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], Evelien M. Adriaenssens[1*]

[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.adriaenssens@quadram.ac.uk

Keywords: *Klebsiella*; bacteriophage; phage; *Przondovirus*; sequencing; assembly.

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; NCBI, National Centre for Biotechnology Information; MLST, multilocus sequence typing; ONT, Oxford Nanopore Technologies; PEG, polyethylene glycol; QC, quality control; QIB, Quadram Institute Bioscience; RNAP, RNA polymerase.

ABSTRACT

Bacteriophages (phages) within the *Przondovirus* genus are T7-like podoviruses belonging to the *Studiervirinae* subfamily, within the *Autographiviridae* family and have a highly conserved genome organisation. The genome size of these phages ranges from 37 kb to 42 kb, 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 data demonstrate the importance of careful curation of phage assemblies before publication, and prior to using them for comparative genomics.

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.

DATA SUMMARY

Phage raw reads are available from the National Centre 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**). 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 **Table S1** and **Table 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 *Przondovirus* genus are T7-like podoviruses, meaning they have a short tail morphotype, belonging to the *Studiervirinae* subfamily, within the *Autographiviridae* family (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 *Autographiviridae* family (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 *Studiervirinae* subfamily is highly conserved: all 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 heterogenous group of Gram-negative bacteria in the Enterobacterales order (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 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).

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

109 MATERIALS AND METHODS

110 Bacterial strains and growth conditions

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

118 Preparation of bacterial DNA and sequencing

119 Genomic DNA for each *Klebsiella* strain was extracted using the AllPrep Bacterial
120 DNA/RNA/Protein kit (QIAGEN) according to the manufacturer's instructions. DNA was
121 quantified as previously described and normalised to 5 ng μL^{-1}.
122 DNA was prepared using the Illumina DNA Prep library preparation kit and was whole-
123 genome sequenced on the Illumina NextSeq500 platform generating 2 x 150 bp paired-end
124 reads by QIB Sequencing Core Services.
125 Additionally, *K. michiganensis* M7 21 2 #35, *K. pneumoniae* M26 18 1, *K. pneumoniae* M26
126 18 2 #21 KpnN, *K. pneumoniae* M26 18 2 #21 KpnA, and *K. pneumoniae* ST38 01 were
127 prepared for enhanced sequencing according to the sequence facilities instructions
128 (MicrobesNG). Genomes were provided assembled and annotated by MicrobesNG.

129 **Bacterial genomics**

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

140 **Isolation and single-plaque purification of phages**

141 Samples from various UK wastewater treatment plants were screened for *Klebsiella*-
142 specific phages using a range of *Klebsiella* strains as hosts for enrichment, adapted from Van
143 Twest *et al.* (36). Briefly, 300 µL filtered wastewater was mixed with 60 µL exponential bacterial
144 culture and used to inoculate 5 mL BHI broth. Enrichments were incubated overnight at 37°C
145 with shaking at 200 rpm. Enrichments were then centrifuged (4000 x *g* for 15 min) and passed
146 through a 0.45 µm filter before spot testing by double agar overlay plaque assay, as previously
147 described (37). All incubations for overlay method were performed over 4-17 h at 37°C. Single
148 plaque purifications were made by extracting single plaques from the soft agar layer using
149 sterile toothpicks and suspended in approximately 300 µL BHI broth. Suspensions were
150 centrifuged (13,000 x *g* for 5 min) and supernatant collected. Ten-fold serial dilutions of the
151 supernatant were performed in phage buffer (75 mM NaCl; 10 mM $MgSO_4$; 10 mM Tris, pH
152 7.5; 0.1 mM $CaCl_2$) and 10 µL of each dilution plated onto double agar overlay and incubated
153 as described above. This process was repeated at least three times to create phage stocks.
154 Phage amplification was performed as for single plaque purification in BHI broth. Once
155 supernatant was collected, approximately 100 µL of phage suspension was spread onto three
156 double agar overlay plates and incubated as before. Phage stocks were prepared by
157 extraction of phage clearance zones. This was achieved by removal of the soft agar layer,
158 which was resuspended in phage buffer, and centrifuged (4000 x *g* for 15 min). Phage
159 supernatant was passed through a 0.45 µm filter into a sterile glass universal and stored at
160 4°C.

161 **Phage host range**

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

165 **Phage DNA extraction and whole-genome sequencing**

166 Phage virions were concentrated by polyethylene glycol (PEG) 8000 (Thermo Fisher)
167 precipitation for DNA extraction. Briefly, phage stock was treated with 1 µL DNase I (10 U μL^{-1})
168 (Merck) and 1 µL RNase A (10 U μL^{-1}) (Merck) per mL of stock and incubated at 37°C for
169 30 min. PEG precipitation was performed with PEG 8000 (10% w/v) and 1 M NaCl and
170 incubated overnight at 4°C. The precipitate was centrifuged (17,000 x *g* for 10 min) and
171 resuspended in 200 µL nuclease-free water. Resuspended phage pellets were treated with
172 proteinase K (50 µg mL^{-1}) (Merck), EDTA (final concentration 20 mM), and 10% SDS (final
173 concentration 0.5% v/v) and incubated at 55°C for 1 h.
174 DNA was extracted using the Maxwell® RSC Viral Total Nucleic Acid Purification kit
175 (Promega), as per the manufacturer's instructions into nuclease-free water. Phage DNA was
176 quantified by Qubit 3.0 fluorometer using the high sensitivity dsDNA kit (Invitrogen). DNA was
177 prepared using Illumina DNA Prep (formerly Nextera Flex) library preparation kit and was
178 whole-genome sequenced on the Illumina NextSeq500 platform generating 2 x 150 bp paired-
179 end reads by QIB Sequencing Core Services. MinION libraries (Oxford Nanopore
180 Technologies, ONT) were constructed without shearing using the short fragment buffer and
181 loaded onto the R9.4.1 flow cell according to the manufacturer's instructions by QIB
182 Sequencing Core Services.
183 Both long-read and short-read raw data for all ten przondoviruses were deposited in NCBI
184 under BioProject number PRJNA914245.

185 **Phage genomics**

186 *Assembly and annotation*

187 All quality control, pre-processing, assembly, and annotation of phage genomes were
188 performed on the QIB Galaxy platform.
189 We checked short-read data for quality using fastQC v0.11.8 (39). Based on this fastQC
190 analysis, reads were pre-processed with fastp v0.19.5 (40), using a hard trim of between 4
191 and 10 bases on both the front and tail to retain at least a per base quality of 28.
192 Long-read data was demultiplexed following sequencing and quality checked with
193 NanoStat v0.1.0 (41). Pre-processing was performed as part of the assembly, and assembled
194 using Flye v2.9 (42) with default settings, which included correction and trimming of reads.
195 Flye was used in the first instance as previously published work has determined it is the most

196 accurate and reliable assembler (43-45). Where Flye was unable to generate a high-quality
197 assembly, Canu v2.2 (46) was used as an alternative. Error correction and trimming were
198 performed as part of the default settings when assembling using Flye or Canu. Flye
199 additionally performed one iteration of long-read polishing by default. We assembled all
200 phages with and without trimming adapter/barcode sequences for long-reads. Trimming was
201 performed with Porechop v0.2.3 (https://github.com/rrwick/Porechop) (47) with default
202 settings.

203 We performed several iterations of long-read and short-read polishing on long-read-only
204 assemblies in a specific order. Firstly, two iterations of long-read polishing were performed
205 using Medaka (48) with default settings, using the previous polished data as the input for the
206 next round of polishing. Secondly, one iteration of short-read polishing was performed using
207 Polypolish (49) with default settings. Finally, a second iteration of short-read polishing was
208 performed using POLCA (50) with default settings. We used raw reads for each iteration of
209 long-read polishing and pre-processed reads for each iteration of short-read polishing.

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

221 Following assembly, the contigs were manually checked for DTRs flanking the genome, as
222 well as with PhageTerm (55) which was unable to identify the DTRs since it does not work
223 well for Nextera-based sequence libraries. Where we could not determine the length and
224 sequence of the DTRs, we performed primer walking. Outward-facing primers were designed
225 to “walk” the genome termini using Sanger sequencing (56). Phage DNA was extracted, and
226 for each phage at least two primers were designed for the reverse strand to walk the beginning
227 of the genome and identify the left terminal repeat, and at least two primers were designed for
228 the forward strand to walk the end of the genome to identify the right terminal repeat. The
229 phage DNA and each primer were then sent for Sanger sequencing separately (Eurofins,
230 Germany). Sanger sequences were visualised in FinchTV v1.5.0
231 (https://digitalworldbiology.com/FinchTV) and compared to the reference phage genome, and

232 DTRs annotated using the Molecular Biology suite on the Benchling platform
233 (https://www.benchling.com/).
234 Assemblies generating multiple contigs were checked for contamination using Kraken 2
235 v2.1.1 (57).
236 Verification of the DTRs and assessment of assembly quality was performed by mapping
237 the raw reads back to the assembled genome using Bowtie 2 v2.3.4.3 (58) and visualised
238 using IGV v2.7.2 (59), and variant calling performed using iVar v1.0.1 (60). Additionally, BWA-
239 MEM v0.7.17.1 (https://github.com/lh3/bwa) was used to map long-reads back to the
240 reference using default settings optimised for ONT reads (61, 62).
241 Assemblies in the reverse orientation were reorientated by reverse complementation of the
242 genome in UGENE v38.0 (63) and uploaded to Benchling. Contigs were then reoriented to
243 begin at the same start point, based on well-curated reference phages and the analysis of the
244 DTRs.
245 Genome annotation was performed using Pharokka v1.2.1 with default settings
246 (https://github.com/gbouras13/pharokka) (64). Specifically, coding sequences were predicted
247 with PHANOTATE (65).

248 *Comparative genomics*

249 Where specified, publicly-available phage genomes used for comparative genomics were
250 derived from these studies (20, 66-78), listed in **Table S6**, and downloaded from the GenBank
251 database.
252 The closest relative for each phage was determined as as the top hit according to maximum
253 score identified by nucleotide BLAST (BLASTn) (https://blast.ncbi.nlm.nih.gov/Blast.cgi) and
254 optimised for somewhat similar sequences (79). Genes associated with specific phage
255 families were identified and used for preliminary taxonomic assignment. Alignments were
256 performed using Mauve v20150226 (80) between the closest relative and phages from the
257 same genera. The intergenomic similarity between przondoviruses in the collection and a
258 selection of publicly-available related phages was calculated using VIRIDIC on the web server
259 (http://rhea.icbm.uni-oldenburg.de/VIRIDIC/) (81).
260 Phylogenetic analyses were performed using the hallmark DNA-directed RNA polymerase
261 (RNAP) amino acid sequence for all phages and a selection of publicly-available
262 phylogenetically-related phages downloaded from the NCBI protein database
263 (https://www.ncbi.nlm.nih.gov/). Multisequence alignment of the RNAP amino acid sequences
264 was performed using the MUSCLE algorithm in MEGA X v10.0.5 (82) with default settings. A
265 maximum-likelihood tree was generated with 500 boostraps using the default Jones-Taylor-
266 Thornton model. Phylogenetic analysis was performed using 35 amino acid sequences, with

267 a total of 684 positions in the final analysis. Tree image rendering was performed using iTOL
268 v6.1.1 (https://itol.embl.de/) (83).
269 Linear mapping of coding sequences for phage final assemblies was performed using
270 Clinker v.0.0.23 (84).

271 **RESULTS AND DISCUSSION**

272 **Phage isolation and host range determination**

273 In this study, we isolated ten lytic T7-like phages from a variety of river water and
274 wastewater samples, using four different *Klebsiella* spp. as isolation hosts (**Table 1**). To
275 examine the host range, we tested the ten phages against a collection of *Klebsiella* spp. from
276 different sources, representing a range of capsule and sequence types. All phages had a
277 narrow host range, with seven being able to infect only a single *Klebsiella* strain within our
278 collection (**Fig. 1**).
279 Three of the ten przondoviruses were used to test and validate the HYPPA workflow: Oda,
280 Toyotomi, and Tokugawa. As the three unifiers of the HYPPA workflow, these were named
281 after the three unifiers of Japan (see **Development of a new workflow for the assembly of**
282 **complete phage genomes**).
283 Only three of our phages were capable of productively infecting more than one *Klebsiella*
284 strain. *Klebsiella* phage Toyotomi was able to infect two different species; a *K. michiganensis*
285 strain (its isolation host) and a *K. quasipneumoniae* strain. *Klebsiella* phages Emom and
286 Amrap were both able to infect two different isolates of *K. oxytoca. Klebsiella* phage Whistle
287 was the only phage capable of productively infecting four strains of *Klebsiella*, spanning three
288 different species (*K. grimontii, K. michiganensis,* and *K. variicola*), and caused lysis without
289 productive infection on a further three *K. pneumoniae* isolates. We could not establish a link
290 between capsular type and host range for these phages.
291 Przondoviruses and other T7-like phages have a relatively small genome of 37 to 42 kb,
292 and this may limit their host expansion capabilities (for taxonomic assignment of the ten
293 phages in this study, see section **Phage genome characterisation and taxonomy**).
294 However, Whistle was capable of infecting multiple hosts, along with Emom, Amrap, and
295 Toyotomi. Previous work has shown that T7-like phages are capable of infecting multiple hosts
296 (66) and that host range is determined by interaction between phage receptor binding proteins,
297 i.e. tail fibre and/or spike proteins, and bacterial cell receptors (14, 66, 85). LPS components
298 are almost always identified as the secondary receptor for irreversible attachment in Gram-
299 negative-targeting podoviruses (6, 14). Whether initial interaction with the outer membrane
300 and degradation of the CPS constitutes a *bone fide* reversible attachment step, or whether

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. CDS	Accession	Closest database relative according to BLASTn		
								Name	Coverage (%)	ID (%)
Oda	River water	K.mi.	41,642	52.64	181	58	OQ579023	*Klebsiella* phage SH-KP152226	92.0	94.62
Toyotomi	Wastewater	K.mi.	41,268	52.64	180	55	OQ579024	*Klebsiella* phage SH-KP152226	92.0	94.71
Mera	Wastewater	K.mi.	41,400	52.58	180	56	OQ579025	*Klebsiella* phage SH-KP152226	92.0	94.34
Speegle	Wastewater	K.mi.	41,395	52.64	180	58	OQ579026	*Klebsiella* phage SH-KP152226	93.0	94.70
Cornelius	Wastewater	K.mi.	40,437	52.72	180	55	OQ579027	*Klebsiella* phage SH-KP152226	94.0	94.84
Tokugawa	Wastewater	K.mi.	41,414	52.64	181	56	OQ579028	*Klebsiella* phage SH-KP152226	92.0	94.70
Saitama	Wastewater	K.qp.	40,741	53.06	181	51	OQ579029	*Klebsiella* phage K11	96.0	95.71
Emom	Wastewater	K.ox.	40,788	52.56	183	53	OQ579030	*Klebsiella* phage KP32	94.0	93.14
Amrap	Wastewater	K.ox.	41,209	52.47	182	57	OQ579031	*Klebsiella* phage KPN3	85.0	95.05
Whistle	Wastewater	K.va.	40,735	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* DSM15968. CDS, coding sequences. Bacterial host species accessions are given in **Table S2**.

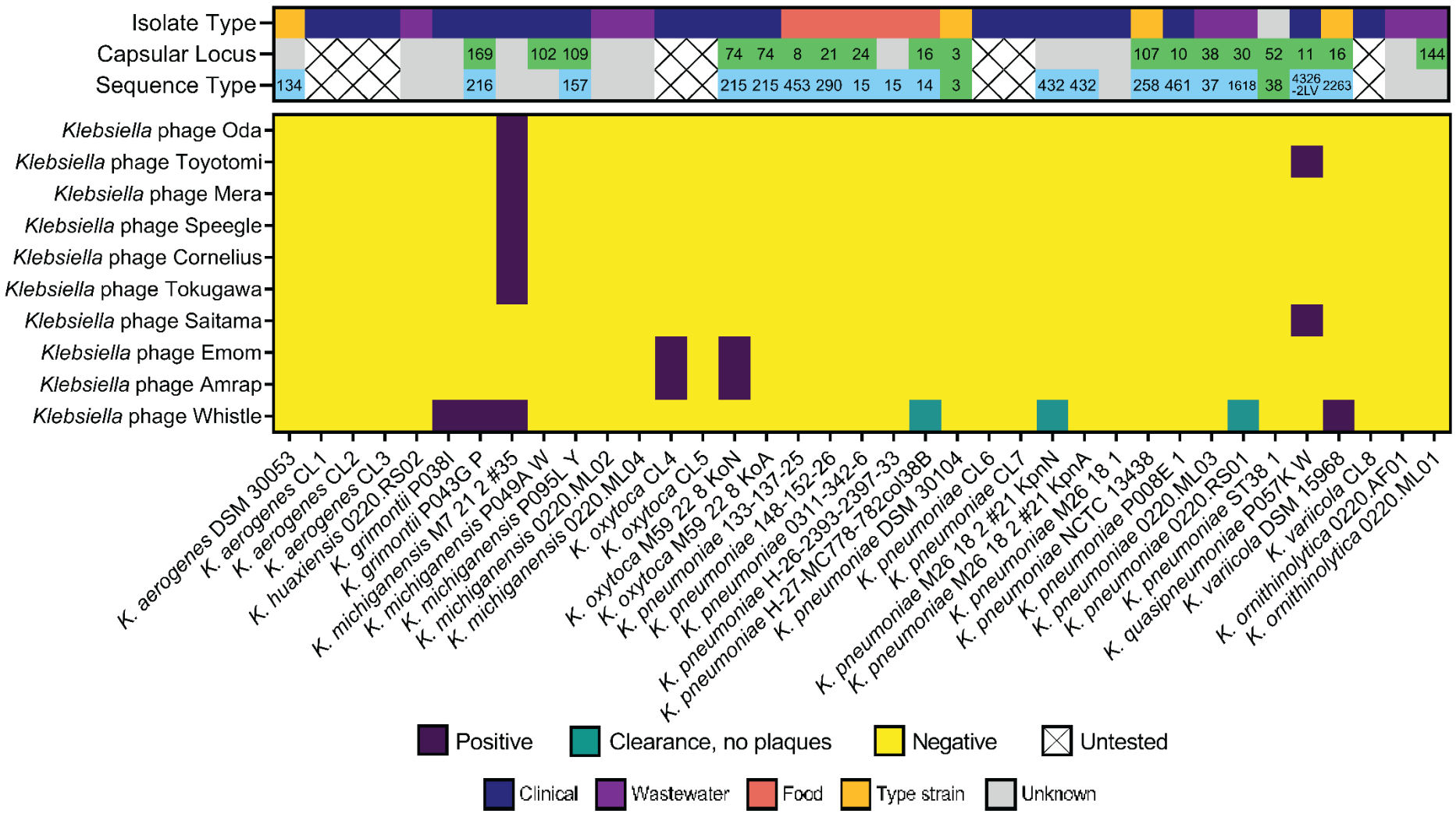

Fig. 1. Heatmap for host range of the przondoviruses in the collection by plaque assay against a diverse range of *Klebsiella* spp. Top panel, isolate type, capsular type, and sequence type. The source of each isolate is given as isolate type, with grey being unknown source. Capsular loci determined by Kaptive and/or Kleborate, green; unknown or no match confidence, grey. Sequence type (ST) determined by multilocus sequence typing, blue; unknown or incomplete matches, grey. No sequencing data available, untested. 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.

304 this is a prerequisite to reversible attachment by the phage to another outer membrane
305 component is yet to be fully elucidated (6, 14, 86, 87).
306 Some phages can be "trained" to increase their host range through co-evolution assays
307 (19, 88). This may be particularly useful in cases of lysis from without, such as observed in
308 Whistle, as they are already capable of binding to host receptors but unable to cause
309 productive infection.
310 Intriguingly, Toyotomi was the only phage capable of infecting two different *Klebsiella*
311 species, that none of its closest relatives from our collection were capable of infecting, despite
312 exceptionally high protein sequence similarity across their tail fibre proteins.
313 Multiple factors affect host range and broadly involve extracellular and intracellular
314 mechanisms. Extracellular mechanisms involve the ability of phages to bind to specific phage
315 receptors on the bacterial cell surface that facilitate DNA ejection (89). Intracellular
316 mechanisms involve evasion of phage defence systems that facilitate phage propagation (89).
317 Expression of diffusible depolymerases facilitate interaction of phages with their primary and
318 secondary receptor. This extracellular mechanism is more likely to explain the ability of Whistle
319 to infect more than one isolate since there is productive infection. Thus, the ability several
320 przondoviruses in our collection to infect different *Klebsiella* isolates could indicate that they
321 share similarities in the chemical composition of their capsules, enabling degradation by a
322 single depolymerase and allowing access to the phage receptors on the bacterial cell.
323 Moreover, the bacterial isolates could share similar sugar motifs within their LPS structures,
324 which are thought to be the secondary receptor of phages within the *Autographiviridae* family
325 (6).

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

327 To generate complete and accurate genomes for these ten phages, which included
328 resolving the defined ends of phage genomes, and correcting sequencing and/or assembly
329 errors, we utilised a long-read-only assembly with sequential polishing steps. This
330 methodology exploited both long-read and short-read sequencing data in a workflow that we
331 have named HYPPA – HYbrid and Poly-polish Phage Assembly (see also **Materials and**
332 **Methods**) before moving onto annotation and comparative genomics (**Fig. 2**). Firstly, the long-
333 reads were assembled using Flye or Canu, followed by two iterations of long-read polishing
334 with Medaka. Next, we performed two iterations of short-read polishing using Polypolish (for
335 the first iteration) and POLCA (for the second iteration).
336 Initially, Flye was used as the primary assembler in our HYPPA workflow and worked
337 particularly well for phages with both very high sequence read coverage (Toyotomi at
338 >117,000x) and very low sequence read coverage, which included Mera (8x), Speegle (23x),
339 and Amrap (27x) (**Table S2**). However, Canu performed better with the other phages as the

340 assemblies in general contained fewer errors. This is contrary to previously published literature
341 that found Flye was more the more accurate assembler using default settings (43-45).
342 As an illustration of the HYPPA workflow, we provided a more detailed description of the
343 process for phage Oda as an exemplar, for which the DTRs were validated with primer
344 walking. Firstly, Oda was assembled using Canu, which yielded one contig of 41,761 bp. After
345 two iterations of long-read polishing followed by two iterations of short-read polishing, the
346 resulting contig was 41,769 bp in size. We were able to identify the terminal repeat regions,
347 but both were flanked by a 64 bp sequence upstream of the left terminal repeat, and
348 downstream of the right terminal repeat after all polishing iterations were complete. The two
349 64 bp sequences were inverted repeats containing adapter sequences of 23 bp, with the
350 remaining sequence being Nanopore barcodes which were manually removed. HYPPA was
351 then used for phage Tokugawa, which after short-read only assembly had included a 79 bp
352 repeat within the genome, but outside of the presumed DTR region (**Fig. S1**). Using HYPPA,
353 the repeat was determined to be an assembly artefact and removed from the assembly. The
354 final curated assembly for phages Oda and Tokugawa was 41,642 bp and 41,414 bp,
355 respectively. Terminal repeats were present for both phages and complete at 181 bp, validated
356 by primer walking and Sanger sequencing (**Fig. S1**).
357 We trimmed the long-reads using Porechop in an attempt to remove the adapter/barcode
358 sequences, but when phage Oda was reassembled and polished using the trimmed reads,
359 the right terminal repeat was missing three bases, but no other single nucleotide
360 polymorphisms (SNPs) or indels were identified.
361 The HYPPA workflow without Porechop-mediated trimming was repeated for the remaining
362 eight przondoviruses, resulting in final genome assemblies ranging between 40-42 kb (**Table**
363 **1**). HYPPA was able to generate a complete genome for phage Toyotomi, where short-read-
364 only, long-read-only, and hybrid assemblies were unable to do so and resulted in fragmented
365 assemblies. Although our HYPPA workflow is a hybrid assembly approach, there is a clear
366 distinction between this and traditional hybrid assembly methods. Importantly, HYPPA used
367 the short-reads for polishing only, not during the genome assembly, whereas traditional hybrid
368 assemblies utilise both long-read and short-read data during the assembly process itself.
369 Moreover, short-read polishing of a long-read-only assembly using Pilon was also unable to
370 resolve the genome of Toyotomi: partial repeat regions were found at the termini but were
371 incomplete, and multiple errors within coding regions persisted. Using HYPPA, we were able
372 to not only preserve the DTRs of Toyotomi, but also correct persistent sequencing and/or
373 assembly errors that occurred in all non-HYPPA assemblies.
374 The genome organisation of genera within the *Autographiviridae* family is highly conserved:
375 all genes are unidirectional and show a high degree of synteny, and genomes are flanked by
376 DTRs (2, 5-7). The DTRs of the przondoviruses described here were 180-183 bp in size,

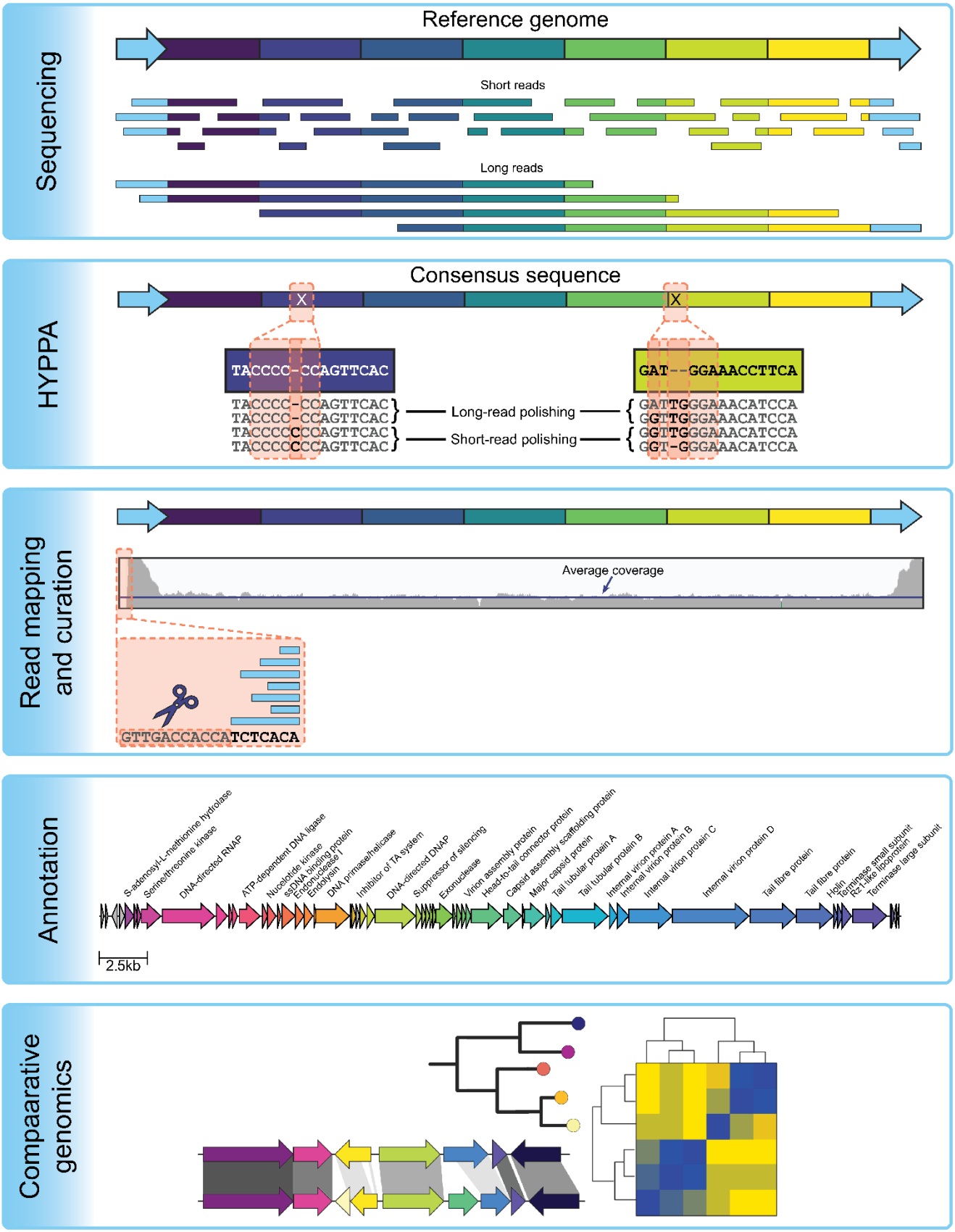

Fig. 2. Workflow for phage genome assembly and analysis. Both long-read and short-read sequencing are recommended, followed by our HYPPA workflow for high quality phage genome assembly. Read mapping for short-read data and manual curation can correct any errors that were missed during polishing, followed by annotation and then comparative genomics.

378 demonstrating sequence similarity of 84.3-99.7%. DTRs are thought to assist circularisation
379 of the phage genome once in the host cytoplasm to prevent host-induced enzymatic digestion
380 (13). Thus, resolution of the DTRs is integral to accurate genomics and understanding of the
381 biology of different phages.

382 **Comparison of HYPPA with traditional short-read-only assembly**

383 When compared to typical short-read-only methodologies of phage genome assembly, in our
384 case using Shovill (51), the HYPPA workflow required significantly less manual curation (**Fig.**
385 **2**). Typically, phage genomes are assembled using short-read only data, and many of these
386 genomes are then published without additional curation, leaving them with potentially
387 significant sequencing and/or assembly errors. Using short-read-only assembly methods for
388 our collection of przondoviruses, we observed that some were in the reverse orientation rather
389 than the forward orientation as is expected for 50% of the assemblies, and some had the
390 DTRs assembled in the middle of the contig. Addressing these issues required manually re-
391 orienting the assemblies and ensuring they all had the same start position, as suggested in
392 the Phage Annotation Guide (90). In contrast, the HYPPA workflow resulted in assemblies
393 with correct start and stop sites, but some were still in the reverse orientation.
394 To check for DTRs in short-read only assemblies, we initially looked for increased reads
395 within the read mapping profiles, which are distinguished by one or two large peaks, and can
396 be automated using the tool PhageTerm (55). If a single peak was observed anywhere other
397 than at either end of the assembly, the assembly had been opened in the middle of the genome
398 and required each to be re-oriented to have the same starting position.
399 Incorrect orientation is a feature of phage genome assembly, and with short-read-only data
400 in particular, may be artificially linearised by the assembler with the DTRs located in the middle
401 of the contig. In many of our own short-read-only assemblies, the przondoviruses described
402 here were linearised in the middle of the genome, and required read mapping to identify where
403 the DTRs may be. In T7-like phages, DNA is concatemeric and requires the assistance of
404 terminases to cut at specific sites to package the DNA into the procapsid (9-11). Although
405 each concatemer contains a single copy of the repeat, a second repeat is synthesised at the
406 other end of the genome to prevent loss of genetic material (9, 12). Since the DTRs are present
407 twice per phage genome, the number of terminal sequences is double following whole genome
408 sequencing and are identified as a single peak of increased reads during read mapping (10-
409 12, 55). Therefore, the DTR and by proxy, the start of the genome can be inferred from the
410 read mapping. Moreover, due to the highly conserved nature of the genomes, all
411 przondoviruses had almost the same starting sequence as the well-curated Enterobacteria
412 phage K30 (accession HM480846) (67), making the beginning relatively easy to find. As a

413 result, considerable time was spent on re-orienting the short-read-only assemblies to be
414 unidirectional and to have the same starting sequence.
415 One of the most problematic aspects using short-reads for phage assembly (both short-
416 read-only and has part of a traditional hybrid assembly) was that the DTRs were deleted,
417 possibly because the assemblers used deem them to be a sequencing artefact. Thus, DTRs
418 need to be manually validated through primer walking and Sanger sequencing validation.
419 However, this was unnecessary when using short-reads for polishing rather than for assembly.
420 Thus, using the HYPPA workflow, the DTRs were present in the final polished assembly in the
421 correct location at the ends and did not have to be manually added.
422 A second type of error that routinely occurred during non-HYPPA phage sequencing and
423 assembly was the introduction of short insertions and/or deletions (indels) that were
424 particularly noticeable in coding regions.
425 For the short-read only assemblies, many sequencing and assembly errors present in
426 coding regions were only found upon annotation of the genomes, including frameshift errors
427 in DNA polymerase (DNAP) and tail fibre protein genes. Often, these frameshift errors were
428 found in homopolymer regions and were introduced during sequencing. Before using HYPPA,
429 these frameshift errors were checked through read mapping followed by variant calling and
430 edited accordingly. Particularly noteworthy were repeat regions of ~79 bp identified close to
431 and sometimes within the DTR regions of seven of the ten phages (See **Development of a**
432 **new workflow for the assembly of complete phage genomes** for description of repeats for
433 Tokugawa), but that did not correlate with the increased reads observed in the read mapping.
434 This suggested that these repeats were introduced in error during assembly and were
435 confirmed to be artefacts in most phages, including Tokugawa through Sanger sequencing
436 (see supplementary **Fig. S1**). Using HYPPA, we found that the two iterations of short-read
437 polishing were able to correct single nucleotide polymorphisms and/or correct indels that
438 resulted in these frameshift errors that long-read polishing was unable to resolve, particularly
439 in homopolymer regions. POLCA was also able to correct indels that Polypolish was unable
440 to resolve.
441 As previously described for Oda, all the przondoviruses contained adapter and barcode
442 DNA upstream and/or downstream of the DTR regions. Initially, as we were trying to
443 reconstruct the linear genome ends, we did not perform adapter and barcode trimming of the
444 Nanopore reads prior to the long-read assembly. We then removed these sequences manually
445 after assembly. To limit the amount of manual curation, Porechop can be used to trim the
446 reads, however, when we attempted this for all the remaining przondoviruses, Porechop-
447 mediated trimming resulted in several further errors. These included trimming bases from the
448 beginning of the left terminal repeat and the end of the right terminal repeat, ranging from 3-
449 18 bp in total; indels; multiple SNPs; and in some cases failure to assemble the phage genome

450 into a single contig, or at all. We would thus recommend manual removal of the
451 adapter/barcodes rather than trimming of long-reads using Porechop, which appears to
452 require more manual curation when compared to using raw Nanopore reads.
453 Multiple sequencing and/or assembly errors were identified in the coding regions of other
454 phages, that again, persisted following traditional methods of phage assembly. Using trial and
455 error, we were able to show that the HYPPA method was superior to other methods of phage
456 assembly, whether hybrid or through using a single sequencing platform in correcting errors
457 (see supplemental **Tables S2-S5** for all assembly details). Moreover, the HYPPA workflow
458 required far fewer manual curation steps than traditional phage assembly methods: while long-
459 read only assemblies were sometimes in the reverse orientation, all were linearised at the
460 starting sequence. This is in contrast with the traditional assembly methods that required re-
461 orienting the genomes to be unidirectional and starting at the same position, manual correction
462 of large assembly errors such as indels, manual correction of homopolymer errors in coding
463 regions, and in some cases, rearrangement of contigs and manual stitching the genome
464 together, followed by primer walking and Sanger sequencing validation to determine the
465 genome termini and DTRs.
466 Errors in homopolymer sequences and repeat regions are particularly common in long-
467 read-only assemblies of bacterial genomes (43, 44, 49), and as we have described here, in
468 phage genomes also. Indeed, two homopolymer errors occurred in the DNAP of Toyotomi,
469 leading to a double frameshift error that resulted in three protein annotations. Short-read
470 polishing can correct errors introduced during long-read-only assemblies (49), as we have
471 demonstrated here. Similarly to using short-read data for polishing, we found that a traditional
472 hybrid assembly using both short- and long-read data for Toyotomi also introduced large
473 deletions in repeat regions, with assembly errors persisting, as has been described previously
474 (44, 54). Assembly metadata showing all previous long-read-only, short-read-only, and hybrid
475 assemblies is provided (see supplemental **Tables S2-S5**).
476 Several limitations of this study include the need for both short-read and long-read data for
477 phage assembly, and specialised knowledge to access and install the software which is all
478 freely available. Which polishing program used and what type of polishing (long-read versus
479 short-read) in what order may give different results of equal validity. While we believe that the
480 HYPPA workflow provides the most accurate phage genome possible, it still may not exactly
481 reflect the DNA that is present within each phage capsid. Additionally, while the highly
482 conserved nature of T7-like phages made it easier to determine the DTR starting sequence,
483 this may not be the case for novel phages.
484
485

486 **Phage genome characterisation and taxonomy**

487 All ten phages were dsDNA phages at 40,336-41,720 bp with a GC content of 52.40-
488 53.06% which is slightly lower than their isolation host GC content of ~55.46-57.59% (**Table**
489 **1, Fig. 3**). The number of predicted coding sequences within the genomes varies from 51 to
490 58, and almost all coding sequences were found in the same orientation on the forward strand.
491 However, five phages had one to four small hypothetical proteins found in opposite orientation.
492 We performed BLASTn on all phages to determine closest relatives in the NCBI GenBank
493 database (as of October 2022). Based on the BLASTn results, which showed high levels of
494 nucleotide similarity with reference phages, the phages in our collection were preliminarily
495 assigned to the *Przondovirus* genus within the *Studiervirinae* subfamily and *Autographiviridae*
496 family, according to the currently established ICTV genus demarcation criterion of 70%
497 nucleotide sequence similarity over the genome length to belong to the same genus (1).
498 The genomic relationships between our novel przondoviruses and a selection of
499 *Autographiviridae* reference phages were explored further by conducting a nucleotide-based
500 intergenomic similarity analysis using VIRIDIC (**Fig. 4**, **Table S2**). Included within the analysis
501 were relatives within the same genus (*Przondovirus*), those within different genera but the
502 same subfamily (*Studiervirinae*), and those within different subfamilies (*Molineuxvirinae*,
503 *Slopekvirinae*) (**Fig. 4**). These data confirmed that the przondoviruses from this study were
504 within the ICTV genus demarcation criterion of 70% nucleotide sequence similarity over the
505 genome length when compared to other przondoviruses. Several genera within the subfamily
506 *Studiervirinae* that were included only shared ~45-57% nucleotide sequence similarity with the
507 przondoviruses in this study (**Fig. 4**).
508 Several przondoviruses clustered more closely together, including *Klebsiella* phages Oda,
509 Toyotomi, Mera, Speegle, Cornelius, and Tokugawa, which were within ~98% nucleotide
510 similarity, except Cornelius which was the most dissimilar at ~95-96% (**Fig. 4**). All
511 aforementioned phages except Oda were isolated from the same wastewater treatment plant
512 at different stages of the treatment process, using the same host. These phages are therefore
513 likely to be different strains of the same new species of phage within the *Przondovirus* genus.
514 Emom and Amrap clustered with their closest relative KP32, but also clustered together with
515 ~92% similarity, and should be assigned to separate species (**Fig. 4**). Saitama and Whistle
516 did not cluster closely with any other phage from our collection, possibly due to differences in
517 their host specificity. Saitama did cluster with its closest relative *Klebsiella* phage K11, and
518 Whistle clustered with its closest relative IME264 (**Fig. 4**). This suggests that Saitama, Emom,
519 Amrap, and Whistle should be assigned to different species within the same genus.
520 After comparative genomic analyses, we observed that several of the closest database
521 relatives were deposited in databases with incomplete genomes. Specifically, the incompleteness

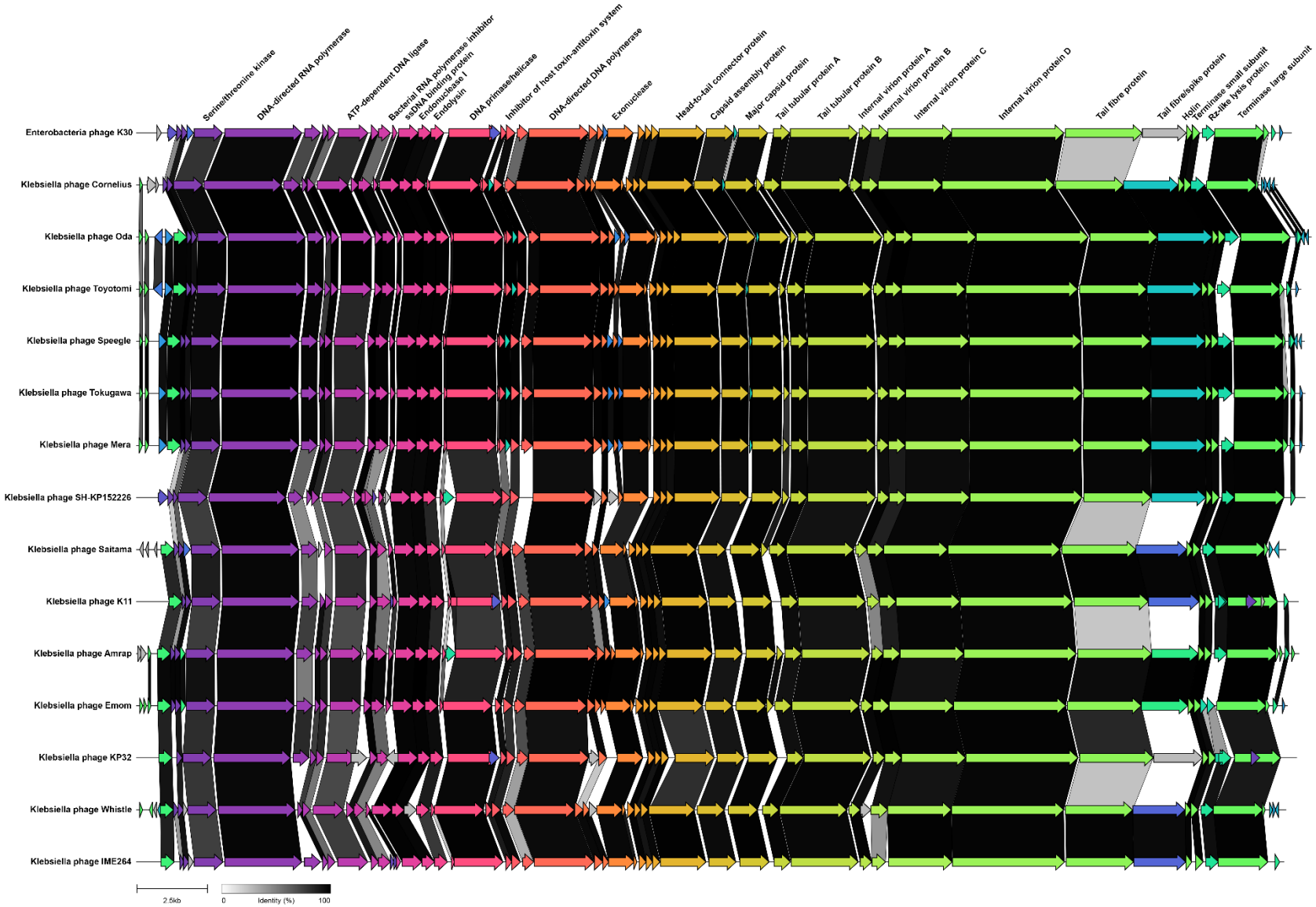

Fig. 3. Genome map and gene clustering for przondoviruses in the collection and a selection of related phages. Arrows represent coding sequences and pairwise comparisons of gene similarities are indicated by percentage identity given as links in greyscale, with darker shading representing areas of higher 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.

484 was most often due to an absence of the DTRs, including *Klebsiella* phages KP32, KPN3, and
485 IME264 (**Table S2, Fig. S2.**). Incomplete genomes could lead to incorrect assignments to
486 species in cases where the reciprocal nucleotide identities are close to the species threshold
487 of 95% similarity across the genome length (1).
488 Additionally, potential errors were noted in phages KPN3 (accession MN101227) and KMI1
489 (accession MN052874) (**Table S2**, **Fig. S2**). For example, KPN3 contained no annotated
490 DNA-directed DNAP, which is conserved across all *Przondovirus* genomes analysed here.
491 KMI1 contained a shorter DNA-directed RNAP annotation that, when included in the
492 phylogenetic analyses, showed higher divergence, which could not be confirmed, and was
493 therefore excluded from our phylogenetic analysis. Without raw short-read and long-read data,
494 it is difficult to determine whether these are genuine errors or whether their differences are a
495 true representation of the genome.
496 To further verify the taxonomic classification of the phages, phylogenetic analysis was
497 performed using the protein sequence of the DNA-dependent RNAP, since it is the hallmark
498 gene of the *Autographiviridae* family, using a selection of publicly available phages from the
499 genera *Apdecimavirus*, *Berlinvirus*, *Przondovirus*, *Teetrevirus*, and *Teseptimavirus*, within the
500 subfamily *Studiervirinae* (**Fig. 5**). As expected, the pzondoviruses clustered together, and there
501 was a clear separation from other phage genera. There were some slight differences between
502 the clustering patterns exhibited in the phylogenetic tree when compared to the VIRIDIC
503 analysis using whole nucleotide data. *Klebsiella* phages Emom and Amrap exhibited relatively
504 high similarity, sharing 91.5% sequence similarity across the whole nucleotide sequence, but
505 this distinction is less obvious in the phylogenetic analysis. As observed previously (91), a
506 single gene phylogenetic tree at the amino acid level will not provide enough resolution to
507 accurately display within-genus relationships, leading to discrepancies between clustering of
508 the phages Saitama, Emom, Amrap and Whistle in the VIRIDIC plot and the phylogenetic tree.

509 **Genome organisation and synteny**

510 We conducted comparative genomic analysis of przondoviruses according to coding
511 sequence similarity with a selection of reference phages (**Fig. 3**). We selected Enterobacteria
512 phage K30 as the representative isolate of the *Przondovirus* genus since its genome is well-
513 curated. Przondoviruses were grouped together with their closest relative according to BLASTn.
514 As expected, all phages share a highly conserved genome organisation, which revealed a
515 high degree of gene synteny, in concordance with the VIRIDIC data (**Fig. 4**).
516 All genomes were found to contain the early, middle, and late genes associated with viral
517 host takeover, DNA replication, and virion assembly and lysis, respectively (**Fig. 3**). The host
518 takeover proteins that were annotated included the S-adenosyl-L-methionine hydrolase, which
519 is a good marker for the start of the genome; serine/threonine kinase; and DNA-directed

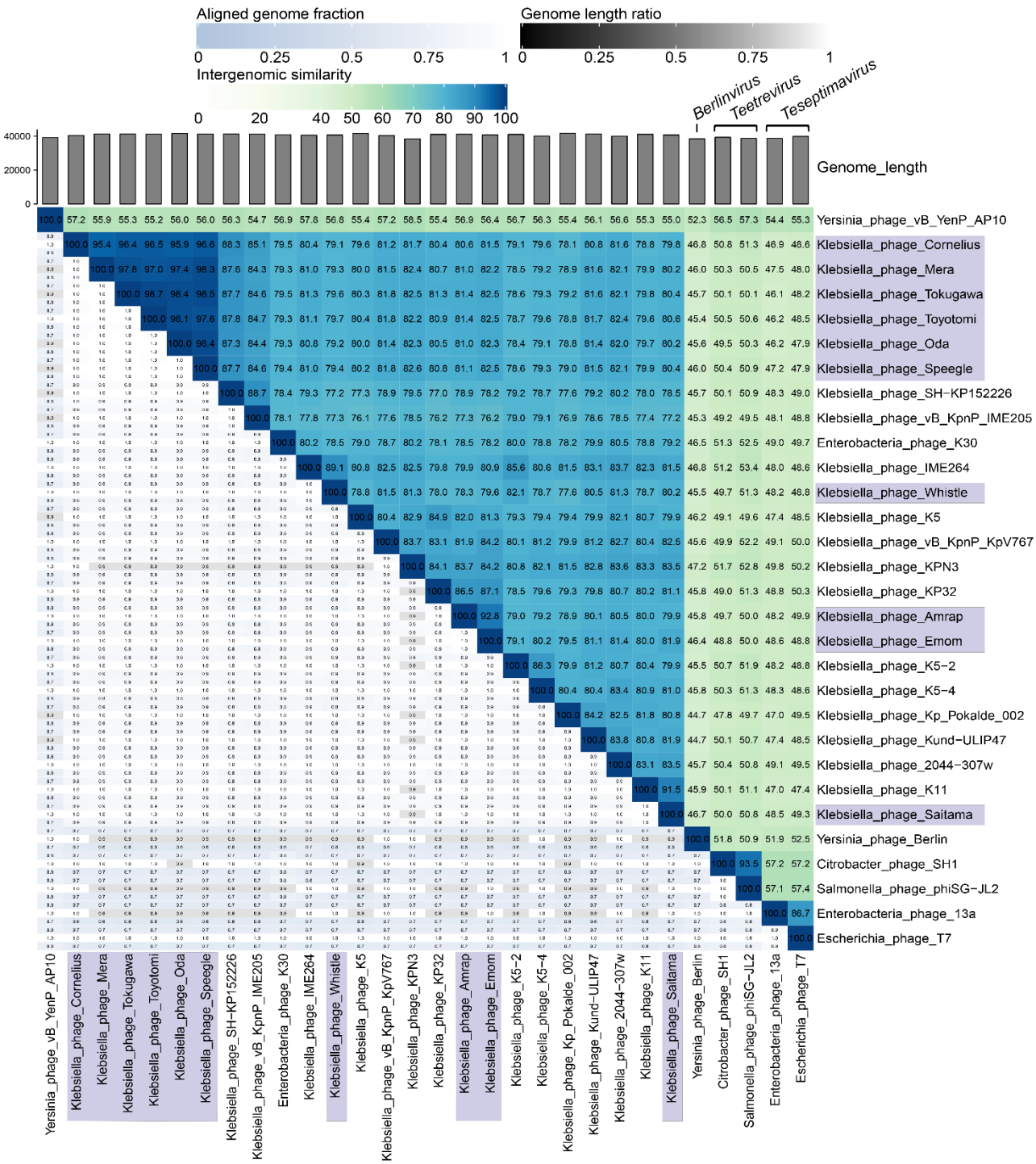

Fig. 4. Nucleotide-based intergenomic similarities of przondoviruses in the collection and a selection of related phages within the *Studiervirinae* subfamily, using VIRIDIC. A heatmap of hierarchical clustering of the intergenomic similarity values was generated and given as percentage values (right 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 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 *Apdecimavirus* genus.

557 RNAP, with the latter being a hallmark of the *Autographiviridae* family (5, 8). The middle
558 proteins annotated were typical for phage DNA replication. The late proteins included all the
559 components necessary for virion assembly, such as capsid proteins and tail-associated
560 proteins, and lysis such as holins and Rz-like lysis proteins. Of the tail-associated proteins,
561 two tail fibre and/or spike proteins were annotated for each przondovirus.

562 Within the *Przondovirus* genus, the main differences were found in the tail proteins (**Fig.**
563 **3**). The tail fibre and tail spike proteins are major determinants for host range, so phages that
564 were isolated against the same *Klebsiella* host strain were expected to have higher sequence
565 similarity across their tail fibre proteins. *Klebsiella* phages Oda, Toyotomi, Mera, Speegle,
566 Cornelius, and Tokugawa, which were isolated against the same *K. michiganensis* strain,
567 shared considerable sequence similarity across their entire genomes, including the tail fibre
568 proteins. However, one difference was that Toyotomi was able to infect two hosts, whereas
569 the remaining phages were not. This is intriguing since the high homology in the tail fibre
570 protein would suggest similar narrow host range capabilities for this subset of przondoviruses
571 in our collection. Emom and Amrap were both isolated against the same *K. oxytoca* strain,
572 where they shared sequence similarity across their entire genomes, including at the tail fibre
573 protein location. The tail fibre protein sequence similarity is complemented by the host range
574 data for these two phages. In contrast, Cornelius and its closest relative *Klebsiella* phage SH-
575 KP152226 still shared a high degree of sequence similarity across their entire genome,
576 including the tail proteins, despite infecting different host species (*K. michiganensis* and *K.*
577 *pneumoniae*, respectively). In fact, all przondoviruses in this study were found to share
578 significant sequence similarity in their tail proteins with their closest relatives, except for
579 Emom, and by proxy Amrap, and *Klebsiella* phage KP32. There was a lower degree of
580 sequence similarity in the first tail fibre protein between Emom and KP32, but there was no
581 sequence similarity in the second tail protein between Emom and that of KP32. This is possibly
582 due to their different isolation hosts, where KP32 had been isolated against a *K. pneumoniae*
583 strain, and Emom/Amrap were isolated against a *K. oxytoca* strain.

584 The most striking differences however, were in the tail proteins between przondoviruses in
585 this study and reference phages that were not their closest BLASTn relatives. For example,
586 Saitama showed sequence similarity with SH-KP152226 in only the initial part of the first tail
587 fibre protein, with no sequence similarity exhibited elsewhere in the tail protein location. A
588 similar pattern was observed for Emom and K11, and Whistle and KP32. This is unsurprising
589 since the isolation host for Emom and K11 are *K. oxytoca* and *K. pneumoniae*, respectively
590 (69, 92). Similarly, Whistle and KP32 infected two different species, *K. variicola* and *K.*
591 *pneumoniae*, respectively. The differences in the tail fibre proteins therefore likely reflect the
592 different isolation hosts for the przondoviruses in our collection and their database relatives.

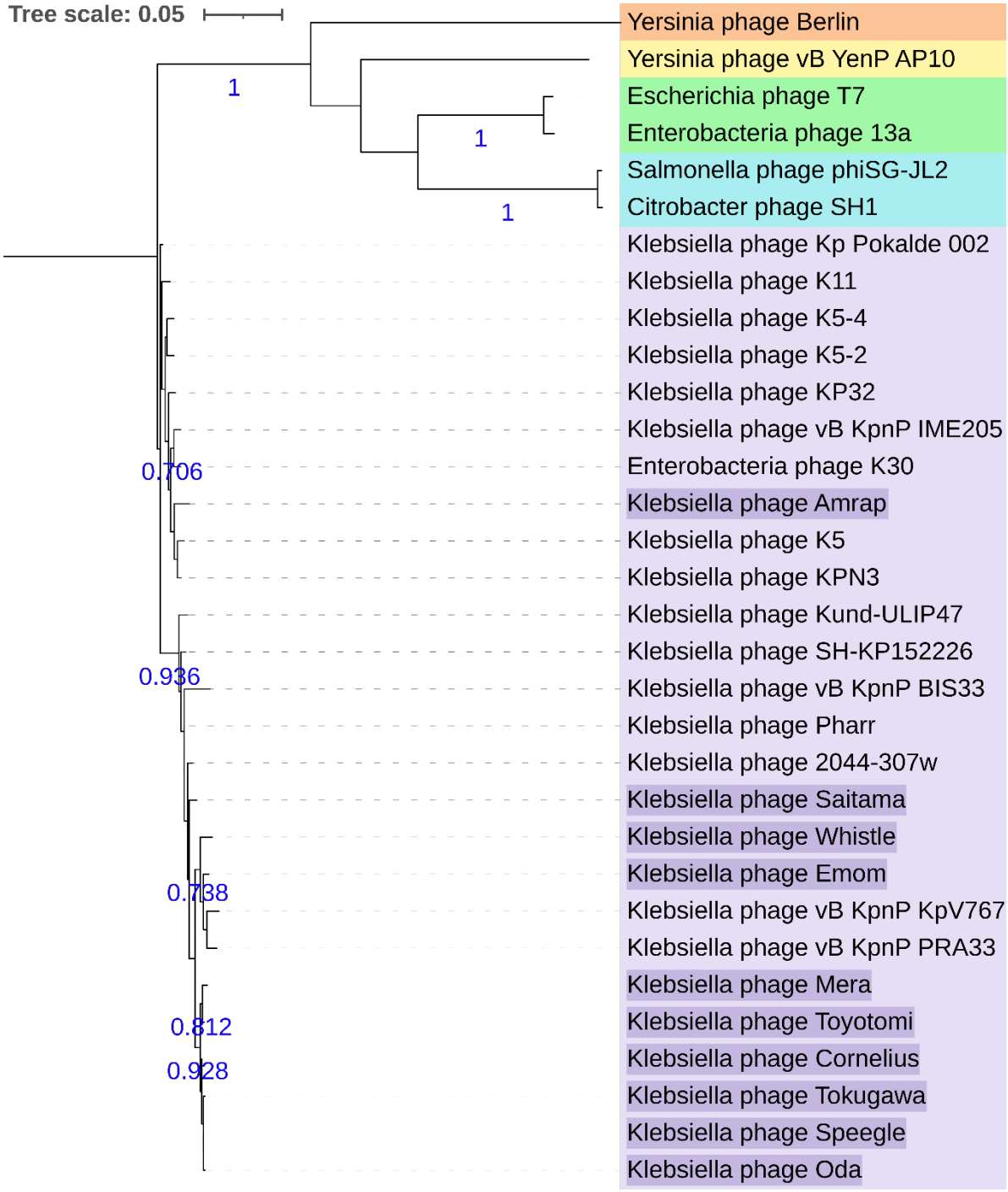

Fig. 5. Maximum-likelihood phylogeny of the RNAP for przondoviruses in this study and a selection of related *Studiervirinae* phages. All phages from this study (purple highlight) clustered with related przondoviruses (purple). Outgroups, *Berlinvirus* (orange), *Apidecimavirus* (yellow), *Teseptimavirus* (green), and *Teetrevirus* (blue). Tree is midpoint rooted. Bootstrap support values at ≥0.7 are given in blue (500 replicates).

603 Other differences between the closely related phages were found in the Rz-like lysis
604 proteins, particularly within the przondoviruses that were within 95-98% similarity to one
605 another. There is high sequence similarity for this protein between Cornelius and Oda, but not
606 between Oda and Toyotomi, for example. Rz-like lysis proteins are involved in the lysis of the
607 inner and outer membrane of Gram-negative bacteria and can be highly diverse (93-95).
608 These proteins may be part of a single-component system, or part of a two-component system:
609 this is where one gene may be embedded within another, overlap another, or exist as separate
610 genes (93-95). These genes encode two different proteins that operate together to disrupt the
611 bacterial membrane, but appear to have distinct evolutionary origins (95). The differences in
612 membrane composition among different *Klebsiella* spp. could explain the differences in the
613 Rz-like proteins, or may simply highlight differences between not only the proteins themselves,
614 but the type of lysis system employed by each phage.

615 **CONCLUSION**

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

627 **FUNDING**

628 CKAE is supported by the Medical Research Council (MRC) and JAFRAL as part of the
629 Doctoral Antimicrobial Research Training (DART) MRC iCASE Programme, grant no.
630 MR/R015937/1. TLB, AT, SKT, and EMA gratefully acknowledge funding by the Biotechnology
631 and Biological Sciences Research Council (BBSRC); this research was funded by the BBSRC
632 Institute Strategic Programme Gut Microbes and Health BB/R012490/1 and its constituent
633 projects BBS/E/F/000PR10353 and BBS/E/F/000PR10356. TLV, DJB, and RE were
634 supported by the Quadram Institute Bioscience BBSRC funded Core Capability Grant (project
635 number BB/CCG1860/1). GT, HAK, RAK, and MAW are supported by the BBSRC Institute
636 Strategic Programme Microbes in the Food Chain BB/R012504/1 and its constituent projects
637 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.

AUTHOR CONTRIBUTIONS

Conceptualisation: CKAE, TLB, EMA.
Data curation: CKAE, TLV, RE, DJB, SKT, GT, HAK, RAK, LJH.
Formal analysis: CKAE, EMA.
Funding acquisition: EMA.
Investigation: CKAE, TLV, GT, HAK.
Methodology: CKAE, TLB, TLV, AT, EMA.
Software: TLV, AT.
Supervision: EMA, MAW.
Validation: CKAE, TLV, EMA.
Visualisation: CKAE.
Writing – original draft: CKAE.
Writing – review and editing: TLB, TLV, RE, DJB, AT, SKT, HAK, GT, RAK, LJH, MAW, EMA.

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.

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).

670 **REFERENCES**

671 1. Turner D, Kropinski AM, Adriaenssens EM. A roadmap for genome-based phage
672 taxonomy. *Viruses*. 2021; 13(3). doi: 10.3390/v13030506.
673 2. Evseev PV, Lukianova AA, Shneider MM, Korzhenkov AA, Bugaeva EN, Kabanova
674 AP *et al.* Origin and evolution of *Studiervirinae* bacteriophages infecting
675 *Pectobacterium*: horizontal transfer assists adaptation to new niches. *Microorganisms*.
676 2020; 8(11). doi: 10.3390/microorganisms8111707.
677 3. ICTV. Current ICTV Taxonomy Release. 2022. Available from:
678 https://ictv.global/taxonomy. (Accessed: 04 Aug 2022).
679 4. Turner D, Shkoporov AN, Lood C, Millard AD, Dutilh BE, Alfenas-Zerbini P *et al.*
680 Abolishment of morphology-based taxa and change to binomial species names: 2022
681 taxonomy update of the ICTV bacterial viruses subcommittee. *Arch Virol*. 2023; 168(2):
682 74-82. doi: 10.1007/s00705-022-05694-2.
683 5. Adriaenssens EM, Sullivan MB, Knezevic P, van Zyl LJ, Sarkar BL, Dutilh BE *et al.*
684 Taxonomy of prokaryotic viruses: 2018-2019 update from the ICTV Bacterial and
685 Archaeal Viruses Subcommittee. *Arch Virol*. 2020; 165(5): 1253-1260. doi:
686 10.1007/s00705-020-04577-8.
687 6. Molineux IJ. The T7 group. In: R Calendar, ed. *The Bacteriophages*. 2nd Ed. 2006.
688 Oxford University Press: Oxford. pp. 277-301.
689 7. Boeckman J, Korn A, Yao G, Ravindran A, Gonzalez C, Gill J. Sheep in wolves'
690 clothing: Temperate T7-like bacteriophages and the origins of the Autographiviridae.
691 *Virology*. 2022; 568: 86-100. doi: 10.1016/j.virol.2022.01.013.
692 8. Lavigne R, Seto D, Mahadevan P, Ackermann H-W, Kropinski AM. Unifying classical
693 and molecular taxonomic classification: analysis of the *Podoviridae* using BLASTP-
694 based tools. *Res Microbiol*. 2008; 159(5): 406-414. doi:
695 https://doi.org/10.1016/j.resmic.2008.03.005.
696 9. Black LW. DNA packaging in dsDNA bacteriophages. *Annu Rev Microbiol*. 1989;
697 43(1): 267-292. doi: 10.1146/annurev.mi.43.100189.001411.
698 10. Li S, Fan H, An X, Fan H, Jiang H, Chen Y *et al.* Scrutinizing virus genome termini by
699 high-throughput sequencing. *PLoS One*. 2014; 9(1): e85806. doi:
700 10.1371/journal.pone.0085806.
701 11. Casjens SR, Gilcrease EB. Determining DNA packaging strategy by analysis of the
702 termini of the chromosomes in tailed-bacteriophage virions. *Methods Mol Biol*. 2009;
703 502: 91-111. doi: 10.1007/978-1-60327-565-1_7.
704 12. Merrill BD, Ward AT, Grose JH, Hope S. Software-based analysis of bacteriophage
705 genomes, physical ends, and packaging strategies. *BMC Genomics*. 2016; 17(1): 679.
706 doi: 10.1186/s12864-016-3018-2.
707 13. Guttman B, Raya R, Kutter E. Basic phage biology. In: E Kutter, A Sulakvelidze, eds.
708 *Bacteriophages: Biology and Applications*. 2005. CRC Press: Boca Raton, Florida. pp.
709 29-66.
710 14. Maffei E, Shaidullina A, Burkolter M, Heyer Y, Estermann F, Druelle V *et al.* Systematic
711 exploration of *Escherichia coli* phage-host interactions with the BASEL phage
712 collection. *PLoS Biol*. 2021; 19(11): e3001424. doi: 10.1371/journal.pbio.3001424.
713 15. Wyres KL, Lam MMC, Holt KE. Population genomics of *Klebsiella pneumoniae*. *Nat*
714 *Rev Microbiol*. 2020. doi: 10.1038/s41579-019-0315-1.
715 16. Paczosa MK, Mecsas J. *Klebsiella pneumoniae*: going on the offense with a strong
716 defense. *Microbiol Mol Biol Rev*. 2016; 80(3): 629-661. doi: 10.1128/MMBR.00078-15.
717 17. Theuretzbacher U. Global antimicrobial resistance in Gram-negative pathogens and
718 clinical need. *Curr Opin Microbiol*. 2017; 39: 106-112. doi: 10.1016/j.mib.2017.10.028.
719 18. Theuretzbacher U, Outterson K, Engel A, Karlén A. The global preclinical antibacterial
720 pipeline. *Nat Rev Microbiol*. 2020; 18(5): 275-285. doi: 10.1038/s41579-019-0288-0.
721 19. Kortright KE, Chan BK, Koff JL, Turner PE. Phage therapy: a renewed approach to
722 combat antibiotic-resistant bacteria. *Cell Host Microbe*. 2019; 25(2): 219-232. doi:
723 10.1016/j.chom.2019.01.014.

20. Kęsik-Szeloch A, Drulis-Kawa Z, Weber-Dąbrowska B, Kassner J, Majkowska-Skrobek G, Augustyniak D *et al.* Characterising the biology of novel lytic bacteriophages infecting multidrug resistant *Klebsiella pneumoniae*. *Virol J*. 2013; 10: 100. doi: 10.1186/1743-422x-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(3): 137-148. doi: 10.1089/phage.2020.0016.
22. Carlson K. Working with bacteriophages: common techniques and methodological approaches. In: E Kutter, A Sulakvelidze, eds. *Bacteriophages: Biology and Applications*. 1. 2005. CRC Press: Boca Raton, Florida. pp. 437-494.
23. Grasis JA. Host-associated bacteriophage isolation and preparation for viral metagenomics. In: V Pantaleo, M Chiumenti, eds. *Viral Metagenomics: Methods and Protocols*. 2018. Springer New York: New York, New York. 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(3): e1010905. doi: 10.1371/journal.pcbi.1010905.
25. Chen Y, Brook TC, Soe CZ, O'Neill I, Alcon-Giner C, Leelastwattanagul O *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. doi: 10.1099/mgen.0.000377.
26. Shin SH, Kim S, Kim JY, Lee S, Um Y, Oh MK *et al.* Complete genome sequence of *Enterobacter aerogenes* KCTC 2190. *J Bacteriol*. 2012; 194(9): 2373-2374. doi: 10.1128/jb.00028-12.
27. Lee JH, Cheon IS, Shim BS, Kim DW, Kim SW, Chun J *et al.* Draft genome sequence of *Klebsiella pneumoniae* subsp. *pneumoniae* DSM 30104T. *J Bacteriol*. 2012; 194(20): 5722-5723. doi: 10.1128/jb.01388-12.
28. Woodford N, Zhang J, Warner M, Kaufmann ME, Matos J, Macdonald A *et al.* Arrival of *Klebsiella pneumoniae* producing KPC carbapenemase in the United Kingdom. *J Antimicrob Chemother*. 2008; 62(6): 1261-1264. doi: 10.1093/jac/dkn396.
29. Chen M, Li Y, Li S, Tang L, Zheng J, An Q. Genomic identification of nitrogen-fixing *Klebsiella variicola*, *K. pneumoniae* and *K. quasipneumoniae*. *J Basic Microbiol*. 2016; 56(1): 78-84. doi: 10.1002/jobm.201500415.
30. Schwengers O, Hoek A, Fritzenwanker M, Falgenhauer L, Hain T, Chakraborty T *et al.* ASA^{3}P: An automatic and scalable pipeline for the assembly, annotation and higher-level analysis of closely related bacterial isolates. *PLoS Comput Biol*. 2020; 16(3): e1007134. doi: 10.1371/journal.pcbi.1007134.
31. Petit RA, 3rd, Read TD. Bactopia: a flexible pipeline for complete analysis of bacterial genomes. *mSystems*. 2020; 5(4). doi: 10.1128/mSystems.00190-20.
32. Jolley KA, Bliss CM, Bennett JS, Bratcher HB, Brehony C, Colles FM *et al.* Ribosomal multilocus sequence typing: universal characterization of bacteria from domain to strain. *Microbiology*. 2012; 158(4): 1005-1015. doi: 10.1099/mic.0.055459-0.
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. doi: 10.12688/wellcomeopenres.14826.1.
34. Lam MMC, Wick RR, Watts SC, Cerdeira LT, Wyres KL, Holt KE. Genomic surveillance framework and global population structure for *Klebsiella pneumoniae*. [Pre-print]. 2021. Doi: 10.1101/2020.12.14.422303.
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(6). doi: 10.1128/JCM.00197-18.
36. Van Twest R, Kropinski AM. Bacteriophage enrichment from water and soil. *Methods Mol Biol*. 2009; 501: 15-21. doi: 10.1007/978-1-60327-164-6_2.
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. doi: 10.1007/978-1-60327-164-6_7.

38. Kutter E. Phage host range and efficiency of plating. In: E Kutter, A Sulakvelidze, eds. *Bacteriophages: Biology and Applications*. 2005. CRC Press: Boca Raton, Florida. pp. 141-149.
39. Andrews S. FastQC: a quality control tool for high throughput sequence data. 2010. Available from: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/. (Accessed: 08 Jan 2021).
40. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*. 2018; 34(17): i884-i890. doi: 10.1093/bioinformatics/bty560.
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(15): 2666-2669. doi: 10.1093/bioinformatics/bty149.
42. Lin Y, Yuan J, Kolmogorov M, Shen MW, Chaisson M, Pevzner PA. Assembly of long error-prone reads using de Bruijn graphs. *Proc Natl Acad Sci USA*. 2016; 113(52): E8396-E8405. doi: 10.1073/pnas.1604560113.
43. Wick R, Holt K. Benchmarking of long-read assemblers for prokaryote whole genome sequencing. *F1000Research*. 2021; 8(2138). doi: 10.12688/f1000research.21782.4.
44. Wick RR, Judd LM, Cerdeira LT, Hawkey J, Méric G, Vezina B *et al.* Trycycler: consensus long-read assemblies for bacterial genomes. *Genome Biol*. 2021; 22(1): 266. doi: 10.1186/s13059-021-02483-z.
45. Chen Y, Zhang Y, Wang AY, Gao M, Chong Z. Accurate long-read *de novo* assembly evaluation with Inspector. *Genome Biology*. 2021; 22(1): 312. doi: 10.1186/s13059-021-02527-4.
46. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res*. 2017; 27(5): 722-736. doi: 10.1101/gr.215087.116.
47. Wick RR. Porechop. 2018. Available from: https://github.com/rrwick/Porechop. (Accessed: 28 Oct 2022).
48. Wright C, Wykes M. Medaka. 2020. Available from: https://github.com/nanoporetech/medaka. (Accessed: 22 Sep 2022).
49. Wick RR, Holt KE. Polypolish: short-read polishing of long-read bacterial genome assemblies. *PLoS Comput Biol*. 2022; 18(1): e1009802. doi: 10.1371/journal.pcbi.1009802.
50. Zimin AV, Salzberg SL. The genome polishing tool POLCA makes fast and accurate corrections in genome assemblies. *PLoS Comput Biol*. 2020; 16(6): e1007981. doi: 10.1371/journal.pcbi.1007981.
51. Seemann T. Shovill. 2018. Available from: https://github.com/tseemann/shovill. (Accessed: 12 Jan 2021).
52. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol*. 2012; 19(5): 455-477. doi: 10.1089/cmb.2012.0021.
53. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S *et al.* Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One*. 2014; 9(11): e112963. doi: 10.1371/journal.pone.0112963.
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(6): e1005595. doi: 10.1371/journal.pcbi.1005595.
55. Garneau JR, Depardieu F, Fortier LC, 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(1): 8292. doi: 10.1038/s41598-017-07910-5.
56. Benes V, Kilger C, Voss H, Pääbo S, Ansorge W. Direct primer walking on P1 plasmid DNA. *BioTechniques*. 1997; 23(1): 98-100.
57. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. *Genome Biol*. 2019; 20(1): 257-269. doi: 10.1186/s13059-019-1891-0.

58. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods*. 2012; 9(4): 357-359. doi: 10.1038/nmeth.1923.
59. Thorvaldsdottir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform*. 2013; 14(2): 178-192. doi: 10.1093/bib/bbs017.
60. Grubaugh ND, Gangavarapu K, Quick J, Matteson NL, De Jesus JG, Main BJ *et al.* An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar. *Genome Biol*. 2019; 20(1): 8. doi: 10.1186/s13059-018-1618-7.
61. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*. 2009; 25(14): 1754-1760. doi: 10.1093/bioinformatics/btp324.
62. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. [Pre-print]. 2013. Doi: 10.48550/ARXIV.1303.3997.
63. Okonechnikov K, Golosova O, Fursov M, team tU. Unipro UGENE: a unified bioinformatics toolkit. *Bioinformatics*. 2012; 28(8): 1166-1167.
64. Bouras G, Nepal R, Houtak G, Psaltis AJ, Wormald P-J, Vreugde S. Pharokka: a fast scalable bacteriophage annotation tool. *Bioinformatics*. 2022; 39(1). doi: 10.1093/bioinformatics/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(22): 4537-4542. doi: 10.1093/bioinformatics/btz265.
66. Hsieh PF, Lin HH, Lin TL, Chen YY, Wang JT. Two T7-like bacteriophages, K5-2 and K5-4, each encodes two capsule depolymerases: isolation and functional characterization. *Sci Rep*. 2017; 7(1): 4624. doi: 10.1038/s41598-017-04644-2.
67. Whitfield C, Lam M. Characterisation of coliphage K30, a bacteriophage specific for *Escherichia coli* capsular serotype K30. *FEMS Microbiol Lett*. 1986; 37(3): 351-355. doi: 10.1111/j.1574-6968.1986.tb01823.x.
68. Teng T, Li Q, Liu Z, Li X, Liu Z, Liu H *et al.* Characterization and genome analysis of novel *Klebsiella* phage Henu1 with lytic activity against clinical strains of *Klebsiella pneumoniae*. *Arch Virol*. 2019; 164(9): 2389-2393. doi: 10.1007/s00705-019-04321-x.
69. Rudolph C, Freund-Mölbert E, Stirm S. Fragments of *Klebsiella* bacteriophage no. 11. *Virology*. 1975; 64(1): 236-246. doi: 10.1016/0042-6822(75)90095-1.
70. Thiry D, Passet V, Danis-Wlodarczyk K, Lood C, Wagemans J, De Sordi L *et al.* New bacteriophages against emerging lineages ST23 and ST258 of *Klebsiella pneumoniae* and efficacy assessment in *Galleria mellonella* larvae. *Viruses*. 2019; 11(5): E411. doi: 10.3390/v11050411.
71. Wu Y, Wang R, Xu M, Liu Y, Zhu X, Qiu J *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. doi: 10.3389/fmicb.2019.02768.
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). doi: 10.1128/genomeA.00919-17.
73. Liu Y, Leung SSY, Huang Y, Guo Y, Jiang N, Li P *et al.* Identification of two depolymerases from phage IME205 and their antivirulent functions on K47 capsule of *Klebsiella pneumoniae*. *Front Microbiol*. 2020; 11: 218. doi: 10.3389/fmicb.2020.00218.
74. Kwon HJ, Cho SH, Kim TE, Won YJ, Jeong J, Park SC *et al.* Characterization of a T7-like lytic bacteriophage (phiSG-JL2) of *Salmonella enterica* serovar Gallinarum biovar Gallinarum. *Appl Environ Microbiol*. 2008; 74(22): 6970-6979. doi: 10.1128/aem.01088-08.
75. Hamdi S, Rousseau GM, Labrie SJ, Kourda RS, Tremblay DM, Moineau S *et al.* Characterization of five *Podoviridae* phages infecting *Citrobacter freundii*. *Front Microbiol*. 2016; 7: 1023. doi: 10.3389/fmicb.2016.01023.

76. Bleriot I, Blasco L, Pacios O, Fernández-García L, Ambroa A, López M *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(1): 4488. doi: 10.1038/s41598-022-08111-5.
77. Dunn JJ, Studier FW. Complete nucleotide sequence of bacteriophage T7 DNA and the locations of T7 genetic elements. *J Mol Biol*. 1983; 166(4): 477-535. doi: 10.1016/s0022-2836(83)80282-4.
78. Dobbins AT, George M, Jr., Basham DA, Ford ME, Houtz JM, Pedulla ML *et al.* Complete genomic sequence of the virulent *Salmonella* bacteriophage SP6. *J Bacteriol*. 2004; 186(7): 1933-1944. doi: 10.1128/jb.186.7.1933-1944.2004.
79. McGinnis S, Madden TL. BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res*. 2004; 32(Web server issue): W20-W25. doi: 10.1093/nar/gkh435.
80. Darling AE, Mau B, Perna NT. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS One*. 2010; 5(6): e11147. doi: 10.1371/journal.pone.0011147.
81. Moraru C, Varsani A, Kropinski AM. VIRIDIC - a novel tool to calculate the intergenomic similarities of prokaryote-infecting viruses. *Viruses*. 2020; 12(11): 1268. doi: 10.3390/v12111268.
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(6): 1547-1549. doi: 10.1093/molbev/msy096.
83. Letunic I, Bork P. Interactive tree of life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res*. 2021. doi: 10.1093/nar/gkab301.
84. Gilchrist CLM, Chooi Y-H. Clinker & clustermap.js: automatic generation of gene cluster comparison figures. *Bioinformatics*. 2021; 37(16): 2473-2475. doi: 10.1093/bioinformatics/btab007.
85. Nobrega FL, Vlot M, de Jonge PA, Dreesens LL, Beaumont HJE, Lavigne R *et al.* Targeting mechanisms of tailed bacteriophages. *Nat Rev Microbiol*. 2018; 16(12): 760-773. doi: 10.1038/s41579-018-0070-8.
86. González-García VA, Pulido-Cid M, Garcia-Doval C, Bocanegra R, van Raaij MJ, Martín-Benito J *et al.* Conformational changes leading to T7 DNA delivery upon interaction with the bacterial receptor. *J Biol Chem*. 2015; 290(16): 10038-10044. doi: 10.1074/jbc.M114.614222.
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(8): 3103-3119. doi: 10.1007/s00253-017-8224-6.
88. Eskenazi A, Lood C, Wubbolts J, Hites M, Balarjishvili N, Leshkasheli L *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(1): 302. doi: 10.1038/s41467-021-27656-z.
89. Piel D, Bruto M, Labreuche Y, Blanquart F, Goudenège D, Barcia-Cruz R *et al.* Phage-host coevolution in natural populations. *Nat Microbiol*. 2022; 7(7): 1075-1086. doi: 10.1038/s41564-022-01157-1.
90. Turner D, Adriaenssens EM, Tolstoy I, Kropinski AM. Phage annotation guide: guidelines for assembly and high-quality annotation. *Phage (New Rochelle)*. 2021; 2(4): 170-182. doi: 10.1089/phage.2021.0013.
91. Barylski J, Enault F, Dutilh BE, Schuller MB, Edwards RA, Gillis A *et al.* Analysis of Spounaviruses as a case study for the overdue reclassification of tailed phages. *Syst Biol*. 2020; 69(1): 110-123. doi: 10.1093/sysbio/syz036.
92. Pan YJ, Lin TL, Chen CT, Chen YY, Hsieh PF, Hsu CR *et al.* Genetic analysis of capsular polysaccharide synthesis gene clusters in 79 capsular types of *Klebsiella* spp. *Sci Rep*. 2015; 5: 15573. doi: 10.1038/srep15573.

93. Summer EJ, Berry J, Tran TAT, Niu L, Struck DK, Young R. Rz/Rz1 lysis gene equivalents in phages of Gram-negative hosts. *J Mol Biol*. 2007; 373(5): 1098-1112. doi: 10.1016/j.jmb.2007.08.045.
94. 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(2): 341-351. doi: 10.1111/j.1365-2958.2008.06408.x.
95. Kongari R, Rajaure M, Cahill J, Rasche E, Mijalis E, Berry J *et al.* Phage spanins: diversity, topological dynamics and gene convergence. *BMC Bioinformatics*. 2018; 19(1): 326. doi: 10.1186/s12859-018-2342-8.