see Development of a new workflow for the assembly of complete phage genomes).

Only two of our phages were capable of productively infecting more than one *Klebsiella* strain: *Klebsiella* phages Emom and Amrap were both able to infect two different isolates of *K. oxytoca*. *Klebsiella* phage Whistle was the only phage that demonstrated lysis without productive infection on a further three *K. pneumoniae* isolates in addition to the isolation host. We could not establish a link between capsular type and host range for these phages.

Przondoviruses and other T7-like phages have a relatively small genome of 37–42 kb, and this may limit their host expansion capabilities (for taxonomic assignment of the ten phages in this study, see section Phage genome characterisation and taxonomy). However, Emom and Amrap were capable of infecting two hosts. Previous work has shown that T7-like phages are capable of infecting multiple hosts [66] and that host range is determined by interaction between phage receptor binding proteins, i.e. tail fibre and/or spike proteins, and bacterial cell receptors [14, 66, 85]. Lipopolysaccharide (LPS) components are almost always identified as the secondary receptor for irreversible attachment in Gram-negative-targeting podoviruses

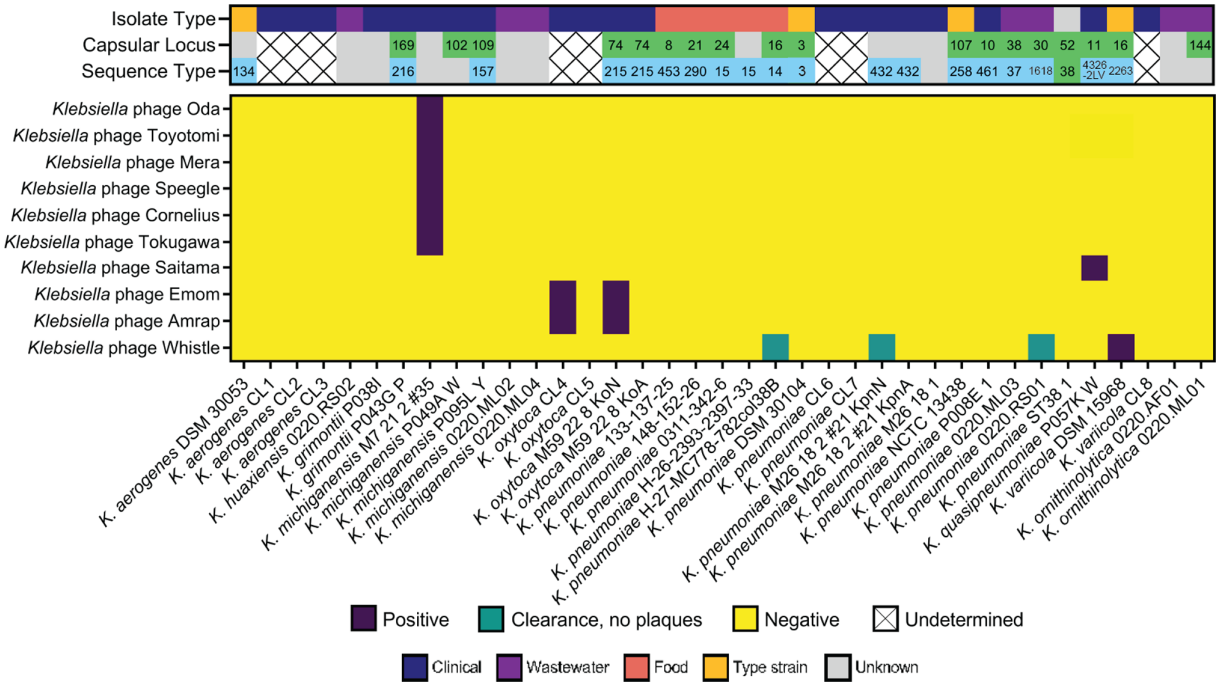

Fig. 1. Heatmap for host range of the przondoviruses in the collection by plaque assay against a diverse range of *Klebsiella* spp. Top panel: isolate type, capsular type and sequence type. The source of each isolate is given as isolate type, with grey indicating an unknown source. Capsular loci determined by Kaptive and/or Kleborate, green; unknown or no match confidence, grey. Sequence type (ST) was determined by MLST, blue; unknown or incomplete matches, grey. No sequencing data available, undetermined. Bottom panel: host range heatmap. Productive infection (positive) is the observation of individual plaques, purple; lysis without productive infection is the observation of clearance without individual plaques, green; no productive infection or clearance (negative), yellow.

[6, 14]. Whether initial interaction with the outer membrane and degradation of the capsular polysaccharide (CPS) constitutes a *bone fide* reversible attachment step, or whether this is a prerequisite to reversible attachment by the phage to another outer membrane component has yet to be fully elucidated [6, 14, 86, 87].

Some phages can be 'trained' to increase their host range through co-evolution assays [19, 88]. This may be particularly useful in cases of lysis from without, such as observed in Whistle, as they are already capable of binding to host receptors but unable to cause productive infection.

Multiple factors affect host range and broadly involve extracellular and intracellular mechanisms. Extracellular mechanisms involve the ability of phages to bind to specific phage receptors on the bacterial cell surface that facilitate DNA ejection [89]. Intracellular mechanisms involve evasion of phage defence systems that facilitate phage propagation [89]. Expression of diffusible depolymerases facilitates interaction of phages with their primary and secondary receptor. This extracellular mechanism is more likely to explain the ability of Emom and Amrap to infect more than one isolate since there is productive infection. Thus, the ability of two przondoviruses in our collection to infect different *Klebsiella* isolates could indicate that they share similarities in the chemical composition of their capsules, enabling degradation by a single depolymerase and allowing access to the phage receptors on the bacterial cell. Moreover, the bacterial isolates could share similar sugar motifs within their LPS structures, which are thought to be the secondary receptor of phages within the family *Autographiviridae* [6]. Without full sequencing data for the *K. oxytoca* CL4 isolate, it is difficult to speculate further.

Development of a new workflow for the assembly of complete phage genomes

To generate complete and accurate genomes for these ten phages, which included resolving the defined ends of phage genomes, and correcting sequencing and/or assembly errors, we utilised a long-read-only assembly with sequential polishing steps. This methodology exploited both long-read and short-read sequencing data in a workflow that we have named HYPPA – HYbrid and Poly-polish Phage Assembly (see also Materials and Methods) before moving onto annotation and comparative genomics (Fig. S1). First, the long reads were assembled using Flye or Canu, followed by two iterations of long-read polishing with Medaka. Next, we performed two iterations of short-read polishing using Polypolish (for the first iteration) and POLCA (for the second iteration).

Initially, Flye was used as the primary assembler in our HYPPA workflow and worked particularly well for phages with both very high sequence read coverage (Toyotomi at >117 000×) and very low sequence read coverage, which included Mera (8×), Speegle (23×) and Amrap (27×) (Table S2). However, Canu performed better with the other phages as the assemblies in general contained fewer errors in their repeat regions. Other types of errors included SNPs, particularly in homopolymer regions, or short insertions and/or deletions (indels), which were especially noticeable in coding regions (Table S3). This is contrary to previously published literature that found Flye was the more accurate assembler using default settings [43–45].

As an illustration of the HYPPA workflow, we provided a more detailed description of the process for phage Oda as an exemplar, for which the DTRs were validated with primer walking. First, Oda was assembled using Canu, which yielded one contig of 41 761 bp. After two iterations of long-read polishing followed by two iterations of short-read polishing, the resulting contig was 41 769 bp in size. We were able to identify the terminal repeat regions, but both were flanked by a 64 bp sequence upstream of the left terminal repeat, and downstream of the right terminal repeat after all polishing iterations were complete. The two 64 bp sequences were inverted repeats containing adapter sequences of 23 bp, with the remaining sequence being Nanopore barcodes which were manually removed. HYPPA was then used for phage Tokugawa, which after short-read-only assembly had included a 79 bp repeat within the genome, but outside of the presumed DTR region (Fig. S2). Using HYPPA, the repeat was determined to be an assembly artefact and removed from the assembly. The final curated assembly for phages Oda and Tokugawa was 41 642 and 41 414 bp, respectively. Terminal repeats were present for both phages and complete at 181 bp, validated by primer walking and Sanger sequencing (Fig. S2).

We trimmed the long reads using Porechop in an attempt to remove the adapter/barcode sequences, but when phage Oda was reassembled and polished using the trimmed reads, the right terminal repeat was missing three bases, but no other SNPs or indels were identified.

The HYPPA workflow without Porechop-mediated trimming was repeated for the remaining eight przondoviruses, resulting in final genome assemblies ranging between 40 and 42 kb (Table 1). HYPPA was able to generate a complete genome for phage Toyotomi, where short-read-only, long-read-only, and hybrid assemblies were unable to do so and resulted in fragmented assemblies. Although our HYPPA workflow is a hybrid assembly approach, there is a clear distinction between this and traditional hybrid assembly methods. Importantly, HYPPA used the short reads for polishing only, not during the genome assembly, whereas traditional hybrid assemblies utilise both long-read and short-read data during the assembly process itself. Moreover, short-read polishing of a long-read-only assembly using Pilon was also unable to resolve the genome of Toyotomi: partial repeat regions were found at the termini but were incomplete, and multiple errors within coding regions persisted. Using HYPPA, we were able to not only preserve the DTRs of Toyotomi, but also correct persistent sequencing and/or assembly errors that occurred in all non-HYPPA assemblies.The genome organisation of genera within the family *Autographiviridae* is highly conserved: all genes are unidirectional and show a high degree of synteny, and genomes are flanked by DTRs [2, 5–7]. The DTRs of the przondoviruses described here were 180–183 bp in size, demonstrating sequence similarity of 84.3–99.7%. DTRs are thought to assist circularisation of the phage genome once in the host cytoplasm to prevent host-induced enzymatic digestion [13]. Thus, resolution of the DTRs is integral to accurate genomics and understanding of the biology of different phages.

Comparison of HYPPA with traditional short-read-only assembly

When compared to typical short-read-only methodologies of phage genome assembly, in our case using Shovill [51], the HYPPA workflow required significantly less manual curation (Fig. S1). Typically, phage genomes are assembled using short-read only data, and many of these genomes are then published without additional curation, leaving them with potentially significant sequencing and/or assembly errors. Using short-read-only assembly methods for our collection of przondoviruses, we observed that some were in the reverse orientation rather than the forward orientation as is expected for 50% of the assemblies, and some had the DTRs assembled in the middle of the contig. Addressing these issues required manually re-orienting the assemblies and ensuring they all had the same start position, as suggested in the Phage Annotation Guide [90]. In contrast, the HYPPA workflow resulted in assemblies with correct start and stop sites, but some were still in the reverse orientation.

To check for DTRs in short-read-only assemblies, we initially looked for increased reads within the read mapping profiles, which are distinguished by one or two large peaks, and can be automated using the tool PhageTerm [55]. If a single peak was observed anywhere other than at either end of the assembly, the assembly had been opened in the middle of the genome and required each to be re-oriented to have the same starting position.

Incorrect orientation is a feature of phage genome assembly, and with short-read-only data in particular, may be artificially linearised by the assembler with the DTRs located in the middle of the contig. In many of our own short-read-only assemblies, the przondoviruses described here were linearised in the middle of the genome, and required read mapping to identify where the DTRs may be. In T7-like phages, DNA is concatemeric and requires the assistance of terminases to cut at specific sites to package the DNA into the procapsid [9–11]. Although each concatemer contains a single copy of the repeat, a second repeat is synthesised at the other end of the genome to prevent loss of genetic material [9, 12]. Since the DTRs are present twice per phage genome, the number of terminal sequences is double following whole genome sequencing and are identified as a single peak

of increased reads during read mapping [10–12, 55]. Therefore, the DTR and, by proxy, the start of the genome can be inferred from the read mapping. Moreover, due to the highly conserved nature of the genomes, all przondoviruses had almost the same starting sequence as the well-curated enterobacterial phage K30 (accession HM480846) [67], making the beginning relatively easy to find. As a result, considerable time was spent on re-orienting the short-read-only assemblies to be unidirectional and to have the same starting sequence.

One of the most problematic aspects using short reads for phage assembly (both short-read-only and as part of a traditional hybrid assembly) was that the DTRs were deleted, possibly because the assemblers used deem them to be a sequencing artefact. Thus, DTRs need to be manually validated through primer walking and Sanger sequencing validation. However, this was unnecessary when using short reads for polishing rather than for assembly. Thus, using the HYPPA workflow, the DTRs were present in the final polished assembly in the correct location at the ends and did not have to be manually added.

A second type of error that routinely occurred during non-HYPPA phage sequencing and assembly was the introduction of short indels that were particularly noticeable in coding regions.

For the short-read-only assemblies, many sequencing and assembly errors present in coding regions were only found upon annotation of the genomes, including frameshift errors in DNA polymerase (DNAP) and tail fibre protein genes. Often, these frameshift errors were found in homopolymer regions and were introduced during sequencing. Before using HYPPA, these frameshift errors were checked through read mapping followed by variant calling and were edited accordingly. Particularly noteworthy were repeat regions of ~79 bp identified close to and sometimes within the DTR regions of seven of the ten phages (see Development of a new workflow for the assembly of complete phage genomes for a description of repeats for Tokugawa), but that did not correlate with the increased reads observed in the read mapping. This suggested that these repeats were introduced in error during assembly and were confirmed to be artefacts in most phages, including Tokugawa, through Sanger sequencing (see Fig. S2). Using HYPPA, we found that the two iterations of short-read polishing were able to correct SNPs and/or correct indels that resulted in these frameshift errors that long-read polishing was unable to resolve, particularly in homopolymer regions. POLCA was also able to correct indels that Polypolish was unable to resolve.

As previously described for Oda, all the przondoviruses contained adapter and barcode DNA upstream and/or downstream of the DTR regions. Initially, as we were trying to reconstruct the linear genome ends, we did not perform adapter and barcode trimming of the Nanopore reads prior to long-read assembly. We then removed these sequences manually after assembly. To limit the amount of manual curation, Porechop can be used to trim the reads, but when we attempted this for all the remaining przondoviruses, Porechop-mediated trimming resulted in several further errors. These included trimming bases from the beginning of the left terminal repeat and the end of the right terminal repeat, ranging from 3 to 18 bp in total; indels; multiple SNPs; and in some cases failure to assemble the phage genome into a single contig, or at all. We would thus recommend manual removal of the adapter/barcodes rather than trimming of long reads using Porechop, which appears to require more manual curation when compared to using raw Nanopore reads.

Multiple sequencing and/or assembly errors were identified in the coding regions of other phages that again persisted following traditional methods of phage assembly. Using trial and error, we were able to show that the HYPPA method was superior to other methods of phage assembly, whether hybrid or through using a single sequencing platform, in correcting errors (see Tables S2–S5 for all assembly details and errors). Moreover, the HYPPA workflow required far fewer manual curation steps than traditional phage assembly methods: while long-read-only assemblies were sometimes in the reverse orientation, all were linearised at the starting sequence. This is in contrast to the traditional assembly methods that required re-orienting the genomes to be unidirectional and starting at the same position, manual correction of large assembly errors such as indels, manual correction of homopolymer errors in coding regions, and in some cases rearrangement of contigs and manual stitching the genome together, followed by primer walking and Sanger sequencing validation to determine the genome termini and DTRs.

Errors in homopolymer sequences and repeat regions are particularly common in long-read-only assemblies of bacterial genomes [43, 44, 49], and as we have described here, in phage genomes also. Indeed, two homopolymer errors occurred in the DNAP of Toyotomi, leading to a double frameshift error that resulted in three protein annotations. Short-read polishing can correct errors introduced during long-read-only assemblies [49], as we have demonstrated here. Similarly to using short-read data for assembly, we found that a traditional hybrid assembly using both short- and long-read data for Toyotomi also introduced large deletions in repeat regions, with assembly errors persisting, as has been described previously [44, 54]. Assembly metadata showing all previous long-read-only, short-read-only, and hybrid assemblies alongside errors are provided (see Tables S2–S6).

Several limitations of this study include the need for both short- and long-read data for phage assembly, and specialised knowledge to access and install the software which is all freely available. Which polishing program used and what type of polishing (long-read versus short-read) in what order may give different results of equal validity. While we believe that the HYPPA workflow provides the most accurate phage genome possible, it still may not exactly reflect the DNA that is present within each phage capsid. Additionally, while the highly conserved nature of T7-like phages made it easier to determine the DTR starting sequence, this may not be the case for novel phages.

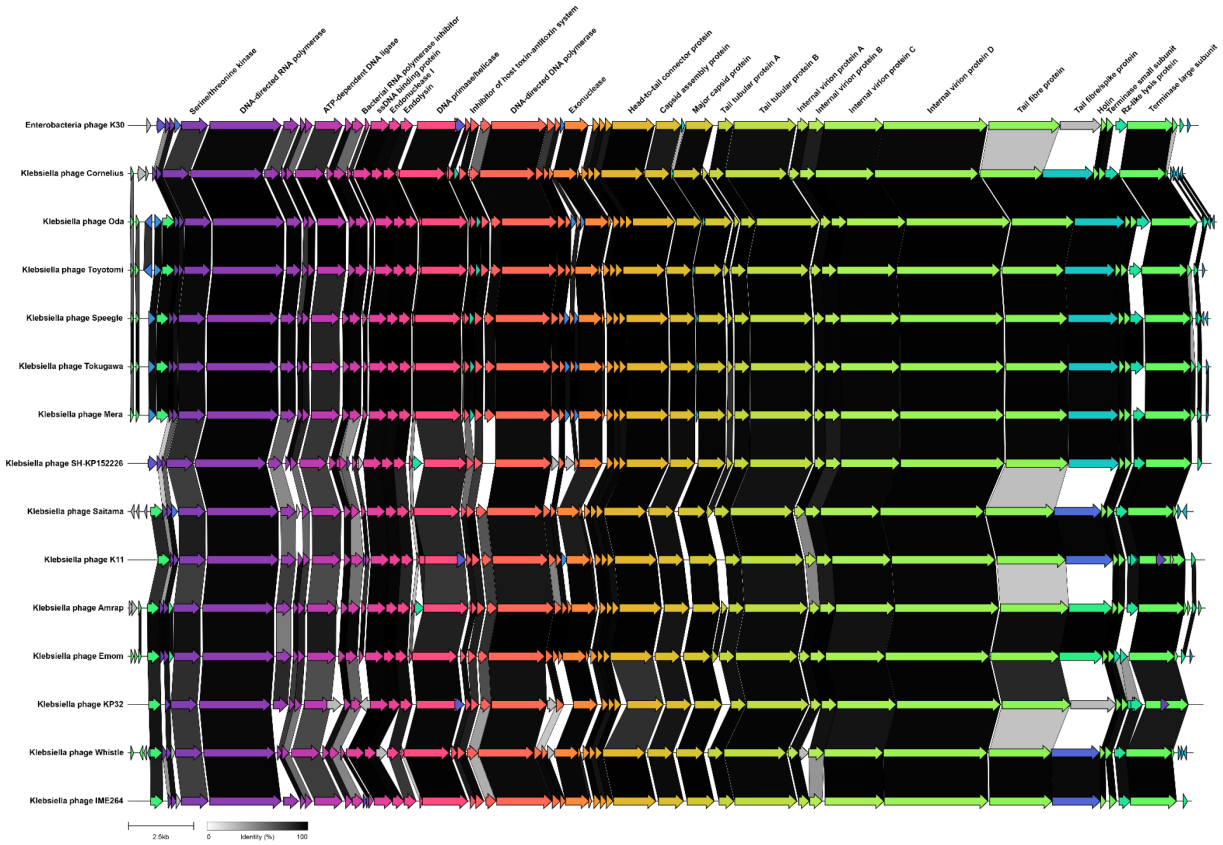

Fig. 2. Genome map and gene clustering for przondoviruses in the collection and a selection of related phages. Arrows represent coding sequences and pairwise comparisons of gene similarities are indicated by percentage identity given as links in greyscale, with darker shading representing areas of higher similarity. Genes without any sequence similarity are indicated without links. Some phages had a hypothetical protein following the tail fibre protein and protein BLAST revealed high homology to tail spike proteins. DTRs are present but not annotated.

Phage genome characterisation and taxonomy

All ten phages were dsDNA phages at 40 336–41 720 bp with a GC content of 52.40–53.06%, which is slightly lower than their isolation host GC content of ~55.46–57.59% (Table 1, Fig. 2). The number of predicted coding sequences within the genomes varied from 51 to 58, and almost all coding sequences were found in the same orientation on the forward strand. However, five phages had one to four small hypothetical proteins found in opposite orientation.

We performed BLASTn on all phages to determine their closest relatives in the NCBI GenBank database (as of October 2022). Based on the BLASTn results, which showed high levels of nucleotide similarity with reference phages, the phages in our collection were preliminarily assigned to the genus *Przondovirus* within the subfamily *Studiervirinae* and family *Autographiviridae*, according to the currently established ICTV genus demarcation criterion of 70% nucleotide sequence similarity over the genome length to belong to the same genus [1].

The genomic relationships between our novel przondoviruses and a selection of *Autographiviridae* reference phages were explored further by conducting a nucleotide-based intergenomic similarity analysis using VIRIDIC (Fig. 3, Table S2). Included within the analysis were relatives within the same genus (*Przondovirus*), those within different genera but the same subfamily (*Studiervirinae*) and those within different subfamilies (*Molineuxvirinae*, *Slopekvirinae*) (Fig. 3). These data confirmed that the przondoviruses from this study were within the ICTV genus demarcation criterion of 70% nucleotide sequence similarity over the genome length when compared to other przondoviruses. Several genera within the subfamily *Studiervirinae* that were included shared only ~45–57% nucleotide sequence similarity with the przondoviruses in this study (Fig. 3) .

Several przondoviruses clustered more closely together, including *Klebsiella* phages Oda, Toyotomi, Mera, Speegle, Cornelius and Tokugawa, which were within ~98% nucleotide similarity, except Cornelius which was the most dissimilar at ~95–96% (Fig. 3). All aforementioned phages except Oda were isolated from the same wastewater treatment plant at different stages of the treatment process, using the same host. These phages are therefore likely to be different strains of the same new species of phage within the genus *Przondovirus*. Emom and Amrap clustered with their closest relative KP32, but also clustered together

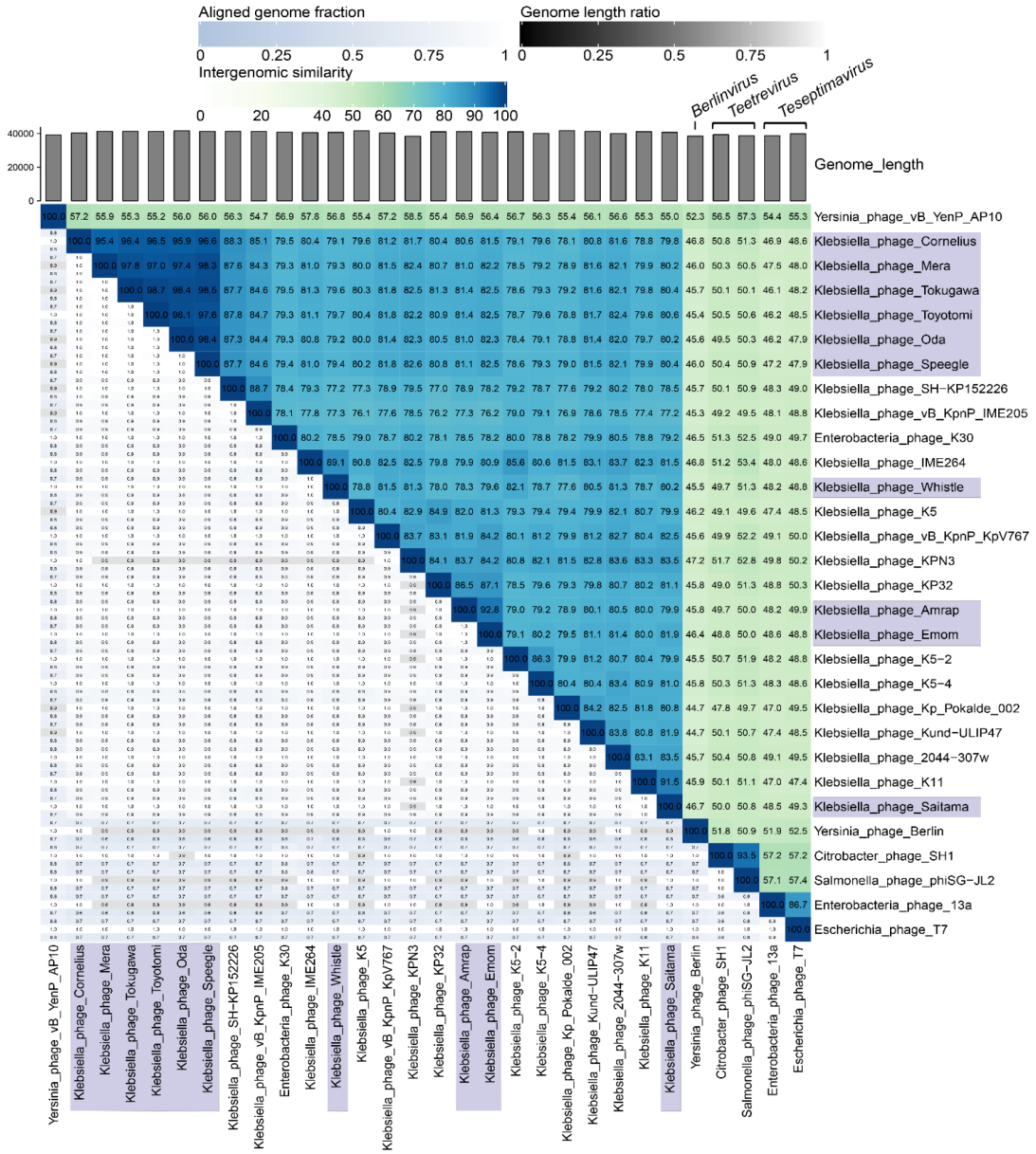

Fig. 3. Nucleotide-based intergenomic similarities of przondoviruses in the collection and a selection of related phages within the subfamily *Studiervirinae*, using VIRIDIC. A heatmap of hierarchical clustering of the intergenomic similarity values was generated and given as percentages (right half, blue–green heatmap). Each genome pair is represented by three values (left half), where the top and bottom (blue scale) represent the aligned genome fraction for the genome in the row and column, respectively, where darker colour indicates that a lower fraction of the genome was aligned. The middle value (grey scale) represents the genome length ratio for each genome pair, where darker colour indicates increasing distance between phages. The przondoviruses within our collection are highlighted in blue–grey. *Yersinia* phage vB_YenP_AP10 is in the genus *Apdecimavirus*.

with ~92% similarity, and should be assigned to separate species (Fig. 3). Saitama and Whistle did not cluster closely with any other phage from our collection, possibly due to differences in their host specificity. Saitama did cluster with its closest relative *Klebsiella* phage K11, and Whistle clustered with its closest relative IME264 (Fig. 3). This suggests that Saitama, Emom, Amrap and Whistle should be assigned to different species within the same genus.

After comparative genomic analyses, we observed that several of the closest database relatives were deposited in databases with incomplete genomes. Specifically, the incompleteness was most often due to an absence of the DTRs, including *Klebsiella* phages KP32, KPN3 and IME264 (Table S6, Fig. S3). Incomplete genomes could lead to incorrect assignments to species in cases where the reciprocal nucleotide identities are close to the species threshold of 95% similarity across the genome length [1].

Additionally, potential errors were noted in phages KPN3 (accession MN101227) and KMI1 (accession MN052874) (Table S6, Fig. S3). For example, KPN3 contained no annotated DNA-directed DNAP, which is conserved across all *Przondovirus* genomes analysed here. KMI1 contained a shorter DNA-directed RNAP annotation that, when included in the phylogenetic analyses, showed higher divergence, which could not be confirmed, and was therefore excluded from our phylogenetic analysis. Without raw short-read and long-read data, it is difficult to determine whether these are genuine errors or whether their differences are a true representation of the genome.

To further verify the taxonomic classification of the phages, phylogenetic analysis was performed using the protein sequence of the DNA-directed RNAP, since it is the hallmark gene of the family *Autographiviridae*, using a selection of publicly available phages from the genera *Apdecimavirus*, *Berlinvirus*, *Przondovirus*, *Teetrevirus* and *Teseptimavirus*, within the subfamily *Studiervirinae* (Fig. S4). As expected, the przondoviruses clustered together, and there was a clear separation from other phage genera.

Genome organisation and synteny

We conducted comparative genomic analysis of przondoviruses according to coding sequence similarity with a selection of reference phages (Fig. 2) . We selected enterobacterial phage K30 as the representative isolate of the genus *Przondovirus* since its genome is well curated. Przondoviruses were grouped together with their closest relative according to BLASTn. As expected, all phages share a highly conserved genome organisation, which revealed a high degree of gene synteny, in concordance with the VIRIDIC data (Fig. 3).

All genomes were found to contain the early, middle and late genes associated with viral host takeover, DNA replication, and virion assembly and lysis, respectively (Fig. 2). The host takeover proteins that were annotated included the *S*-adenosyl-L-methionine hydrolase, which is a good marker for the start of the genome; serine/threonine kinase; and DNA-directed RNAP, with the last being a hallmark of the family *Autographiviridae* [5, 8]. The middle proteins annotated were typical for phage DNA replication. The late proteins included all the components necessary for virion assembly, such as capsid proteins and tail-associated proteins, and lysis such as holins and Rz-like lysis proteins. Of the tail-associated proteins, two tail fibre and/or spike proteins were annotated for each przondovirus.

Within the genus *Przondovirus*, the main differences were found in the tail proteins (Fig. 2). The tail fibre and tail spike proteins are major determinants for host range, so phages that were isolated against the same *Klebsiella* host strain were expected to have higher sequence similarity across their tail fibre proteins. *Klebsiella* phages Oda, Toyotomi, Mera, Speegle, Cornelius and Tokugawa, which were isolated against the same *K. michiganensis* strain, shared considerable sequence similarity across their entire genomes, including the tail fibre proteins. Emom and Amrap were both isolated against the same *K. oxytoca* strain, where they shared sequence similarity across their entire genomes, including at the tail fibre protein location. The tail fibre protein sequence similarity is complemented by the host range data for these two phages. In contrast, Cornelius and its closest relative *Klebsiella* phage SH-KP152226 still shared a high degree of sequence similarity across their entire genome, including the tail proteins, despite infecting different host species (*K. michiganensis* and *K. pneumoniae*, respectively). In fact, all przondoviruses in this study were found to share significant sequence similarity in their tail proteins with their closest relatives, except for Emom, and by proxy Amrap, and *Klebsiella* phage KP32. There was a lower degree of sequence similarity in the first tail fibre protein between Emom and KP32, but there was no sequence similarity in the second tail protein between Emom and that of KP32. This is possibly due to their different isolation hosts, where KP32 had been isolated against a *K. pneumoniae* strain, and Emom/Amrap were isolated against a *K. oxytoca* strain.

The most striking differences, however, were in the tail proteins between przondoviruses in this study and reference phages that were not their closest BLASTn relatives. For example, Saitama showed sequence similarity with SH-KP152226 in only the initial part of the first tail fibre protein, with no sequence similarity exhibited elsewhere in the tail protein location. A similar pattern was observed for Emom and K11, and for Whistle and KP32. This is unsurprising since the isolation hosts for Emom and K11 were *K. oxytoca* and *K. pneumoniae*, respectively [69, 91]. Similarly, Whistle and KP32 infected two different species, *K. variicola* and *K. pneumoniae*, respectively. The differences in the tail fibre proteins therefore probably reflect the different isolation hosts for the przondoviruses in our collection and their database relatives.

Other differences between the closely related phages were found in the Rz-like lysis proteins, particularly within the przondoviruses that were within 95–98% similarity to one another. There is high sequence similarity for this protein between Cornelius and Oda, but not between Oda and Toyotomi, for example. Rz-like lysis proteins are involved in the lysis of the inner and outer membrane of Gram-negative bacteria and can be highly diverse [92–94]. These proteins may be part of a single-component system, or part of a two-component system: this is where one gene may be embedded within another, overlap another or exist as separate genes [92–94]. These genes encode two different proteins that operate together to disrupt the bacterial membrane, but appear to have distinct evolutionary origins [94]. The differences in membrane composition among different *Klebsiella* spp. could explain the differences in the Rz-like proteins, or may simply highlight differences between not only the proteins themselves, but the type of lysis system employed by each phage.

Our rationale for using ONT for our long-read sequencing was due to it being more widespread and affordable than other long-read sequencing technologies, such as PacBio. We performed a search of PacBio-assembled *Autographiviridae* and none of the nine genomes we found provided the raw reads for either long-read-only sequencing data or both long- and short-read data. The lack of publicly available long-read raw data for phage genomes makes validating this work using either ONT or other long-read technologies more challenging.

Similarly, our choice for short-read library preparation kit and sequencing technologies were selected based on their low cost and low DNA input requirements. However, the Illumina DNA Prep (formerly Nextera Flex) is a transposome-based library preparation kit, which does not allow the capture of the physical ends of linear genomes, but does allow the capture of the majority of the DTR sequence since there are two. While failure to capture the DTRs could be overcome using a different library preparation kit, this would not solve the assembly issue that HYPPA addresses, where the DTRs are being assembled in the middle of the genome. This issue would be much more difficult to resolve without HYPPA should the phage be novel, whereby presence or absence, length and sequence of potential DTRs are unknown and/or undetermined.

Conclusion

Here, we developed the HYPPA workflow for generating high-quality phage genomes that require minimal manual curation, and is most representative of what is actually biologically present within the phage capsid. We tested and validated the workflow using ten przondoviruses, negating the need for laborious primer walking and Sanger sequencing validation. Accurate phage genomes provide the necessary foundation for a mechanistic understanding of infection biology, which itself is integral to the use of phages within a phage therapy setting. Moreover, accurate phage genomes provide better understanding of the nucleotide and proteomic structure and how they fit into current taxonomic classification of phages. This is particularly important when performing comparative genomic analyses. We acknowledge that the production of high-quality phage genomes using this workflow requires sequencing and bioinformatic capabilities, and may be a limiting factor for some.

Funding information

CKAE is supported by the Medical Research Council (MRC) and JAFRAL as part of the Doctoral Antimicrobial Research Training (DART) MRC iCASE Programme, grant no. MR/R015937/1. TLB, AT, SKT and EMA gratefully acknowledge funding by the Biotechnology and Biological Sciences Research Council (BBSRC); this research was funded by the BBSRC Institute Strategic Programme Gut Microbes and Health BB/R012490/1 and its constituent projects BBS/E/F/000PR10353 and BBS/E/F/000PR10356. TLV, DJB and RE were supported by the Quadram Institute Bioscience BBSRC funded Core Capability Grant (project number BB/CCG1860/1). GT, HAK, RAK and MAW are supported by the BBSRC Institute Strategic Programme Microbes in the Food Chain BB/R012504/1 and its constituent projects BBS/E/F/000PR10348 and BBS/E/F/000PR10349. LJH is supported by Wellcome Trust Investigator Awards 100974 /C/13/Z and 220876/Z/20/Z; by the BBSRC Institute Strategic Programme Gut Microbes and Health BB/R012490/1, and its constituent projects BBS/E/F/000PR10353 and BBS/E/F/000PR10356.

Acknowledgements

We would like to thank Dr Oliver Charity for assistance with comparative genomics. We gratefully acknowledge CLIMB-BIG-DATA infrastructure (MR/T030062/1) support for high-performance computing. We would like to thank Dr Kata Farkas and Prof. Davey Jones at Bangor University for their assistance in procuring wastewater samples. We would like to thank Dr James Soothill at Great Ormond Street Hospital for additional clinical *Klebsiella* strains. We would like to thank all other members of the Adriaenssens Group, including Luke Acton at Quadram Institute Bioscience, for their feedback and support.

Author contributions

Conceptualization: C.K.A.E., TL.B., E.M.A. Data curation: C.K.A.E., T.L.V., R.E., D.J.B., S.K.T., G.T., H.A.K., R.A.K., L.J.H. Formal analysis: C.K.A.E., E.M.A. Funding acquisition: E.M.A. Investigation: C.K.A.E., T.L.V., G.T., H.A.K. Methodology: C.K.A.E., T.L.B., T.L.V., A.T., E.M.A. Software: T.L.V., A.T. Supervision: E.M.A., M.A.W. Validation: C.K.A.E., T.L.V., E.M.A. Visualization: C.K.A.E. Writing – original draft: C.K.A.E. Writing – review and editing: T.L.B., T.L.V., R.E., D.J.B., A.T., S.K.T., H.A.K., G.T., R.A.K., L.J.H., M.A.W., E.M.A.

Conflicts of interest

The authors declare that there are no conflicts of interest.

Ethical statement

An ethical statement for this study was not necessary since no clinical samples were processed. However, the ethical statement for the preterm infant isolates is available from the original study [25].

References

1. Turner D, Kropinski AM, Adriaenssens EM. A roadmap for genome-based phage taxonomy. *Viruses* 2021;13:506.
2. Evseev PV, Lukianova AA, Shneider MM, Korzhenkov AA, Bugaeva EN, *et al*. Origin and evolution of *Studiervirinae* bacteriophages infecting *Pectobacterium*: horizontal transfer assists adaptation to new niches. *Microorganisms* 2020;8:1707.
3. ICTV. Current ICTV Taxonomy Release. 2022; 2022. https://ictv.global/taxonomy [accessed 4 August 2022].
4. 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;168:74–82.
5. Adriaenssens EM, Sullivan MB, Knezevic P, van Zyl LJ, Sarkar BL, *et al*. Taxonomy of prokaryotic viruses: 2018-2019 update from the ICTV Bacterial and Archaeal Viruses subcommittee. *Arch Virol* 2020;165:1253–1260.
6. Molineux IJ. The T7 group. In: R Calendar (eds). *The Bacteriophages*, 2nd edn. Oxford: Oxford University Press; 2006. pp. 277–301.
7. Boeckman J, Korn A, Yao G, Ravindran A, Gonzalez C, *et al*. Sheep in wolves' clothing: temperate T7-like bacteriophages and the origins of the Autographiviridae. *Virology* 2022;568:86–100.

8. Lavigne R, Seto D, Mahadevan P, Ackermann H-W, Kropinski AM. Unifying classical and molecular taxonomic classification: analysis of the *Podoviridae* using BLASTP-based tools. *Res Microbiol* 2008;159:406–414.
9. Black LW. DNA packaging in dsDNA bacteriophages. *Annu Rev Microbiol* 1989;43:267–292.
10. Li S, Fan H, An X, Fan H, Jiang H, *et al*. Scrutinizing virus genome termini by high-throughput sequencing. *PLoS One* 2014;9:e85806.
11. Casjens SR, Gilcrease EB. Determining DNA packaging strategy by analysis of the termini of the chromosomes in tailed-bacteriophage virions. *Methods Mol Biol* 2009;502:91–111.
12. Merrill BD, Ward AT, Grose JH, Hope S. Software-based analysis of bacteriophage genomes, physical ends, and packaging strategies. *BMC Genomics* 2016;17:679.
13. Kutter E, Sulakvelidze A. *Bacteriophages*. In: Kutter E and Sulakvelidze A (eds). *Basic Phage Biology*. Boca Raton, Florida: CRC Press; 2004. pp. 29–66.
14. Maffei E, Shaidullina A, Burkolter M, Heyer Y, Estermann F, *et al*. Systematic exploration of *Escherichia coli* phage-host interactions with the BASEL phage collection. *PLoS Biol* 2021;19:e3001424.
15. Wyres KL, Lam MMC, Holt KE. Population genomics of *Klebsiella pneumoniae*. *Nat Rev Microbiol* 2020;18:344–359.
16. Paczosa MK, Mecsas J. *Klebsiella pneumoniae*: going on the offense with a strong defense. *Microbiol Mol Biol Rev* 2016;80:629–661.
17. Theuretzbacher U. Global antimicrobial resistance in Gram-negative pathogens and clinical need. *Curr Opin Microbiol* 2017;39:106–112.
18. Theuretzbacher U, Outterson K, Engel A, Karlén A. The global preclinical antibacterial pipeline. *Nat Rev Microbiol* 2020;18:275–285.
19. Kortright KE, Chan BK, Koff JL, Turner PE. Phage therapy: a renewed approach to combat antibiotic-resistant bacteria. *Cell Host Microbe* 2019;25:219–232.
20. Kęsik-Szeloch A, Drulis-Kawa Z, Weber-Dąbrowska B, Kassner J, Majkowska-Skrobek G, *et al*. Characterising the biology of novel lytic bacteriophages infecting multidrug resistant *Klebsiella pneumoniae*. *Virol J* 2013;10:100.
21. Olsen NS, Hendriksen NB, Hansen LH, Kot W. A New High-Throughput Screening Method for Phages: Enabling Crude Isolation and Fast Identification of Diverse Phages with Therapeutic Potential. *Phage* 2020;1:137–148.
22. Carlson K. Working with bacteriophages: common techniques and methodological approaches. In: E Kutter and A Sulakvelidze (eds). *Bacteriophages: Biology and Applications*, vol. 1. Boca Raton, Florida: CRC Press; 2005. pp. 437–494.
23. Grasis JA. Host-associated bacteriophage isolation and preparation for viral metagenomics. In: Pantaleo V and Chiumenti M (eds). *Viral Metagenomics: Methods and Protocols*. New York: Springer; 2018. pp. 1–25.
24. Wick RR, Judd LM, Holt KE. Assembling the perfect bacterial genome using Oxford Nanopore and Illumina sequencing. *PLoS Comput Biol* 2023;19:e1010905.
25. Chen Y, Brook TC, Soe CZ, O'Neill I, Alcon-Giner C, *et al*. Preterm infants harbour diverse *Klebsiella* populations, including atypical species that encode and produce an array of antimicrobial resistance- and virulence-associated factors. *Microb Genom* 2020;6:e000377.
26. Shin SH, Kim S, Kim JY, Lee S, Um Y, *et al*. Complete genome sequence of *Enterobacter aerogenes* KCTC 2190. *J Bacteriol* 2012;194:2373–2374.
27. Lee JH, Cheon IS, Shim B-S, Kim DW, Kim SW, *et al*. Draft genome sequence of *Klebsiella pneumoniae* subsp. *pneumoniae* DSM 30104T. *J Bacteriol* 2012;194:5722–5723.
28. Woodford N, Zhang J, Warner M, Kaufmann ME, Matos J, *et al*. Arrival of *Klebsiella pneumoniae* producing KPC carbapenemase in the United Kingdom. *J Antimicrob Chemother* 2008;62:1261–1264.
29. Chen M, Li Y, Li S, Tang L, Zheng J, *et al*. Genomic identification of nitrogen-fixing *Klebsiella Variicola K. pneumoniae* and *K. quasipneumoniae*. *J Basic Microbiol* 2016;56:78–84.
30. Schwengers O, Hoek A, Fritzenwanker M, Falgenhauer L, Hain T, *et al*. ASA3P: an automatic and scalable pipeline for the assembly, annotation and higher-level analysis of closely related bacterial isolates. *PLoS Comput Biol* 2020;16:e1007134.
31. Petit RA, Read TD. Bactopia: a flexible pipeline for complete analysis of bacterial genomes. *mSystems* 2020;5:e00190-20.
32. Jolley KA, Bliss CM, Bennett JS, Bratcher HB, Brehony C, *et al*. Ribosomal multilocus sequence typing: universal characterization of bacteria from domain to strain. *Microbiology* 2012;158:1005–1015.
33. Jolley KA, Bray JE, Maiden MCJ. Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. *Wellcome Open Res* 2018;3:124.
34. Lam MMC, Wick RR, Watts SC, Cerdeira LT, Wyres KL, *et al*. Genomic surveillance framework and global population structure for *Klebsiella pneumoniae*. *Genomics* 2021.
35. Wick RR, Heinz E, Holt KE, Wyres KL. Kaptive web: user-friendly capsule and lipopolysaccharide serotype prediction for *Klebsiella genomes*. *J Clin Microbiol* 2018;56:e00197-18.
36. Van Twest R, Kropinski AM. Bacteriophage enrichment from water and soil. *Methods Mol Biol* 2009;501:15–21.
37. Kropinski AM, Mazzocco A, Waddell TE, Lingohr E, Johnson RP. Enumeration of bacteriophages by double agar overlay plaque assay. *Methods Mol Biol* 2009;501:69–76.
38. Kropinski AM. Bacteriophages. In: Kutter E and Sulakvelidze A (eds). *Phage Host Range and Efficiency of Plating*. Totowa, NJ: CRC Press; 2009. pp. 141–149.
39. Andrews S. FastQC: a quality control tool for high throughput sequence data; 2010. http://www.bioinformatics.babraham.ac.uk/projects/fastqc [accessed 8 January 2021].
40. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 2018;34:i884–i890.
41. De Coster W, D'Hert S, Schultz DT, Cruts M, Van Broeckhoven C. NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics* 2018;34:2666–2669.
42. Lin Y, Yuan J, Kolmogorov M, Shen MW, Chaisson M, *et al*. Assembly of long error-prone reads using de Bruijn graphs. *Proc Natl Acad Sci*2016;113:E8396–E8405.
43. Wick RR, Holt KE. Benchmarking of long-read assemblers for prokaryote whole genome sequencing. *F1000Res* 2019;8:2138.
44. Wick RR, Judd LM, Cerdeira LT, Hawkey J, Méric G, *et al*. Trycycler: consensus long-read assemblies for bacterial genomes. *Genome Biol* 2021;22:266.
45. Chen Y, Zhang Y, Wang AY, Gao M, Chong Z. Accurate long-read de novo assembly evaluation with Inspector. *Genome Biol* 2021;22:312.
46. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, *et al*. Canu: scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. *Genome Res* 2017;27:722–736.
47. WickRR. Porechop; 2018. https://github.com/rrwick/Porechop [accessed 28 October 2002].
48. Wright C, Wykes M. Medaka; 2020. https://github.com/nanoporetech/medaka [accessed 22 September 2022].
49. Wick RR, Holt KE. Polypolish: short-read polishing of long-read bacterial genome assemblies. *PLoS Comput Biol* 2022;18:e1009802.
50. Zimin AV, Salzberg SL. The genome polishing tool POLCA makes fast and accurate corrections in genome assemblies. *PLoS Comput Biol* 2020;16:e1007981.
51. Seemann T. Shovill; 2018. https://github.com/tseemann/shovill [accessed 12 January 2021].
52. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, *et al*. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 2012;19:455–477.

53. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, *et al*. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* 2014;9:e112963.
54. Wick RR, Judd LM, Gorrie CL, Holt KE. Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput Biol* 2017;13:e1005595.
55. Garneau JR, Depardieu F, Fortier L-C, Bikard D, Monot M. PhageTerm: a tool for fast and accurate determination of phage termini and packaging mechanism using next-generation sequencing data. *Sci Rep* 2017;7:8292.
56. Benes V, Kilger C, Voss H, Pääbo S, Ansorge W. Direct primer walking on P1 plasmid DNA. *Biotechniques* 1997;23:98–100.
57. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. *Genome Biol* 2019;20:257–269.
58. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 2012;9:357–359.
59. Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative genomics viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* 2013;14:178–192.
60. Grubaugh ND, Gangavarapu K, Quick J, Matteson NL, De Jesus JG, *et al*. An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar. *Genome Biol* 2019;20:8.
61. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009;25:1754–1760.
62. Li H. Aligning sequence reads, clone sequences and assembly Contigs with BWA-MEM. *biorxiv* 2013.
63. Okonechnikov K, Golosova O, Fursov M. team tU. Unipro UGENE: a unified bioinformatics toolkit. *Bioinformatics* 2012;28:1166–1167.
64. Bouras G, Nepal R, Houtak G, Psaltis AJ, Wormald P-J, *et al*. Pharokka: a fast scalable bacteriophage annotation tool. *Bioinformatics* 2022;39:btac776.
65. McNair K, Zhou C, Dinsdale EA, Souza B, Edwards RA. PHANOTATE: a novel approach to gene identification in phage genomes. *Bioinformatics* 2019;35:4537–4542.
66. Hsieh P-F, Lin H-H, Lin T-L, Chen Y-Y, Wang J-T. Two T7-like bacteriophages, K5-2 and K5-4, each encodes two capsule depolymerases: isolation and functional characterization. *Sci Rep* 2017;7:4624.
67. Whitfield C, Lam M. Characterisation of coliphage K30, a bacteriophage specific for *Escherichia coli* capsular serotype K30. *FEMS Microbiol Lett* 1986;37:351–355.
68. Teng T, Li Q, Liu Z, Li X, Liu Z, *et al*. Characterization and genome analysis of novel *Klebsiella* phage Henu1 with lytic activity against clinical strains of *Klebsiella pneumoniae*. *Arch Virol* 2019;164:2389–2393.
69. Rudolph C, Freund-Mölbert E, Stirm S. Fragments of *Klebsiella* bacteriophage no. 11. *Virology* 1975;64:236–246.
70. Thiry D, Passet V, Danis-Wlodarczyk K, Lood C, Wagemans J, *et al*. New bacteriophages against emerging lineages ST23 and ST258 of *Klebsiella pneumoniae* and efficacy assessment in *Galleria mellonella* larvae. *Viruses* 2019;11:411.
71. Wu Y, Wang R, Xu M, Liu Y, Zhu X, *et al*. A novel polysaccharide depolymerase encoded by the phage SH-KP152226 confers specific activity against multidrug-resistant *Klebsiella pneumoniae* via biofilm degradation. *Front Microbiol* 2019;10:2768.
72. Labudda Ł, Strapagiel D, Karczewska-Golec J, Golec P. Complete annotated genome sequences of four *Klebsiella pneumoniae* phages isolated from Sewage in Poland. *Genome Announc* 2017;5:45.
73. Liu Y, Leung SSY, Huang Y, Guo Y, Jiang N, *et al*. Identification of two depolymerases from phage IME205 and their antivirulent functions on K47 capsule of *Klebsiella pneumoniae*. *Front Microbiol* 2020;11:218.
74. Kwon H-J, Cho S-H, Kim T-E, Won Y-J, Jeong J, *et al*. Characterization of a T7-like lytic bacteriophage (phiSG-JL2) of *Salmonella enterica* serovar gallinarum biovar gallinarum. *Appl Environ Microbiol* 2008;74:6970–6979.
75. Hamdi S, Rousseau GM, Labrie SJ, Kourda RS, Tremblay DM, *et al*. Characterization of five podoviridae phages infecting *Citrobacter freundii*. *Front Microbiol* 2016;7:1023.
76. Bleriot I, Blasco L, Pacios O, Fernández-García L, Ambroa A, *et al*. The role of PemIK (PemK/PemI) type II TA system from *Klebsiella pneumoniae* clinical strains in lytic phage infection. *Sci Rep* 2022;12:4488.
77. Dunn JJ, Studier FW. Complete nucleotide sequence of bacteriophage T7 DNA and the locations of T7 genetic elements. *J Mol Biol* 1983;166:477–535.
78. Dobbins AT, George M, Basham DA, Ford ME, Houtz JM, *et al*. Complete genomic sequence of the virulent *Salmonella* bacteriophage SP6. *J Bacteriol* 2004;186:1933–1944.
79. McGinnis S, Madden TL. BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res* 2004;32:W20–W25.
80. Darling AE, Mau B, Perna NT, Stajich JE. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS One* 2010;5:e11147.
81. Moraru C, Varsani A, Kropinski AM. VIRIDIC- a novel tool to calculate the intergenomic similarities of prokaryote-infecting viruses. *Viruses* 2020;12:1268.
82. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: molecular evolutionary genetics analysis across computing platforms. *Mol Biol Evol* 2018;35:1547–1549.
83. Letunic I, Bork P. Interactive tree of life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res* 2021;49:W293–W296.
84. Gilchrist CLM, Chooi Y-H. Clinker & clustermap.js: automatic generation of gene cluster comparison figures. *Bioinformatics* 2021;37:2473–2475.
85. Nobrega FL, Vlot M, de Jonge PA, Dreesens LL, Beaumont HJE, *et al*. Targeting mechanisms of tailed bacteriophages. *Nat Rev Microbiol* 2018;16:760–773.
86. González-García VA, Pulido-Cid M, Garcia-Doval C, Bocanegra R, van Raaij MJ, *et al*. Conformational changes leading to T7 DNA delivery upon interaction with the bacterial receptor. *J Biol Chem* 2015;290:10038–10044.
87. Latka A, Maciejewska B, Majkowska-Skrobek G, Briers Y, Drulis-Kawa Z. Bacteriophage-encoded virion-associated enzymes to overcome the carbohydrate barriers during the infection process. *Appl Microbiol Biotechnol* 2017;101:3103–3119.
88. Eskenazi A, Lood C, Wubbolts J, Hites M, Balarjishvili N, *et al*. Combination of pre-adapted bacteriophage therapy and antibiotics for treatment of fracture-related infection due to pandrug-resistant *Klebsiella pneumoniae*. *Nat Commun* 2022;13:302.
89. Piel D, Bruto M, Labreuche Y, Blanquart F, Goudenège D, *et al*. Phage-host coevolution in natural populations. *Nat Microbiol* 2022;7:1075–1086.
90. Turner D, Adriaenssens EM, Tolstoy I, Kropinski AM. Phage annotation guide: guidelines for assembly and high-quality annotation. *Phage*2021;2:170–182.
91. Pan Y-J, Lin T-L, Chen C-T, Chen Y-Y, Hsieh P-F, *et al*. Genetic analysis of capsular polysaccharide synthesis gene clusters in 79 capsular types of *Klebsiella* spp. *Sci Rep* 2015;5:15573.
92. Summer EJ, Berry J, Tran TAT, Niu L, Struck DK, *et al*. Rz/Rz1 lysis gene equivalents in phages of Gram-negative hosts. *J Mol Biol* 2007;373:1098–1112.
93. Berry J, Summer EJ, Struck DK, Young R. The final step in the phage infection cycle: the Rz and Rz1 lysis proteins link the inner and outer membranes. *Mol Microbiol* 2008;70:341–351.
94. Kongari R, Rajaure M, Cahill J, Rasche E, Mijalis E, *et al*. Phage spanins: diversity, topological dynamics and gene convergence. *BMC Bioinformatics* 2018;19:326.