MICROBIAL GENOMICS

RESEARCH ARTICLE

Papudeshi *et al.*, *Microbial Genomics* 2023;9:001100

DOI 10.1099/mgen.0.001100

MICROBIOLOGY SOCIETY

OPEN DATA OPEN ACCESS

Host interactions of novel *Crassvirales* species belonging to multiple families infecting bacterial host, *Bacteroides cellulosilyticus* WH2

Bhavya Papudeshi[1,*], Alejandro A. Vega[2,3], Cole Souza[2], Sarah K. Giles[1], Vijini Mallawaarachchi[1], Michael J. Roach[1], Michelle An[2], Nicole Jacobson[2], Katelyn McNair[4], Maria Fernanda Mora[2], Karina Pastrana[2], Lance Boling[2], Christopher Leigh[5], Clarice Harker[1], Will S. Plewa[1], Susanna R. Grigson[1], George Bouras[6], Przemysław Decewicz[1,7], Antoni Luque[4,8,†], Lindsay Droit[9], Scott A. Handley[9], David Wang[9], Anca M. Segall[2], Elizabeth A. Dinsdale[1] and Robert A. Edwards[1]

Abstract

Bacteroides, the prominent bacteria in the human gut, play a crucial role in degrading complex polysaccharides. Their abundance is influenced by phages belonging to the *Crassvirales* order. Despite identifying over 600 *Crassvirales* genomes computationally, only few have been successfully isolated. Continued efforts in isolation of more *Crassvirales* genomes can provide insights into phage-host-evolution and infection mechanisms. We focused on wastewater samples, as potential sources of phages infecting various *Bacteroides* hosts. Sequencing, assembly, and characterization of isolated phages revealed 14 complete genomes belonging to three novel *Crassvirales* species infecting *Bacteroides cellulosilyticus* WH2. These species, *Kehishuvirus* sp. 'tikkala' strain Bc01, *Kolpuevirus* sp. 'frurule' strain Bc03, and 'Rudgehvirus jaberico' strain Bc11, spanned two families, and three genera, displaying a broad range of virion productions. Upon testing all successfully cultured *Crassvirales* species and their respective bacterial hosts, we discovered that they do not exhibit co-evolutionary patterns with their bacterial hosts. Furthermore, we observed variations in gene similarity, with greater shared similarity observed within genera. However, despite belonging to different genera, the three novel species shared a unique structural gene that encodes the tail spike protein. When investigating the relationship between this gene and host interaction, we discovered evidence of purifying selection, indicating its functional importance. Moreover, our analysis demonstrated that this tail spike protein binds to the TonB-dependent receptors present on the bacterial host surface. Combining these observations, our findings provide insights into phage-host interactions and present three *Crassvirales* species as an ideal system for controlled infectivity experiments on one of the most dominant members of the human enteric virome.

DATA SUMMARY

The genomes used in this research are available on Sequence Read Archive (SRA) within the project, PRJNA737576. *Bacteroides cellulosilyticus* WH2, *Kehishuvirus* sp. 'tikkala' strain Bc01, *Kolpuevirus* sp. 'frurule' strain Bc03, and 'Rudgehvirus jaberico' strain

Received 14 March 2023; Accepted 10 August 2023; Published 04 September 2023

Author affiliations: [1]Flinders Accelerator for Microbiome Exploration, College of Science and Engineering, Flinders University, Bedford Park, Adelaide SA, 5042, Australia; [2]Department of Biology, San Diego State University, 5500 Campanile Drive, San Diego, CA, 92182, USA; [3]David Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA, USA; [4]Computational Science Research Center, San Diego State University, 5500 Campanile Drive, San Diego, CA, 992182, USA; [5]Adelaide Microscopy, University of Adelaide, Adelaide, SA, 5005, Australia; [6]Adelaide Medical School, Faculty of Health and Medical Sciences, The University of Adelaide, Adelaide, SA, 5005, Australia; [7]Department of Environmental Microbiology and Biotechnology, Institute of Microbiology, Faculty of Biology, University of Warsaw, Miecznikowa 1, Warsaw, 02-096, Poland; [8]Department of Mathematics and Statistics, San Diego State University, 5500 Campanile Drive, San Diego, CA, 992182, USA; [9]Department of Pathology & Immunology, Washington University School of Medicine, St. Louis, MO, 63110, USA.

***Correspondence:** Bhavya Papudeshi, nala0006@flinders.edu.au

Keywords: *Crassvirales*; wastewater; phage-host-interaction; tail spike protein; TonB-depdenent receptors; purifying selection; co-evolution.

Abbreviations: ANI, average nucleotide identity; BHISMg, Brain Heart Infusion media supplemeted with Magnesium sulphate and Magnesium chloride; MCP, major capsid protein; MSA, multiple sequence alignment; PFU, plaque forming units.

†Present address: Department of Biology, University of Miami, Coral Gables, Florida, USA.

Data statement: All supporting data, code and protocols have been provided within the article or through supplementary data files.Five supplementary figures and five supplementary tables are available with the online version of this article.

001100 © 2023 The Authors

This is an open-access article distributed under the terms of the Creative Commons Attribution License. This article was made open access via a Publish and Read agreement between the Microbiology Society and the corresponding author's institution.

Impact Statement

Bacteriophages play a crucial role in shaping microbial communities within the human gut. Among the most dominant bacteriophages in the human gut microbiome are *Crassvirales* phages, which infect Bacteroides. Despite being widely distributed, only a few *Crassvirales* genomes have been isolated, leading to a limited understanding of their biology, ecology, and evolution. This study isolated and characterized three novel *Crassvirales* genomes belonging to two different families, and three genera, but infecting one bacterial host, *Bacteroides cellulosilyticus* WH2. Notably, the observation confirmed the phages are not co-evolving with their bacterial hosts, rather have a shared ability to exploit similar features in their bacterial host. Additionally, the identification of a critical viral protein undergoing purifying selection and interacting with the bacterial receptors opens doors to targeted therapies against bacterial infections. Given *Bacteroides* role in polysaccharide degradation in the human gut, our findings advance our understanding of the phage-host interactions and could have important implications for the development of phage-based therapies. These discoveries may hold implications for improving gut health and metabolism to support overall well-being.

Bc11 are all available on GenBank with accessions NZ_CP072251.1 (*B. cellulosilyticus* WH2), OQ198717.1 (Bc01), OQ198718.1 (Bc03), and OQ198719.1 (Bc11), and we are working on making the strains available through ATCC. The 3D protein structures for the three *Crassvirales* genomes are available to download at doi.org/10.25451/flinders.21946034.

INTRODUCTION

The intricate relationship between gut microbiomes and human health is characterized by the diverse microbial communities that help with digestion, regulate the immune system, and alter brain function [1–3]. Metagenomics, a culture-independent technique is used to capture microbial diversity in a sample [4, 5], has transformed our understanding of bacteria and the corresponding bacteriophages in the environment [6–9]. These metagenomic datasets have revealed a correlation between bacterial and bacteriophage populations, which suggests bacteriophages play a role in modulating bacterial populations [10, 11]. In particular, the human gut microbiome exhibits varying bacterial densities, including a high abundance of Bacteroidota (formerly Bacteroidetes) [12–14], which are shaped by the phages belonging to the *Crassvirales* order. These dsDNA bacteriophages have a podovirus-like morphology, genomes ranging between 100 and 200 kb, and conserved gene order [15–17]. They are widespread, constituting a stable component of an individual's microbiome, and do not appear to be associated with human health or disease states [16, 18].

The first phage within *Crassvirales* order was computationally discovered by cross-assembly of DNA sequence reads from human gut microbiome samples [19]. Since, nearly 600 *Crassvirales* genomes have been identified computationally, leading to the International Committee of Taxonomy of Viruses (ICTV) formally classifying the *Crassvirales* order into four families, ten subfamilies, 42 new genera, and 72 new species [17, 20]. The classification relied on phylogenetic analysis of conserved structural genes, including the major capsid protein (MCP), terminase large subunit (*terL*), and portal protein (portal). Additionally, the average nucleotide identity (ANI) species cutoff was set to 95% identity over 85% genome coverage.

The identification of numerous *Crassvirales* genomes has advanced our understanding of this viral order. Similar to other phages, these genomes contain three discernible regions encoding for 1) structural proteins involved in producing the capsid and tail genes, 2) transcription proteins and 3) replication proteins crucial for successful phage replication in different infection stages [15]. Gene homology analysis showed that majority of the genes are highly variable when compared with other genomes from this order. Comparative genomics further displayed unique biological characteristics of *Crassvirales* species [21, 22], including switching DNA polymerases, alternative coding strategies [23–26], and the variable intron density across lineages [15, 26]. Overall, functional annotation of *Crassvirales* genomes remain challenging, with majority of the genes annotated as hypothetical proteins lacking known biological function, with little to no similarity with sequences in reference databases. These challenges can be addressed through experimental approaches that can help elucidate the functions of these uncharacterized genes or proteins.

The first step in experimental approaches is phage isolation, which requires the knowledge of their host species, and the ability to culture them. This has led to only four successful isolates obtained so far, including *Kehishuvirus primarius* (crAss001) infecting *Bacteroides intestinalis* APC919/174 [27], *Wulfhauvirus bangladeshii* DAC15 and DAC17 from wastewater effluent infecting *Bacteroides thetaiotaomicron* VPI-5482 [28], and *Jahgtovirus secundus* (crAss002) infecting *Bacteroides xylanisolvens* APCS1/XY [29]. All these isolates exhibited host specialist morphotypes that can be maintained in the continuous host culture, but none possess lysogeny-related genes [27, 30]. The proposed mechanism of persistence includes where the bacterial host cycles between sensitive and resistant states through altering the genes encoding surface transporters and capsular polysaccharide structures (CPS) on the bacterial surface [30, 31]. Further to improve the annotations, cryogenic-electron microscopy of *K. primarius* provided functional assignments to the virion proteins, and an insight into the infection mechanism, revealing how the capsid

and tail store cargo proteins aid in initial host infection [32]. Continued efforts in isolation of more *Crassvirales* genomes can provide insights into phage-host-evolution, comprehensive protein annotation and elucidation of infection mechanisms.

Here we present the successful isolation of 14 *Crassvirales* isolates from wastewater, that include three novel species, belonging to families *Steigviridae* and *Intestiviridae*. Remarkably, all these isolates infect the same host, *Bacteroides cellulosilyticus* WH2. We investigate the genes playing a role in host interaction, providing insights into the evolution of these dominant phages, and how their interactions shape the gut microbiome.

METHODS

Phage sampling

Untreated sewage water (influent) was collected from a waste treatment plant in Cardiff, CA in one litre Nalgene bottles. An aliquot of 30 ml influent was centrifuged at 5000 RCF for 5 min to pellet the debris. The supernatant was decanted and passed through a 0.22 µm pore size Sterivex filter. The filtrate was used as a phage source and stored between 2 to 8 °C.

Host bacteria cultivation

Bacterial species, *B. cellulosilyticus* WH2 [33] received as glycerol stocks from Washington University, St. Louis, *B. fragilis* NCTC 9343 (ATCC 25285), *B. stercoris* CC31F (ATCC 43183), and *B. uniformis* (ATCC 8492) were received as glycerol stocks from BEI resources were used as bacterial hosts. All the bacteria were grown in brain-heart infusion media supplemented with 2 mM $MgSO_4$, and 10 mM $MgCl_2$ we denote as BHISMg. Culture plates were supplemented with 1.5% w/v agar and incubated at 37 °C for 48 h under anaerobic conditions with 5% H_2, 5% CO_2, and 90% N_2. Following incubation, an isolated colony was transferred into a 12 h deoxygenated BHISMg broth. Following anaerobic incubation at 37 °C for 24 h the liquid cultures were further sub-cultured into another BHISMg broth and incubated overnight.

Plaque assays

BHISMg plates were deoxygenated for 12 h in the anaerobic chamber and pre-warmed before use. For top agar plates were prepared by adding cooled molten BHISMg with 0.7 % w/v agar was inoculated with 500 µl of bacteria, and between 2 µl and 50 µl of processed phage influent. The plates were cooled for 15 min before incubating at 37 °C for up to 5 days. Plates were assessed daily for the development of plaques.

Lysate preparation

Plaque from each plate was inoculated into 200 µl of SM buffer and homogenized to diffuse the phage from the agar to the buffer. A 200 µl aliquot of the phage was added to *B. cellulosilyticus* WH2 bacteria in the log-growth phase and grown at 37 °C anaerobically, overnight. The tubes containing the bacteria and phage were manually shaken every 30 min for the first 3 h of incubation. Post incubation, tubes were centrifuged at 4500 ***g*** for 5 min, and the supernatant was collected and concentrated using a 50000 Da MWCO Vivaspin ultrafiltration unit (Sartorius). Phage lysate was stored at 4 °C.

Phage titering enumeration

Phage titre were enumerated using the molten agar overlay method described above. A 200 µl aliquot of the lysate was diluted ten-fold in sterile SM buffer, and 10 µl was spotted onto a BHISMg plate. The plates were incubated for 24–48 h at 37 °C. After incubation, the plates were analysed by counting the plaques obtained to determining the titre.

Viral DNA extraction and sequencing

Phage DNA was extracted using a Promega Wizard DNA clean-up system (Catalog #A7280) using the method outlined by Summer et al [34]. In short, 15 ml of phage lysate was DNase and RNase treated, lysed, and treated with Proteinase K. The sample was added to a Wizard minicolumn and washed. DNA was eluted in 100 µl of preheated water at 80 °C. The DNA obtained was quantified using a Qubit 1 x dsDNA high-sensitivity assay kit (Invitrogen, Life Technologies) per manufacturer's instructions. Oxford Nanopore MinION sequencing was undertaken the manufacturer's instructions. In short, a maximum of 400 ng of sample DNA was used for the library preparation using Oxford Nanopore Rapid Barcoding Sequencing Kit (SQK-RBK0004); samples were barcoded, pooled and cleaned. The pooled samples were loaded and run on a Flowcell R9.4.1 (FLO-MIN106) following the manufacturer's instructions. The Illumina sequencing libraries were prepared by extracting the total nucleic acid (RNA and DNA) using the COBAS AmpliPrep instrument (Roche), with NEBNext library construction and sequenced on Illumina MiSeq using the paired-end 2×250 bp protocol as described in [35]. The sequencing data were deposited to Sequence Read Archive in Bioproject, PRJNA737576.

For the Nanopore sequenced isolates, basecalling was performed with Guppy v6.0.1 with model dna_r9.4.1_450bps_hac. The reads were then processed with Filtlong v0.2.20 [36] to remove reads less than 1000 bp in length and exclude 5% of the lowest-quality

reads. Similarly, Illumina sequences were processed with prinseq++ v0.0.20.4 [37], filtering reads less than 60 bp in length, reads with quality scores less than 25, and exact duplicates.

Genome assembly

To assemble the genomes, a pipeline based on Snakemake using Snaketool [38] was employed. Nanopore reads were assembled using Flye v2.9 [39], while Illumina reads were assembled using MEGAHIT v1.2.9 [40]. These assemblers were selected as they provide assembly graphs, which are useful for completing fragmented genome assemblies [41–44].

To evaluate assembly quality, the resulting contigs were processed with ViralVerify v1.1 to detect viral contigs [45], read coverage was calculated using CoverM v0.6.1 [46], and the assembly graph was examined. The assembly graph provides information on connecting unitigs (high-quality contigs) representing the longest non-branching paths joined together to form contigs.

From each assembly, unitigs meeting specific criteria were selected as complete phage genomes. These unitigs had a length greater than 90 kb, were identified as viral, exhibited the highest read coverage and were classified as complete using CheckV v1.0.1. To ensure representation, one unitg per sample was selected as the complete phage assembly. In the end, the assemblies were polished with high-coverage Illumina reads using Polca to reduce sequencing-related errors [47].

Among the 41 phage genomes, 14 phages infecting *B. cellulosilyticus* were identified as belonging to *Crassvirales* order. These genomes were approximately 90 to 120 kbp in length and aligned against known *Crassvirales* genomes. Among these phages, eight samples were sequenced on both Nanopore and Illumina sequencing platforms (Bc01 to Bc03, Bc05 to Bc08), while four were sequenced only on Nanopore platforms (Bc09 to Bc11), and the remaining four were sequenced only on Illumina platforms (Bc04, Bc12 to Bc14).

Taxonomic and functional annotation

The isolates in this study were processed with CrassUS [48], a specialized tool for annotating *Crassvirales* genomes, providing taxonomic, functional annotation along with direct terminal repeats (DTR), and average nucleotide identities (ANI) of similar reference genomes. Taxonomic annotations from CrassUS followed ICTV [20, 22] criteria for *Crassvirales* order demarcation. Phylogenetic trees were constructed using conserved genes (MCP, portal, *terL*) with MAFFT v7.49 [49] for alignment, followed by trimal v1.4.1 for trimming, and FastTree v2.1.10 [50] for inference. Trees were built using JTT model, CAT approximation with 20 rate categories, and visualized using iTol [51].

To compare the predicted genes and their arrangement across species, clinker plots [52] were used after re-circularizing the genes to start at the *terL*. This allowed for the examination of synteny across genomes. Additionally, tRNA genes encoded by the phages, which evade host translation machinery, were predicted with tRNA-scanSE [53].

Phage-host co-phylogenetic analysis

To determine if the phage co-evolve with bacterial hosts, we performed a cophylogenetic analysis using Parafit [54] via the ape v5. 6-2 R package. The distance matrix of the trimmed multiple sequence alignment (MSA) using MAFFT v7.520 [49] and trimal v1.4.1 of the seven *Crassvirales* species ('*K.* tikkala' Bc01, '*K.* frurule' Bc03, 'R. jaberico' Bc11, *K. primarius, J. secundus, W. bangladeshii* DAC15, *W. bangladeshii* DAC17) portal gene, was generated using EMBOSS distmat v6.6.0. These steps were repeated for the associated bacterial hosts, *B. cellulosilyticus, B. intestinalis, B. thetaiotomicron* and *B. xylanisolvens*. The two distance matrices were compared using Parafit [54] with 1000 permutations, and Cailliez eigen value correction. The trimmed multiple sequence alignment was used to generate the two phylogenetic trees using FastTree v2.1.10 using JTT model, CAT approximation with 20 rate categories, and visualized using iTol [51].

Transmission electron microscopy imaging

Crassvirales phages were grown using the phage overlay method described above. To prepare the phage lysates, they were diluted 1:10, and 5 µl of the diluted phage lysate was applied to a plasma-cleaned grid for 2 min at room temperature. The grids used were formvar and carbon coated 200 mesh grids and they were plasma cleaned using the Gatan (Solarus) Advanced plasma system for 30 s prior to use. The excess phage lysate sample was wicked off with Whatman filter paper and the grid was washed with 5 µl of water. The sample was negatively stained with 5 µl of the 2 % w/v uranyl acetate for 1 min. The excess stain was wicked off with filter paper to dry the sample on the grid. The grid was then imaged using a Tecnai G2 Spirit TEM operated at 120 kV at a magnification of 49000× and the images were recorded on an AMT Nanosprint 15 digital camera using software v7.0.1.

Phage measurements were conducted using the ImageJ v1.53t software [55]. The capsid diameter was calculated by measuring the diameter of the circle circumscribing the capsid, such that the more distant vertices of the projected capsid contacted the circle (Fig. S1). The length of the tail was calculated from the base of the capsid to the end of the visible tail, including the collar section of the phage structure. Tail fibres or appendages was calculated (Fig. S1). Average measurements from five phages were

calculated and reported. The TEM image was further edited for publication using the GNU Image Manipulation Programme (GIMP) v2.10 [56].

The packing genome density was predicted by correcting the measured radius from the expected capsid thickness and calculating the internal volume assuming an icosahedral model inferred from prior tailed phage capsid studies [57, 58]. The results can be reproduced using the Colab notebook available at link, https://shorturl.at/enAKS.

Evolutionary analyses

The 14 *Crassvirales* isolated and assembled genomes in this study and the four reference pure culture isolates, *K. primarius*, *J. secundus, W. bangladeshii* DAC15 and DAC17 were assessed together for this analysis. Orthologous genes were identified from genes predicted from the above 18 genomes, using Orthofinder v2.5.4 default settings to determine signatures for host interactions. The default settings use diamond for sequence search, MAFFT for alignment, FastTree for tree inference and STAG species tree method [59]. Orthogroups that included genes present only in phages from the host *B. cellulosilyticus* WH2 were examined further.

These orthogroups were aligned using Muscle [60] codon-based multiple sequence alignment in MEGA v11.0 [61]. To test for codon-based positive selection, we calculated the probability synonymous (d_S) and non-synonymous (d_N) mutations occuring. The null hypothesis of strict neutrality ($d_N=d_S$) was rejected, and in favour of the alternate hypothesis ($d_N>d_S$). The d_N/d_S values were calculated from the MSA using MEGA v11.0 [62], with the Li-Wu-Luo method [63]. The variance of the difference was computed using bootstraps, set to 100 replicates. As this analysis can be misleading in the presence of recombination breakpoints, orthogroups were run through Genetic Algorithm for Recombination Detection (GARD) analysis [64], with default settings. This method utilizes a combination of phylogenetic and statistical approaches to detect recombination signals.

Predicting proteins 3D structure and docking

The 3D structures of the proteins from '*K.* tikkala' strain Bc01, '*K.* frurule' strain Bc03, and 'R. jaberico' strain Bc11 were predicted using Colabfold v1.4.0 [65] on the Gadi server at the National Computational Infrastructure (NCI). To determine structural similarity, the protein structures were run through pairwise structure alignment using the Flexible structure AlignmenT by Chaining Aligned fragment pairs allowing Twists (FATCAT) that allows for flexible protein structure comparison [66, 67].

To predict the phage protein interaction with the bacterial host, the previously predicted 3D protein structures of all the proteins for *B. cellulosilyticus* WH2 were downloaded from the AlphaFold Protein Structure Database via the Google Cloud Platform [68]. All protein pairs were docked using hdock-lite v1.1 [69] on the Gadi server. The results from hdock were sorted based on the binding score (hdock-scores) in the output file to identify the highest-quality binding predictions for each phage protein. In general, lower HDock-scores are indicative of more favourable or stronger protein-protein interactions, suggesting a higher likelihood of a stable complex formation. Higher HDock-scores, on the other hand, may suggest weaker or less favourable interactions. The 3D structure of the proteins was visualized using Chimera.

RESULTS

Search for *Crassvirales* phages

We obtained a total of 41 phages from wastewater infecting four different *Bacteroides* species, *B. cellulosilyticus* WH2, *B. fragilis* NCTC 9343, *B. stercoris* CC31F, and *B. uniformis* ATCC 8492. The phages were sequenced using Oxford Nanopore or Illumina MiSeq platforms, and the resulting sequences were assembled. We performed BLASTN searches of the assembled phages against the non-redundant nucleotide (nr/nt) NCBI database for taxonomic assignment. As a result, we identified 14 phages infecting *B. cellulosilyticus* WH2 belong to *Crassvirales* order. Each of these phages were labelled with code ranging from Bc01 to Bc14.

Isolation and taxonomic classification of *Crassvirales* isolates

Crassvirales isolates formed distinct clear circular plaques with a uniform diameter of 1 mm on soft agar overlays. We performed shotgun sequencing on the 14 isolates phages, with Bc01 to Bc03, Bc05 to Bc11 sequenced on Oxford Nanopore platform, and Bc01 to Bc08, Bc12 to Bc14 sequenced on the Illumina platform. The assemblies produced multiple contigs, and our selection criteria to identify complete phage genomes was based on presence of viral genes, highest read coverage and unitigs (high quality contig) size of approximately 100 kb (Table 1). This resulted in complete genomes for each of the 14 phages, that were polished with Illumina reads correcting for substitution, insertion, and deletion errors.

For taxonomic classification of these isolates, we applied the ICTV report guidelines for defining taxonomy within *Crassvirales* order. Phylogenetic clustering of the conserved portal gene and average nucleotide identity (ANI) species cutoff (95% identity over 85% genome coverage) identified three distinct clusters (Fig. 1a). We selected the highest confidence genomes: Bc01, Bc03, and Bc11 from each cluster. These three isolates were compared against all known *Crassvirales* genomes through phylogenetic

Table 1. Taxonomic classification of the 14 *Crassvirales* genomes isolated from wastewater infecting *Bacteroides cellulosilyticus* WH2

Phage isolate	Sequencing platform	Genome length (bp)	Taxonomy	Biosample ID
Bc01	MinION, MiSeq	100,722	*Kehishuvirus* sp. 'tikkala' strain Bc01	SAMN20326212
Bc02	MinION, MiSeq	98,905	*Kolpuevirus* sp. 'frurule' strain Bc02	SAMN20326213
Bc03	MinION, MiSeq	99,379	*Kolpuevirus* sp. 'frurule' strain Bc03	SAMN20326214
Bc04	MiSeq	99,033	*Kolpuevirus* sp. 'frurule' strain Bc04	SAMN29929441
Bc05	MinION, MiSeq	97,832	*Kolpuevirus* sp. 'frurule' strain Bc05	SAMN20326216
Bc06	MinION, MiSeq	99,845	*Kolpuevirus* sp. 'frurule' strain Bc06	SAMN20326217
Bc07	MinION, MiSeq	98,518	*Kolpuevirus* sp. 'frurule' strain Bc07	SAMN20326218
Bc08	MinION, MiSeq	98,067	*Kolpuevirus* sp. 'frurule' strain Bc08	SAMN20326219
Bc09	MinION	98,788	*Kolpuevirus* sp. 'frurule' strain Bc09	SAMN20326220
Bc10	MinION	96,329	*Kolpuevirus* sp. 'frurule' strain Bc10	SAMN20326221
Bc11	MinION	90,458	'Rudgehvirus jaberico' strain Bc11	SAMN20326222
Bc12	MiSeq	96,952	*Kolpuevirus* sp. 'frurule' strain Bc12	SAMN29929442
Bc13	MiSeq	90,716	'Rudgehvirus jaberico' strain Bc13	SAMN29929443
Bc14	MiSeq	98,803	*Kehishuvirus* sp. 'tikkala' strain Bc14	SAMN29929444

clustering of conserved genes, major capsid protein (MCP), portal, terminase large subunits (*terL*) genes to determine that Bc01 and Bc03 clusters belong to the *Steigviridae* family, and Bc11 to the *Intestiviridae* family (Figs 1b and S2, available in the online version of this article). Confirmation of the genus assignment was obtained through ANI and shared protein information, which identified Bc01 to *Kehishuvirus*, Bc03 to *Kolpuevirus*, and Bc11 to a novel genus group that we propose to name 'Rudgehvirus'.

All three isolates represent novel species exhibited less than 95% identity and 85% coverage to any known *Crassvirales* genomes. Bc01 is most similar to the reference genome *Kehishuvirus primarius* (Genbank ID: MH675552) with 95.5 % identity across 79.1% genome coverage. Bc03 aligns with *Kolpuevirus hominis* (Genbank ID: MT774391) with 82.8 % identified across 53.73% query coverage. Bc11 aligns with the reference genome *Jahgtovirus intestinalis* (Genbank ID: OGOL01000109) with 74.7% identity across only 9.9% query coverage. We proposed names for these novel species as *Kehishuvirus* sp. 'tikkala' strain Bc01, *Kolpuevirus* sp. 'frurule' strain Bc03, and 'Rudgehvirus jaberico' strain Bc11.

Genome characteristics of the novel *Crassvirales* species

Kehishuvirus sp. 'tikkala' strain Bc01 is 100,841 bp, with 104 proteins, 24 tRNAs and GC content of 35.09% (Table 2) which is lower than the bacterial host GC content of 42.8%. These isolates formed clear, uniform circular spot plaques approximately 1 mm in diameter, forming 9.3×10^9 PFU ml^{-1} (Fig. 2a). Transmission electron microscopy (TEM) revealed they have podovirus-like morphology, displaying polyhedral capsids with a diameter of 94 ± 3 nm and tails with collar structures of 34 ± 3 nm length (Fig. 2b). From the calculated capsid size and genome length, we found this phage packages its DNA at a density of 0.54 bp/nm^3. This genome lacks direct terminal repeat sequences, stop-codon reassignment, and lysogeny-related genes (Table S1).

Kolpuevirus sp. 'frurule' strain Bc03 shares the *Steigviridae* family with '*K.* tikkala' strain Bc01. This genome is 99,523 bp long with GC content of 33%, 108 genes and four tRNA genes encoding arginine, asparagine, and tyrosine (Table 2). Similar to '*K.* tikkala' strain Bc01, this phage also formed clear, uniform circular spot plaques but forming 2.3×10^9 PFU ml^{-1} (Fig. 2a). Displaying a podovirus morphology, the virion was slightly larger than '*K.* tikkala' strain Bc01 with capsids of diameter 97 ± 3 nm, tail with collar structures of 33 ± 3 nm (Fig. 2b), packaging its DNA at a lower density of 0.48 bp/nm^3. Annotation of genes confirmed absence of direct terminal repeats, stop-codon reassignments, and lysogeny-related genes (Table S2). This is the first isolate within its genus.

'Rudgehvirus jaberico' strain Bc11 belongs to *Intestiviridae* family in a novel genus. This genome is 90575 bp long with 29.15% GC content, encoding 84 genes and lacks tRNA genes (Table 2). Unlike the above two species, this isolate formed plaques with a circular halo, indicating depolymerase activity, forming 3.75×10^3 PFU ml^{-1} (Fig. 2a). This isolate's virion was relatively smaller in size compared to the other two isolates, with capsid size diameter of 90 ± 4 nm, tails measuring 25 ± 4 nm (Fig. 2b). Despite the smaller capsid size, and genome length, this phage packages its DNA at a density of 0.56 bp/nm^3, comparable to '*K.* tikkala' strain Bc01. Similar to the other two genomes, direct terminal repeats, stop-codon reassignments, and lysogeny-related genes (Table S3) were absent.

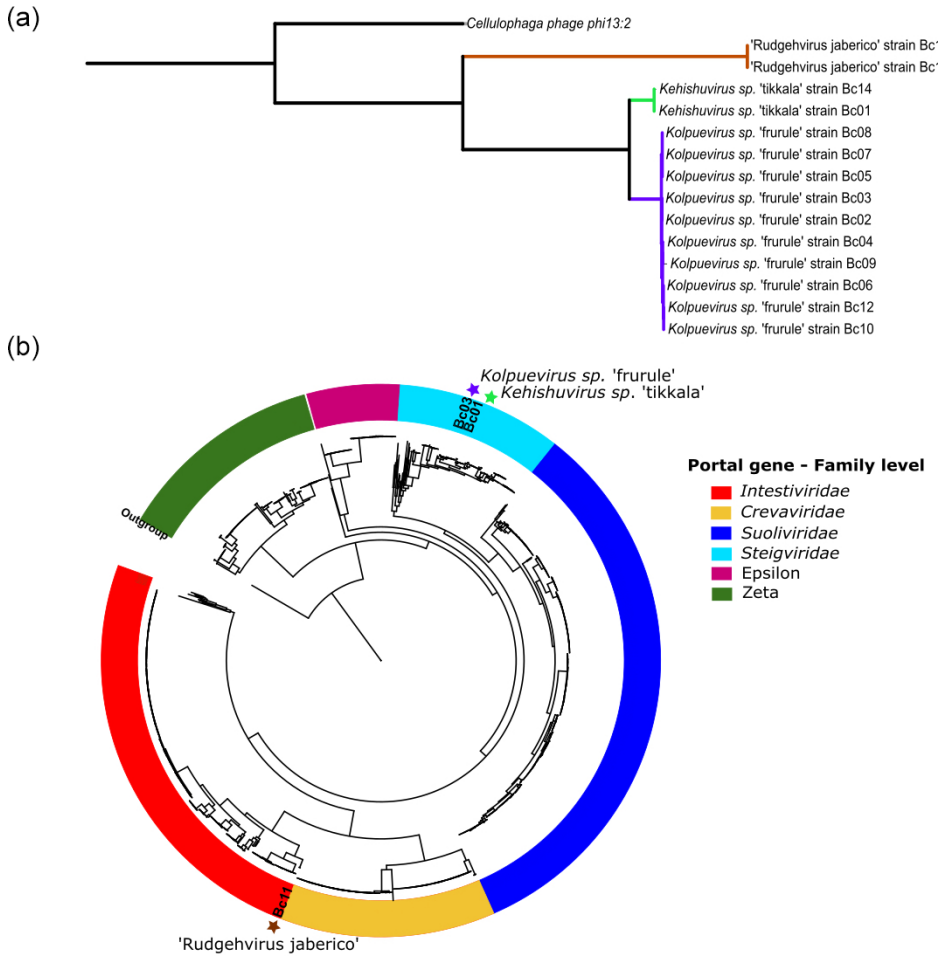

Fig. 1. : Phylogenetic tree constructed using the portal protein using JTT model, CAT approximation with 20 rate categories and outgroup set to Cellulophaga phage phi13:2. (a) Phylogenetic tree of the 14 *Crassvirales* isolates with the branches colour-coded to represent the three species, *Kehishuvirus* in light green, *Kolpuevirus* in purple, and 'Rudgehvirus' in brown. (b) Clustering of all known *Crassvirales* genomes confirming that isolate *Kehishuvirus* sp. 'tikkala' strain Bc01 and *Kolpuevirus* sp. 'frurule' strain Bc03 belong to the family *Steigviridae* (cyan), and 'Rudgehvirus jaberico' strain Bc11 to *Intestiviridae* (red).

Comparative analysis across the three isolates shows that '*K.* tikkala' strain Bc01 and '*K.* frurule' strain Bc03, belonging to the same family (Steigviridae) exhibit higher gene similarity with each other. In contrast, 'R. jaberico' strain Bc11 from *Intestiviridae* family displays distinct gene arrangements (Fig. 2c). Notably, all three genomes share two structural genes encoding tail spike proteins with a domain encoding for polysaccharide degrading enzymes such as glycoside hydrolase. Structural protein one encompassing the '*K.* tikkala' strain Bc01 protein (WEU69744.1) shared 97% sequence similarity with the '*K.* frurule' strain Bc03 protein (WEY17522.1), while collectively these sequences share greater than 39% similarity with 'R. jaberico' strain Bc11 protein (WEU69859.1) (Fig. S3). Similarly, structural protein two encompassing '*K.* tikkala' strain Bc01 protein (WEU69745.1) shared 59% sequence similarity with '*K.* frurule' strain Bc03 protein (WEY17523.1), and together share more than 46% similarity with 'R. jaberico' strain Bc11 protein (WEU69857.1) (Fig. S4).

Table 2. Genome characteristics of the three novel *Crassvirales* species

Genome	Length (bp)	GC %	Coding density	no. of CDS	Unknown function	tRNA	DTR
Kehishuvirus sp. 'tikkala' strain Bc01	100,841	35.09	91.84	104	58	24	False
Kolpuevirus sp. 'frurule' strain Bc03	99,523	33.00	92.06	108	63	5	False
'Rudgehvirus jaberico' strain Bc11	90,575	29.15	87.45	84	48	0	False

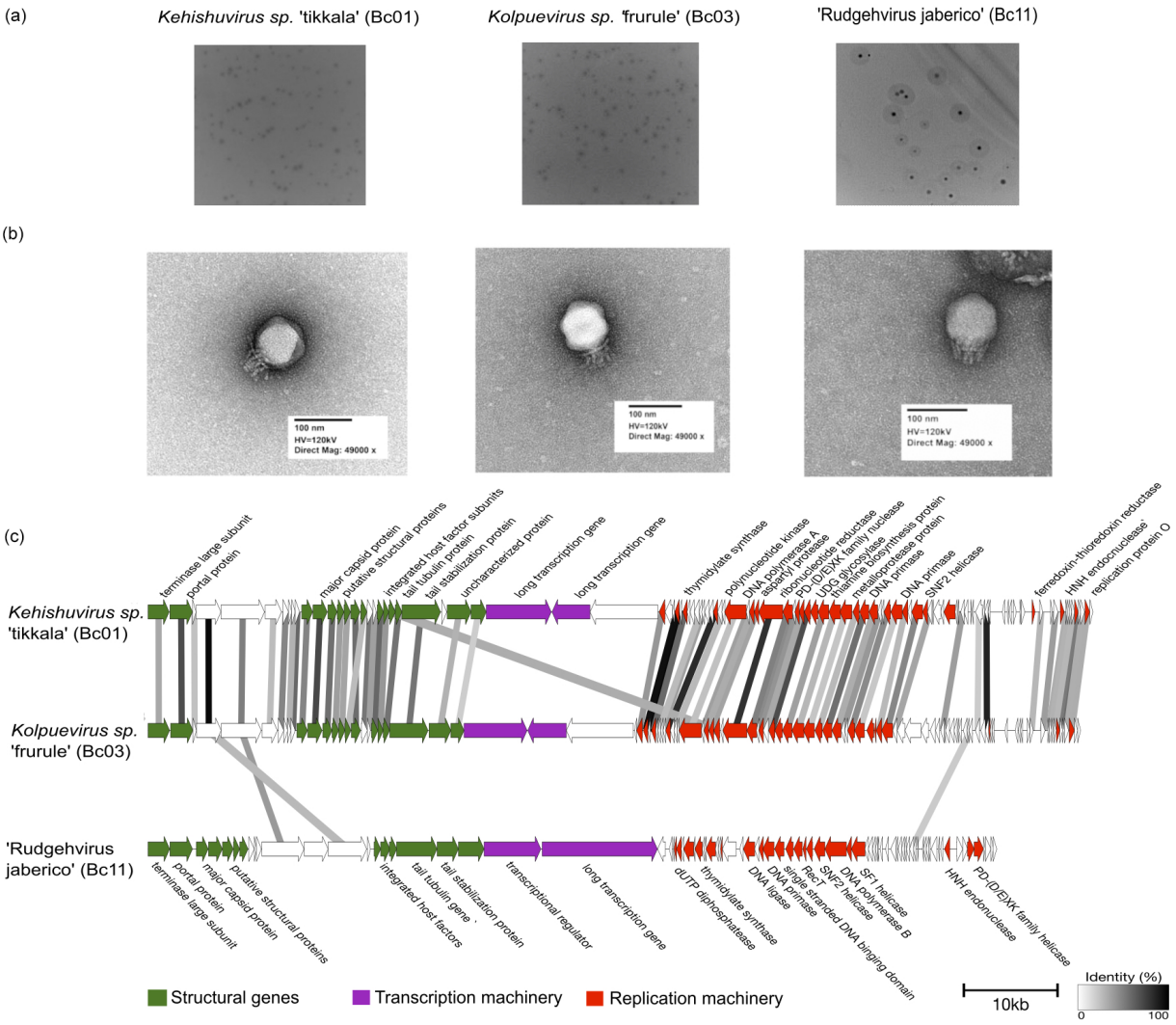

Fig. 2. (a) Plaque morphology of three species, '*K.* tikkala' strain Bc01, '*K.* frurule' strain Bc03, and 'R. jaberico' strain Bc11. (b) Transmission electron microscopy images negatively stained with uranyl acetate of the three isolates. (c) Gene arrangement and functional annotation of the three genomes colour-coded based on their functional modules and hypothetical genes represented in white. The direction of the arrows represents the direction of the gene read from the genome, and the arrows themselves represent individual genes. The links connecting the genes indicate sequence identity, ranging from 30 % (grey) to 100% (black).

Synteny across all seven *Crassvirales* species successfully isolated

The comparison of the three novel species from this study that infect the same bacterial host with the four isolate *Crassvirales* genomes that infect other *Bacteroides* hosts, showed expected gene similarity based on their taxonomic assignment. Among the *Steigviridae* genomes, '*K.* tikkala' strain Bc01 was most similar to *K. primarius* sharing 76 of 106 genes and the two genomes from *Wulfhauvirus* genus (strains DAC15 and DAC17) shared 115 of the 121 genes with greater than 30% similarity. '*K.* frurule' strain Bc03 belonging to a unique genus, *Kolpuevirus* exhibited intermediate similarity, sharing 68 genes with *Kehishuvirus* and 71 genes with *Wulfhauvirus* genus (Fig. 3a). Within the *Intestiviridae* family, 'R. jaberico' strain Bc11 was compared to the *J. secundus,* and they shared 37 genes, including 11 structural genes, three transcription genes, and 23 replication-related genes (Fig. 3b).

The exception to the taxa-based similarity was the two structural genes, encoding tail spike proteins that were shared only among isolates infecting the same host, '*K.* tikkala' strain Bc01, '*K.* frurule' strain Bc03, 'R. jaberico' strain Bc11 despite belonging to different genera (Fig. 2c).

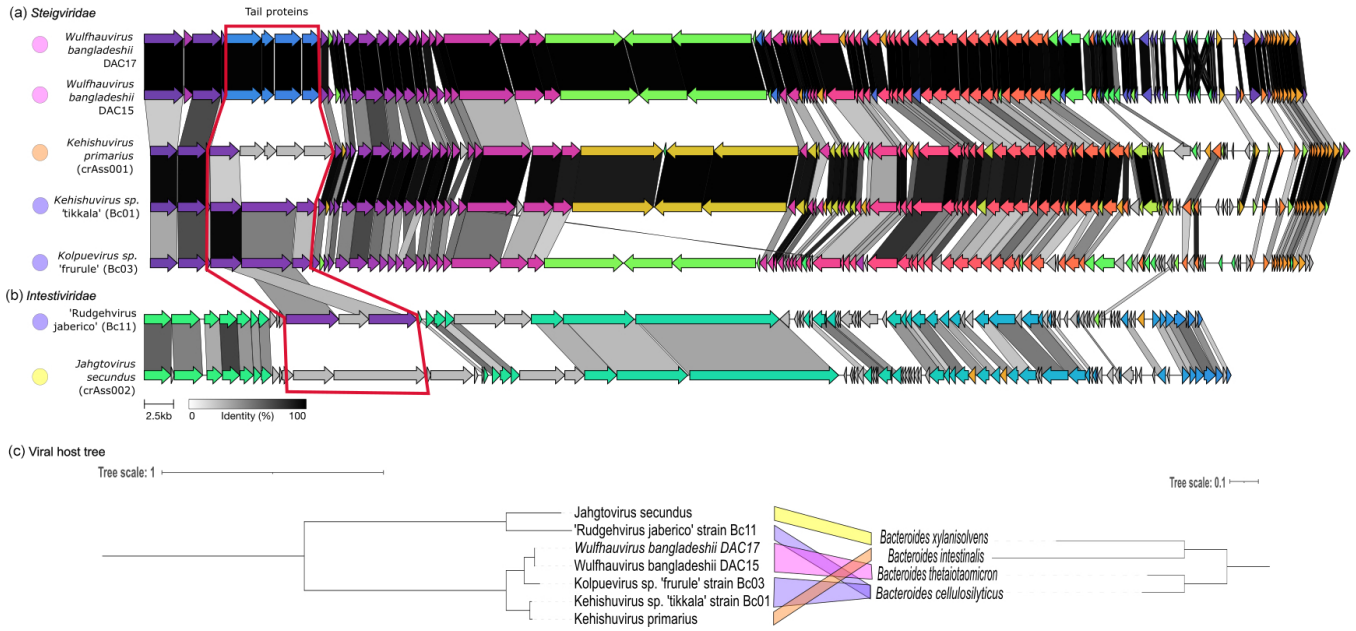

Fig. 3. Gene synteny across seven pure culture isolates across two *Crassvirales* families. (a) *Steigviridae* family comprising five isolates spanning across three genera. (b) *Intestiviridae* family comprising two isolates from two genera. Genes are represented as arrows, with their direction indicating the gene direction, and their colour indicating cluster group with the grey-coloured arrows representing unique genes that didn't form any clusters. Finally, the links connecting the genes are colour-coded based on sequence similarity, ranging from grey (30%) to black (100%). The tail proteins that were shared between the three isolates from this study are highlighted with a red box. A dot is added next to each of the phage to represent the bacterial host, *B. intestinalis* in orange, *B. cellulosilyticus* in purple, *B. thetaiotaomicron* in pink and *B. xylanisolvens* in yellow. (c) Viral host-tree constructed using the portal gene for *Crassvirales* species and 16S gene for the bacterial hosts, with unique colours connecting the phage to its bacterial host.

As there were multiple *Crassvirales* species infecting multiple bacterial hosts (Fig. 1a), we performed a coevolutionary test using Parafit [54] that supported random association between *Crassvirales* phages with their bacterial hosts (Parafit Global= 3.33, *p*-value >0.05).

Structural proteins playing a role in host interaction

To investigate the phage genes playing a role in host interaction, we compared all 1887 genes across the 18 *Crassvirales* genomes, including 14 from this study that infect bacterial host, *B. cellulosilyticus* WH2 and the four *Crassvirales* isolates that infect four different bacterial hosts, *B. intestinalis, B. thetaiotaomicron* and *B. xylanisolvens*. Together, from 18 *Crassvirales* isolates 1766 genes were categorized into 383 orthologous groups (Table S4), while the remaining 121 genes remained singletons. To reinforce the validity of this analysis, we corroborated the species tree inferred from orthogroups (Fig. S5) to have the same species level clustering as observed in phylogenetic tree (Fig. 1a). There was one exception, *J. secundus* which belongs to *Intestiviridae* family was grouped with the *Steigviridae* isolates instead of its relative 'R. jaberico' strains, due to gene duplication or recombination events in this genomes.

Following the species level clustering, we identified 64 orthogroups (193 genes) that were specific to *Kehishuvirus,* 55 orthogroups (564 genes) specific to *Kolpuevirus,* 89 orthogroups (187 genes) specific to *Wulfhauvirus,* 73 orthogroups (148 genes) specific to 'Rudgehvirus', and five orthogroups (ten genes) specific to *Jahgtovirus* genera (Table S4). Within these groups, only two orthogroups – OG000000 (including Bc01: WEU69745.1, Bc03: WEY17523.1, Bc11: WEU69858.1) and OG000008 (including Bc01: WEU69744.1, Bc03: WEY17522.1, Bc11: WEU69859.1) included genes only from the 14 *Crassvirales* isolates that infect the same bacterial host, *B. cellulosilyticus* WH2. However, in orthogroup OG000000, four gene duplication events occurred with at least 50% of the descendant species having retained both the gene duplicates, therefore this orthogroup was not investigated further. Conversely, there were no gene duplication events within OG000008.

To determine if the genes in OG000008 is undergoing selection pressure, we calculated the number of synonymous (d_S) and non-synonymous (d_N) mutations occurring (d_N/d_S <1). Averaging all the sequence pairs, we used the codon-based z-test to identify genes under selection and found that OG000008 rejected the null hypothesis (z-score=0.56, *p*-value<0.001), suggesting that these genes are under purifying selection. As recombination can impact this analysis, we ran Genetic Algorithm for Recombination

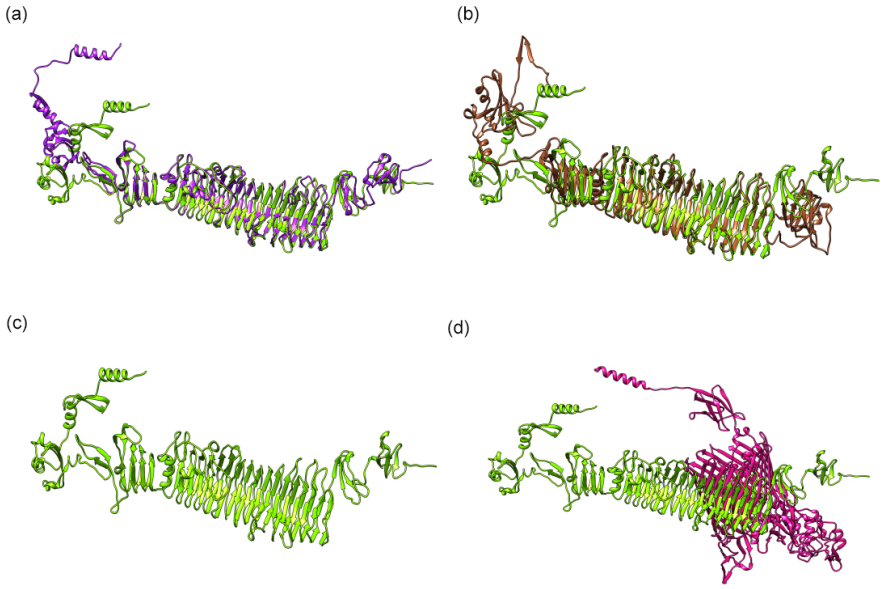

Fig. 4. 3D structure of tail spike proteins visualized using Chimaera. (a) Structural alignment of tail spike protein *Kehishuvirus* sp. 'tikkala' strain Bc01 (WEU69744.1 in green) with *Kolpuevirus* sp. 'frurule' strain Bc03 (WEY17522.1 in purple). (b) Structural alignment of tail spike protein *Kehishuvirus* sp. 'tikkala' strain Bc01 (WEU69744.1 in green) with 'Rudgehvirus jaberico' strain Bc11 (WEU69859.1 in brown). (c) 3D Structure of *Kehishuvirus* sp. 'tikkala' strain Bc01 (WEU69744.1 in green). (d) *Kehishuvirus* sp. 'tikkala' strain Bc01 (WEU69744.1 in green) docked with *Bacteroides cellulosilyticus* WH2 TonB-dependent receptor (A0A0P0GGA2 in pink).

Detection (GARD) to detect recombination, which identified five recombination breakpoints of which none were detected to be significant by the genetic algorithm. We therefore investigated the tail spike protein structure and role in host interaction.

Tail spike protein interacts with TonB-dependent receptors

To identify the potential host interactions, we predicted structure of all 103 proteins from '*K.* tikkala' strain Bc01, 109 proteins from '*K.* frurule' strain Bc03, and 83 proteins 'R. jaberico' strain Bc11 were generated using Colabfold [65] (protein structures available at doi.org/10.25451/flinders.21946034). Specifically, we compared the folded structures of tail spike proteins belonging to orthogroup OG000008 (Bc01: WEU69744.1, Bc03: WEY17522.1, Bc11: WEU69859.1) using Flexible structure AlignmenT by Chaining Aligned fragment pairs allowing Twists (FATCAT-rigid) method to show '*K.* tikkala' strain Bc01 is similar to '*K.* frurule' strain Bc03 with root-mean square deviation (RMSD) of 5.86 and only 74% of paired residues in the structural alignment (Fig. 4a). On the other hand, the tail spike protein of '*K.* tikkala' strain Bc01 exhibited a RMSD of 6.61 and 60% identity when compared to 'R. jaberico' strain Bc11 (Fig. 4b).

Each of the tail spike protein structures was individually docked against all 3223 predictions from the *B. cellulosilyticus* WH2 proteome available in the AlphaFold database using hdock-lite (). '*K.* tikkala' strain Bc01 tail spike protein (WEU69744.1) (Fig. 4c) interacted best with TonB-dependent receptors (UniProt ID: A0A0P0GGA2, hdock-score: -700) (Fig. 4d), '*K.* frurule' strain Bc03 protein (WEY17522.1) with another TonB-dependent receptor (UniProt ID: A0A0P0GR14, hdock-score: −694), and 'R. jaberico' strain Bc11 protein (WEU69859.1) with a different TonB-dependent receptor (UniProt ID: A0A0P0FZA4, hdock-score: −574).

DISCUSSION

The role of *Crassvirales* genomes in the human gut is enigmatic, which has been hindered by the limited number of cultured *Crassvirales* phages. Here, we address this gap by successfully isolating three novel *Crassvirales* species infecting *Bacteroides cellulosilyticus* WH2, belonging to different genera and families. This observation confirmed that these phages are not co-evolving with their bacterial hosts, rather they have a shared ability to exploit similar features in their bacterial host. Notably, we identified a unique tail spike protein shared among isolates infecting the same bacterial host, undergoing purifying selection and interacting with the TonB-dependent receptors on the bacterial surface.

The *Crassvirales* order is currently comprised of vast and diverse collection of genomes. Despite this, the study of these phages has been limited due to the scarcity of pure isolates. The challenge associated with successful isolation underscores the difficulty in identifying and predicting their associated bacterial hosts. In our study, we addressed this challenge through focusing on wastewater samples, a source for phages infecting different *Bacteroides* hosts. Employing this approach, we successfully

isolated 14 novel *Crassvirales* isolates specifically infecting *B. cellulosilyticus* WH2. These isolates were sequenced on different sequencing platforms, including Oxford Nanopore, Illumina Miseq or a combination of both. Nanopore assemblies provided high-quality and complete assemblies, however required polishing the assembly with Illumina reads to correct for frameshift errors that can fragment genes [70–72]. As a result, the 14 complete genomes were classified at the family and genus levels, denoted as three novel species (Fig. 1a), while the remaining isolates were grouped as strains of the same species (Table 1). The highest confidence isolate was selected for each species, *Kehishuvirus* sp. 'tikkala' strain Bc01, *Kolpuevirus* sp. 'frurule' strain Bc03 and 'Rudgehvirus jaberico' strain Bc11 and examined further.

Taxonomic assignment of the three novel species showed they belong to two families. *Kehishuvirus* sp. 'tikkala' strain Bc01 and *Kolpuevirus* sp. 'frurule' strain Bc03 were assigned to the *Steigviridae* family. This family also comprised of three other *Crassvirales* phages, *Kehishuvirus primarius, Wulfhauvirus bangladeshii* DAC15, and *Wulfhauvirus bangladeshii* DAC17 infecting other *Bacteroides* hosts. Notably, the two novel isolates exhibited clear, uniform circular spot morphology distinct from the turbid plaques observed in *K. primarius*, despite their close relationship within the same family. The third novel species, 'Rudgehvirus jaberico' strain Bc11 belonged to *Intestivirdae* family, along with other *Crassvirales* species, *Jahgtovirus secundus*. 'R. jaberico' strain Bc11 presented plaques with a circular halo surrounding the cleared spot, indicating depolymerase activity to break down the polysaccharides found on the bacterial cell wall.

Furthermore, comparing the three novel species, we found that their virion production, estimated from the number of plaques formed, was correlated to the number of tRNA genes within the genome [73]. However, it is possible there are other factors such as gene regulation, and host immune responses that could also be influencing virion production. Additionally, we conducted genome density analysis in association with capsid sizes and genome lengths, revealing an inconsistency with prior isolated *K. primarius* and *J. secundus* species. The capsid diameters of the new three novel *Crassvirales* species of virions (90 to 97 nm) were apparently 20% larger in size than those reported for *K. primarius* and *J. secundus* virions (77 nm) [27, 29]. However, considering that the reported values for *K. primarius* and *J. secundus* corresponded to the inscribed rather than the circumscribed diameters, a geometric correction of 22% that brought the genome density near 0.5 bp/nm^3. This correction aligned with a larger diameter measured in the recently published cryo-EM reconstruction of *K. primarius* [32]. The finding highlights the importance of accurately assessing virion dimensions and genome density to ensure consistency in the classification of *Crassvirales* phages.

The addition of the three novel *Crassvirales* species spanning multiple families infecting one bacterial host *B. cellulosilyticus* WH2 indicated these species may not be co-evolving with their bacterial hosts. We therefore tested all the successfully cultured *Crassvirales* species and their respective bacterial hosts to discover that they do not exhibit co-evolutionary patterns, rather support random association. These findings imply that the phage-host association within *Crassvirales* group are shaped by the environment and host interactions [54, 74]. Additionally, genome comparison of the known *Crassvirales* species showed greater shared similarity within genera. However, the three *Crassvirales* species despite belonging to three different genera shared two unique structural genes. Evolutionary analysis confirmed one of the two structural genes, encoding tail spike protein (comprising Bc01: WEU69744.1, Bc03: WEY17522.1, Bc11: WEU69859.1) formed an orthologous group, and is undergoing purifying selection pressure. Tail spike proteins have been shown to play crucial for binding to specific membrane receptors on the bacteria in tailed bacteriophages [75]. Therefore, through preserving this gene function, the phage can successfully infect and replicate within the host.

We found the tail spike proteins of the three novel *Crassvirales* species to interact with different TonB-dependent receptors on the bacterial surface, provide significant insights into the mechanism of phage-host interactions. The bacterial host, *B. cellulosilyticus* WH2 possesses a substantial repertoire of up to 112 TonB-receptors on its surface. These receptors are typically used by *Bacteroides* to take up starches [76], and have been associated with phage sensitivity [30, 31]. The tail spike protein also encodes for polysaccharide degrading enzymes such as glycoside hydrolase domain that target the capsular polysaccharides on the bacterial surface, allowing for phage-host interaction and lead to infection. This interaction therefore ensures successful propagation, highlighting the evolutionary adaptation between the *Crassvirales* phage and their bacterial hosts.

Overall, our study on the three novel *Crassvirales* species infecting *Bacteroides cellulosilyticus* WH2 revealed critical insights into their evolutionary dynamics and interactions with the bacterial host. The novel phages belonging to different genera but infect the same host provide a valuable model system for studying the interactions that occur within one of the dominant members of the gut microbiome.

Funding information
This work was supported by an award from NIH NIDDK RC2DK116713 and an award from the Australian Research Council DP220102915. PD's contribution was supported by the Polish National Agency for Academic Exchange (NAWA) Bekker Program fellowship no. BPN/BEK/2021/1/00416. AL's contribution was supported by National Science Foundation (NSF) Award 1951678.

Acknowledgements
This research/project was undertaken with the assistance of resources and services from Flinders University and the National Computational Infrastructure (NCI), which is supported by the Australian Government.

Author contributions
B.P. performed bioinformatics analysis and wrote the paper. A.A.V., C.S., S.K.G., M.A., N.J., L.B., C.H. and W.S.P. collected samples, isolated and cultured phages. M.F.M., M.A., K.P. and L.D. sequenced phages. C.L. and S.K.G. took TEM images. K.M., M.J.R., P.D., S.R.G., V.M., G.B. and A.L. performed bioinformatics analysis. S.H., D.W., A.M.S., E.A.D. and R.A.E. conceived the project, performed the bioinformatics, and wrote the paper with input from all authors.

Conflicts of interest
The authors declare no conflict of interest.

References

1. Hou K, Wu Z-X, Chen X-Y, Wang J-Q, Zhang D, *et al*. Microbiota in health and diseases. *Signal Transduct Target Ther* 2022;7:135.
2. Integrative HMP (iHMP) Research Network Consortium. The integrative human microbiome project. *Nature* 2019;569:641–648.
3. Shamash M, Maurice CF. Phages in the infant gut: a framework for virome development during early life. *ISME J* 2022;16:323–330.
4. Hugenholtz P, Goebel BM, Pace NR. Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. *J Bacteriol* 1998;180:6793– .
5. Pace NR, Stahl DA, Lane DJ, Olsen GJ. The analysis of natural microbial populations by ribosomal RNA sequences. In: Marshall KC (eds). *Advances in Microbial Ecology [Internet]*. Boston, MA: Springer US; 1986. pp. 1–55.
6. Inglis LK, Edwards RA. How metagenomics has transformed our understanding of bacteriophages in microbiome research. *Microorganisms* 2022;10:1671.
7. Roach MJ, Beecroft SJ, Mihindukulasuriya KA, Wang L, Paredes A, *et al*. Hecatomb: an end-to-end research platform for viral metagenomics. *bioRxiv* 2022:2022. DOI: 10.1101/2022.05.15.492003.
8. Hesse RD, Roach M, Kerr EN, Papudeshi B, Lima LFO, *et al*. Phage diving: an exploration of the carcharhinid shark epidermal virome. *Viruses* 2022;14:1969.
9. Anthenelli M, Jasien E, Edwards R, Bailey B, Felts B, *et al*. Phage and bacteria diversification through a prophage acquisition ratchet. *bioRxiv* 2020. DOI: 10.1101/2020.04.08.028340.
10. Knowles B, Silveira CB, Bailey BA, Barott K, Cantu VA, *et al*. Lytic to temperate switching of viral communities. *Nature* 2016;531:466–470.
11. Chevallereau A, Pons BJ, van Houte S, Westra ER. Interactions between bacterial and phage communities in natural environments. *Nat Rev Microbiol* 2022;20:49–62.
12. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, *et al*. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 2010;464:59–65.
13. HMP Consortium. Structure, function and diversity of the healthy human microbiome. *Nature* 2012;486:207–214.
14. Pargin E, Roach MJ, Skye A, Papudeshi B, Inglis LK, *et al*. The human gut virome: composition, colonization, interactions, and impacts on human health. *Front Microbiol* 2023;14:963173. DOI: 10.3389/fmicb.2023.963173.
15. Yutin N, Benler S, Shmakov SA, Wolf YI, Tolstoy I, *et al*. Analysis of metagenome-assembled viral genomes from the human gut reveals diverse putative CrAss-like phages with unique genomic features. *Nat Commun* 2021;12:1044.
16. Edwards RA, Vega AA, Norman HM, Ohaeri M, Levi K, *et al*. Global phylogeography and ancient evolution of the widespread human gut virus crAssphage. *Nat Microbiol* 2019;4:1727–1736.
17. Rossi A, Treu L, Toppo S, Zschach H, Campanaro S, *et al*. Evolutionary study of the crassphage virus at gene level. *Viruses* 2020;12:1035.
18. Norman JM, Handley SA, Baldridge MT, Droit L, Liu CY, *et al*. Disease-specific alterations in the enteric virome in inflammatory bowel disease. *Cell* 2015;160:447–460.
19. Dutilh BE, Cassman N, McNair K, Sanchez SE, Silva GGZ, *et al*. A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. *Nat Commun* 2014;5:4498.
20. Shkoporov AN. Create one new order (Crassvirales) including four new families, ten new subfamilies, 42 new genera and 73 new species (Caudoviricetes); 2021. https://ictv.global/ictv/proposals/2021.022B.R.Crassvirales.zip
21. Dutilh BE, Varsani A, Tong Y, Simmonds P, Sabanadzovic S, *et al*. Perspective on taxonomic classification of uncultivated viruses. *Curr Opin Virol* 2021;51:207–215.
22. Walker PJ, Siddell SG, Lefkowitz EJ, Mushegian AR, Adriaenssens EM, *et al*. Recent changes to virus taxonomy ratified by the International Committee on Taxonomy of Viruses (2022). *Arch Virol* 2022;167:2429–2440.
23. Borges AL, Lou YC, Sachdeva R, Al-Shayeb B, Penev PI, *et al*. Widespread stop-codon recoding in bacteriophages may regulate translation of lytic genes. *Nat Microbiol* 2022;7:918–927. DOI: 10.1038/s41564-022-01128-6.
24. Ivanova NN, Schwientek P, Tripp HJ, Rinke C, Pati A, *et al*. Stop codon reassignments in the wild. *Science* 2014;344:909–913.
25. Crisci MA, Chen L-X, Devoto AE, Borges AL, Bordin N, *et al*. Closely related Lak megaphages replicate in the microbiomes of diverse animals. *iScience* 2021;24:102875.
26. Peters SL, Borges AL, Giannone RJ, Morowitz MJ, Banfield JF, *et al*. Experimental validation that human microbiome phages use alternative genetic coding. *Nat Commun* 2022;13:5710.
27. Shkoporov AN, Khokhlova EV, Fitzgerald CB, Stockdale SR, Draper LA, *et al*. ΦCrAss001 represents the most abundant bacteriophage family in the human gut and infects bacteroides intestinalis. *Nat Commun* 2018;9:4781.
28. Hryckowian AJ, Merrill BD, Porter NT, Van Treuren W, Nelson EJ, *et al*. Bacteroides thetaiotaomicron-infecting bacteriophage isolates inform sequence-based host range predictions. *Cell Host Microbe* 2020;28:371–379..
29. Guerin E, Shkoporov AN, Stockdale SR, Comas JC, Khokhlova EV, *et al*. Isolation and characterisation of ΦcrAss002, a crAss-like phage from the human gut that infects bacteroides xylanisolvens. *Microbiome* 2021;9:89.
30. Shkoporov AN, Khokhlova EV, Stephens N, Hueston C, Seymour S, *et al*. Long-term persistence of crAss-like phage crAss001 is associated with phase variation in bacteroides intestinalis. *BMC Biol* 2021;19:163.
31. Porter NT, Hryckowian AJ, Merrill BD, Fuentes JJ, Gardner JO, *et al*. Phase-variable capsular polysaccharides and lipoproteins modify bacteriophage susceptibility in bacteroides thetaiotaomicron. *Nat Microbiol* 2020;5:1170–1181.
32. Bayfield OW, Shkoporov AN, Yutin N, Khokhlova EV, Smith JLR, *et al*. Structural atlas of a human gut crassvirus. *Nature* 2023;617:409–416.
33. McNulty NP, Wu M, Erickson AR, Pan C, Erickson BK, *et al*. Effects of diet on resource utilization by a model human gut microbiota containing bacteroides cellulosilyticus WH2, a symbiont with an extensive glycobiome. *PLoS Biol* 2013;11:e1001637.

34. Summer EJ. Preparation of a phage DNA fragment library for whole genome shotgun sequencing. *Methods Mol Biol* 2009;502:27–46.
35. Kim AH, Armah G, Dennis F, Wang L, Rodgers R, *et al*. Enteric virome negatively affects seroconversion following oral rotavirus vaccination in a longitudinally sampled cohort of Ghanaian infants. *Cell Host & Microbe* 2022;30:110–123..
36. Wick RR. Filtlong: Tool for filtering long reads by quality; 2018. https://github.com/rrwick/Filtlong/
37. Cantu VA, Sadural J, Edwards R. PRINSEQ++, a multi-threaded tool for fast and efficient quality control and preprocessing of sequencing datasets. *PeerJ Preprints* 2019. DOI: 10.7287/peerj.preprints.27553v1.
38. Roach MJ, Pierce-Ward NT, Suchecki R, Mallawaarachchi V, Papudeshi B, *et al*. Ten simple rules and a template for creating workflows-as-applications. *PLoS Comput Biol* 2022;18:e1010705.
39. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. *Nat Biotechnol* 2019;37:540–546.
40. Li D, Luo R, Liu C-M, Leung C-M, Ting H-F, *et al*. MEGAHIT v1.0: a fast and scalable metagenome assembler driven by advanced methodologies and community practices. *Methods* 2016;102:3–11.
41. Wick RR, Schultz MB, Zobel J, Holt KE. Bandage: interactive visualization of de novo genome assemblies. *Bioinformatics* 2015;31:3350–3352.
42. Mallawaarachchi V, Wickramarachchi A, Lin Y. GraphBin: refined binning of metagenomic contigs using assembly graphs. *Bioinformatics* 2020;36:3307–3313.
43. Mallawaarachchi VG, Lin Y. MetaCoAG: binning metagenomic contigs via composition, coverage and assembly graphs. *Res Comput Mol Biol* 2022.
44. Mallawaarachchi VG, Wickramarachchi AS, Lin Y. Improving metagenomic binning results with overlapped bins using assembly graphs. *Algorithms Mol Biol* 2021;16:3.
45. Raiko M. viralVerify: viral contig verification tool; 2021. https://github.com/ablab/viralVerify
46. Woodcroft BJ. CoverM:DNA read coverage and relative abundance calculator; 2021. https://github.com/wwood/CoverM
47. Zimin AV, Salzberg SL, Ouzounis CA. The genome polishing tool POLCA makes fast and accurate corrections in genome assemblies. *PLoS Comput Biol* 2020;16:e1007981.
48. Carrillo D. CrassUS - Crassvirales Uncovering Software; 2022. https://github.com/dcarrillox/CrassUS
49. Nakamura T, Yamada KD, Tomii K, Katoh K, Hancock J. Parallelization of MAFFT for large-scale multiple sequence alignments. *Bioinformatics* 2018;34:2490–2492.
50. Price MN, Dehal PS, Arkin AP. FastTree 2--approximately maximum-likelihood trees for large alignments. . *PLoS One* 2010;5:e9490.
51. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res* 2019;47:W256–W259.
52. Gilchrist CLM, Chooi Y-H. clinker & clustermap.js: automatic generation of gene cluster comparison figures. *Bioinformatics* 2021;37:2473–2475.
53. Chan PP, Lowe TM. tRNAscan-SE: searching for tRNA genes in genomic sequences. *Methods Mol Biol* 2019;1962:1–14.
54. Legendre P, Desdevises Y, Bazin E, Page RDM. A statistical test for host-parasite coevolution. *Syst Biol* 2002;51:217–234.
55. Schneider CA, Rasband WS, Eliceiri KW. NIH Image to ImageJ: 25 years of image analysis. *Nat Methods* 2012;9:671–675.
56. The GIMP Development Team. GIMP; 2019. https://www.gimp.org
57. Luque A, Benler S, Lee DY, Brown C, White S. The missing tailed phages: prediction of small capsid candidates. *Microorganisms* 2020;8:1944.
58. Lee DY, Bartels C, McNair K, Edwards RA, Swairjo MA, *et al*. Predicting the capsid architecture of phages from metagenomic data. *Comput Struct Biotechnol J* 2022;20:721–732.
59. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol* 2019;20:238.
60. Edgar RC. High-accuracy alignment ensembles enable unbiased assessments of sequence homology and phylogeny. *Bioinformatics*:2021. DOI: 10.1101/2021.06.20.449169.
61. Stecher G, Tamura K, Kumar S. Molecular Evolutionary Genetics Analysis (MEGA) for macOS. *Mol Biol Evol* 2020;37:1237–1239.
62. Tamura K, Stecher G, Kumar S, Battistuzzi FU. MEGA11: Molecular Evolutionary Genetics Analysis Version 11. *Mol Biol Evol* 2021;38:3022–3027.
63. Li WH, Wu CI, Luo CC. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Mol Biol Evol* 1985;2:150–174.
64. Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SDW. GARD: a genetic algorithm for recombination detection. *Bioinformatics* 2006;22:3096–3098.
65. Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, *et al*. ColabFold: making protein folding accessible to all. *Nat Methods* 2022;19:679–682.
66. Li Z, Jaroszewski L, Iyer M, Sedova M, Godzik A. FATCAT 2.0: towards a better understanding of the structural diversity of proteins. *Nucleic Acids Res* 2020;48:W60–W64.
67. Ye Y, Godzik A. Flexible structure alignment by chaining aligned fragment pairs allowing twists. *Bioinformatics* 2003;19:ii246–ii255.
68. Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, *et al*. AlphaFold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res* 2022;50:D439–D444.
69. Yan Y, Tao H, He J, Huang S-Y. The HDOCK server for integrated protein-protein docking. *Nat Protoc* 2020;15.
70. Nanoporetech Consortium. medaka: Sequence correction provided by ONT Research; 2022. https://github.com/nanoporetech/medaka/releases
71. Arumugam K, Bağcı C, Bessarab I, Beier S, Buchfink B, *et al*. Annotated bacterial chromosomes from frame-shift-corrected long-read metagenomic data. *Microbiome* 2019;7.
72. Cook R, Brown N, Rihtman B, Michniewski S, Redgwell T, *et al*. The long and short of it: benchmarking viromics using Illumina, Nanopore and PacBio sequencing technologies. *bioRxiv* 2023.
73. Delesalle VATankeNTVill AC, Krukonis GP. Testing hypotheses for the presence of tRNA genes in mycobacteriophage genomes. *Bacteriophage* 2016;6:e1219441.
74. Papudeshi B, Rusch DB, VanInsberghe D, Lively CM, Edwards RA, *et al*. Host association and spatial proximity shape but do not constrain population structure in the mutualistic symbiont *Xenorhabdus bovienii*. *mBio* 2023;14:e0043423.
75. 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.
76. Pollet RM, Martin LM, Koropatkin NM. TonB-dependent transporters in the bacteroidetes: unique domain structures and potential functions. *Mol Microbiol* 2021;115:490–501.