Novel crAssphage isolates exhibit conserved gene order and purifying selection of the host specificity protein

New crAssphage isolates infecting *Bacteroides cellulosilyticus WH2*

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

[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]Adelaide Microscopy, University of Adelaide, Adelaide, SA, 5005, Australia

[4]Department of Environmental Microbiology and Biotechnology, Institute of Microbiology, Faculty of Biology, University of Warsaw, Miecznikowa 1, Warsaw, 02-096, Poland

[5]Department of Pathology & Immunology, Washington University School of Medicine, St. Louis, MO, 63110, USA

[6]Adelaide Medical School, Faculty of Health and Medical Sciences, The University of Adelaide, Adelaide, SA, 5005, Australia.

[7]Department of Mathematics and Statistics, San Diego State University, 5500 Campanile Drive, San Diego, CA, 992182, USA

[8]Computational Science Research Center, San Diego State University, 5500 Campanile Drive, San Diego, CA, 992182, USA

*Corresponding author: Bhavya Papudeshi, nala0006@flinders.edu.au

Abstract

Bacteroidota are the most common bacteria in the human gut and are responsible for degrading complex polysaccharides that would otherwise remain undigested. The abundance of Bacteroides in the gut is shaped by phages such as crAssphages that infect and kill them. While close to 600 genomes have been identified computationally, only four have been successfully cultured. Here, we identify and characterize three novel crAssphage species isolated from wastewater and infecting the bacterial host *Bacteroides cellulosilyticus* WH2*.* We named the novel species, *Kehishuvirus winsdale* (Bc01), *Kolpuevirus frurule* (Bc03), and *Rudgehvirus redwords* (Bc11) which span two different families and three genera. These phages may not have co-evolved with their respective bacterial hosts. The phages had a conserved gene arrangement with known crAssphages, but gene similarity within phages belonging to the same taxa was highly variable. Across the three species, only two structural genes encoding a hypothetical protein and a tail spike protein were similar. Evolutionary analysis revealed the tail spike protein is undergoing purifying selection and was predicted to bind to a TonB-dependent transporter on the host cell surface, suggesting a role for host specificity. This study expands the known crAssphage isolates and reveals insights into the crAssphage infection mechanism. The availability of pure cultures of multiple crAssphage infecting the same host provides an opportunity to perform controlled experiments on one of the most dominant members of the human enteric virome.

Introduction

There is an intricate relationship between gut microbiomes and human health. A healthy gut microbiome contains a high diversity of microbes that help with digestion, regulate the immune system, and alter brain function (1–3). Application of the metagenomics technique, a culture-independent technique that captures the diversity of the microbial community in a sample (4,5) has transformed our understanding of the prevalence of bacteria and the corresponding bacteriophages in the environment (6–9). From metagenomic datasets, we have observed a correlation between bacterial and bacteriophage populations which suggests bacteriophages play a role in controlling the ratios of different bacteria (10,11). For instance, within an environment with high or low bacterial densities, phages favor integration into the bacterial host, while at intermediary densities the phages favor host lysis (12). The human gut microbiome has high bacterial densities, including a high abundance of Bacteroidota (formerly Bacteroidetes) (13–15), which are shaped by the phages that infect and kill them such as crAssphages. The dsDNA crAssphages have a podovirus-like morphology, genomes ranging between 100 and 200 kb, and conserved gene order (16–18). These bacteriophages are ubiquitous, can stably colonize an individual, and do not appear to be associated with health or disease states (17,19). This phage was first discovered computationally by cross-assembling DNA sequence reads from human gut microbiome samples (20), and to date, there have been close to 600 crAssphage genomes identified computationally, but only four cultured phages (16,21).

The isolation of crAssphages *in vitro* remains a challenge with only four successful pure isolates after eight years of experimental research. In 2018, *Kehishuvirus primarius* (crAss001) was isolated from *Bacteroides intestinalis* (22). A brace of phages, *Wulfhauvirus bangladeshii* DAC15 and DAC17, were isolated from wastewater effluent infecting *Bacteroides thetaiotaomicron* (23). Finally, *Jahgtovirus secundus* (crAss002) was isolated on *Bacteroides xylanisolvens* (24). All the pure isolates exhibited host specialist morphotypes that can be maintained in the continuous host culture, but none possess lysogeny-related genes to suggest a lysogeny lifestyle (22,25). Alternative mechanism proposed is that these phages host populations cycle between sensitive and resistant states through phase variation of capsular polysaccharides or through exhibiting pseudolysogeny or a carrier state phenotype (25). Transmission electron microscopy (TEM) of these phages has confirmed podovirus-like morphology (22,24,26).

Analysis of all known crAssphage that includes the pure cultures and the computationally identified crAssphage led to grouping of these phages to distinct group. Recently, International Committee of Taxonomy of Viruses (ICTV) published a formal report on classification of Crassvirales order into four families, ten subfamilies, 42 new genera, and 72 new species (18,27). This classification was based on phylogenetic analysis of conserved structural genes, such as major capsid protein (MCP), terminase large subunit (*terL*), and portal protein (portal). Despite this progress and expansion of the Crassvirales order, functional annotation of viral genome remains challenging, with lots of genes annotated as hypothetical proteins with no known biological function. However, subsequent analysis with the Crassvirales has improved our understanding of phage diversity, and revealed their unique biological characteristics (28,29). For instance, 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 that are involved in the successful replication of the phage that is activated in the different stages of phage infection (16). They also revealed unique characteristics of the lineage including switching DNA polymerases, alternative coding strategies (30–33), and the variable density of introns across taxa (16,33). Further, cryogenic-electron microscopy of *K. primarius* (crAss001) added the structural basis of the genes and description of the mechanisms of assembly and infection (26). An evolutionary study of crAssphages showed that capsid protein is the most conserved gene, followed by two other uncharacterized proteins within the structural module (18). The remaining modules are highly variable with little homology to other known proteins.

Here we present the isolation and characterization of three novel crAssphage isolates from wastewater. We present three new species within *Steigviridae* and *Intestiviridae* families and how they infect the same host, *Bacteroides cellulosilyticus* WH2*.* We computationally predict the genes playing a role in host specificity, providing insights into the evolution of these dominant phages, and how their interactions shape the gut microbiome.

Results

Isolation and assembly of phage genomes *B. cellulosilyticus WH2*

We isolated and sequenced 16 *B. cellulosilyticus WH2* phage isolates, with phage (Bc01 to Bc12) genomes sequenced on the Oxford Nanopore platform assembled using Flye and Unicycler, while genomes (Bc01 to Bc08, Bc13 to Bc16) sequenced on the

Illumina platform were assembled using MEGAHIT assembly. We used multiple assemblers as bioinformatics assembly tools have some level of algorithmic nuances as well as sensitivities to different features (Table S1). For each genome, our selection criterion (viral, highest read coverage, unitig approximately 100kb) identified as complete phage genome for 14 of the 16 phage samples across both sequencing platforms (Table S1). Two genomes (Bc04, Bc12) did not assemble due to their coverage profile (Figure S1), were also identified to be replicates of other assembled genomes and therefore left out of this analysis. Finally, the 14 phage genomes from the Nanopore based assembly were polished with Illumina reads correcting for substitution, insertion, and deletion errors (Table S1). These assembled unitigs were then processed to calculate the average nucleotide identity (ANI) (part of CrassUS workflow), and genomes with more than 95% ANI were grouped as likely phage replicates (Table S1). This reduced our phage genomes to three clusters of phages from which we selected the highest confidence genomes: Bc01, Bc03, and Bc11.

Taxonomic assignment of the three novel isolates

To identify the taxa of the three representative isolates, BLASTN searches were performed against the non-redundant (nr/nt) NCBI database confirming that all three representative isolates (Bc01, Bc03 and Bc11) were crAssphage (34). The phylogenetic tree built from conserved proteins including the major capsid protein (MCP), portal, terminase large subunits (*terL*) genes from all 14 assembled phage genomes demonstrated that Bc01 and Bc03 belong to the *Steigviridae* family, and Bc11 to the *Intestiviridae* family (Figure 1, Figure S2). The taxonomic classification follows the ICTV report with suggestions on defining a taxonomy for phage genomes belonging to the

Crassvirales order. Taxonomic classification of the three genomes was also confirmed with ANI and shared protein information, identifying Bc01 to belong to *Kehishuvirus*, Bc03 to *Kolpuevirus*, and Bc11 to a novel genus group that we propose to call *Rudgehvirus*.

All three isolates are novel species exhibited less than 95% identity to any known crAssphage genome. Bc01 is most similar to the reference genome *Kehishuvirus primarius* (crAss001, MH675552) with 95.5% identity across 79.1% genome coverage. Bc03 aligns with *Kolpuevirus hominis* (crAssphage cr126_1, MT774391) with 82.8% identified across 53.73% query coverage. Bc11 aligns with the reference genome *Jahgtovirus intestinalis* (OGOL01000109) with 74.7% identity across only 9.9% query coverage. We named these novel species *Kehishuvirus winsdale* (Bc01), *Kolpuevirus frurule* (Bc03), and *Rudgehvirus redwords* (Bc11) (Table 1).

Table 1: Genome characteristics of the three novel crAssphage representative isolates

Figure 1: Phylogenetic tree of the portal protein showing that isolate *K. winsdale* (Bc01) and *K. frurule* (Bc03) belong to the family *Steigviridae* (cyan), and *R. redwords* (Bc11) to *Intestiviridae* (red). The outgroup is set to *Cellulophaga phage phi13:2.*

Genome characteristics of the novel species

Kehishuvirus winsdale (Bc01) is 100,841 bp, with 104 proteins and 24 tRNAs identified within the genome (Table 1). The cumulative GC% of this genome is 35.09% (Table 1) which is lower than the bacterial host GC% of 42.8%. This genome is 95% similar to the isolate *K. primarius* (crAss001). In a few crAssphage genomes, a direct terminal repeat has been proposed to be involved in genome packaging (including *K. primarius*) but no direct terminal repeat was identified in *K. winsdale*. There was also no evidence of stop codon reassignment in this genome as observed in closely related reference *K. primarius* (Table S2). While 24 tRNA genes were identified within a contiguous segment from 88,172 to 93,727 bp on the genome, no tRNA suppressors were found that would be used to stop codon reassignment (Table S2). Functional gene annotations were performed by aligning the 104 predicted genes across all the known Caudovirales genomes and generating functional assignments for 48 genes (Table S3). From the functional assignments, we were able to divide the genome into three regions (structural, transcription, and replication) as observed across all the other crAssphage genomes (Figure 2). There was no lysogeny module or integrase-related genes detected within the phage genome and there was no evidence of temperate replication in the plaque morphology.

Kolpuevirus frurule (Bc03) belonged to the same family, *Steigviridae* as *K. winsdale* (Bc01). This genome was shorter in length (99,523 bp) but with a similar GC content of 33% (Table 1). This genome is 82% similar to a reference genome to this isolate *Kolpuevirus hominis* (CrAssphage cr126_1). Neither direct terminal repeat genome packaging, stop-codon reassignment, nor integration and lysogeny were observed (Table S2). Functional characterization of this genome predicted 108 genes, of which 45 were assigned a function (Table S4). In this genome, only four tRNA genes were predicted, two coding for arginine (TCT anticodon), and the other two encoding asparagine (GTT anticodon), and tyrosine (GTA anticodon).

Finally, *Rudgehvirus redwords* (Bc11) is 90,575 bp in length with 29.15% GC, and there were 84 genes encoded within this genome (Table 1). Taxonomic assignment of this genome placed this isolate within the *Intestiviridae* family, but in a novel genus. Currently, this genus includes only this isolate (Figure 1), with the neighboring clade including *Jahgtovirus* crAssphages*.* We named the new genus *Rudgehvirus* following the ICTV Caudovirales order naming suggestions. The reference genome that is most closely related to this genome is *Jahgtovirus intestinalis* (uncultured phage cr54_1) with 75%. In addition, pure culture isolate crass002 (*J. secundus*) also belongs to the closely related clade. There were 84 predicted genes with standard stop codons, as there was no evidence of codon reassignment (Table S2), and 36 of these genes were assigned a function (Table S4). Similar to one-third of crAssphage genomes currently known, there were no tRNA genes (Table S2) identified in this genome.

We compared the gene arrangement across the three novel species by comparing gene similarity with clinker (only greater than 30%) and demonstrated that *Steigviridae* (Bc01, Bc03) isolates shared more gene similarity compared to *Intestitviridae* (Bc11) in congruence with taxa classification (Figure 2). Across the three isolates, there were only three shared genes: two encoding structural proteins, and one hypothetic gene within replication machinery. *Steigviridae* isolates shared 68 genes: 24 belonging to the structural region; none within the transcription machinery; and the remaining 44 genes within the replication machinery. Finally, *Kehishuvirus* (Bc01) and *Rudgehvirus* (Bc11), and *Kolpuevirus* (Bc03) and *Rudgehvirus* (Bc11) shared two structural genes along with one hypothetical gene within the replication machinery.

Figure 2: A) Transmission electron microscopy images negatively stained with uranyl acetate of the three isolates, *K. winsdale* (Bc01), *K. frurule* (Bc03), and *R. redwords* (Bc11). B) Gene arrangement and functional annotation of the three genomes color-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).

Podovirus-like morphology

Transmission electron microscopy (TEM) of the three species revealed all have a podovirus-like morphology (Figure 2). *K. winsdale* (Bc01) displayed polyhedral capsids with a diameter of 94 ± 3 nm, tails with collar structures that were 34 ± 3 nm with several tail fibers with variable lengths. *K. frurule* (Bc03) included capsids of diameter 97 ± 3 nm, a tail with collar structures of 33 ± 3 nm, and several tail appendages of variable length. Finally, *R. redwords* (Bc11) was slightly smaller in size with capsids of size 90 ± 4 nm, tails with collar structures measuring 25 ± 4 nm, with tail appendages that were of variable length but different tail structure than the other two crAssphage genomes.

Synteny across crAssphages

Comparing the three new species from this study with the four reference crAssphage isolates available showed gene similarity expected from the taxonomic assignment. Within the *Steigviridae* genomes including two new species and three reference genomes, *K. winsdale* (Bc01) was most similar to *K. primarius* (crAss001) sharing 76 of 106 genes, and two isolates belonging to the *Wulfhauvirus* genus (DAC15 and DAC17) shared 115 of the 121 genes with the highest similarity. Meanwhile, *K. frurule* (Bc03) was equidistant from the species of *Kehishuvirus* (68 genes shared) and *Wulfhauvirus* genus (71 genes) (Figure 3A). The third isolate *R. redwards* (Bc11) was compared to the *J. secundus* (crAss002) as these two isolates belonged to an *Intestiviridae* family, and they shared 37 genes including 11 structural genes, three transcription genes, and 23 replication-related genes (Figure 3B). However, they are both overall dissimilar to genomes from *Steigviridae* (Figure 3A). A notable characteristic was that tail-related genes were only shared by those isolates infecting the same host. *K. winsdale* (Bc01), *K. frurule* (Bc03) and *R.redwords* (Bc11) shared unique tail-related genes although they belonged to different genera (Figure 2).

Figure 3. Gene synteny across seven pure culture isolates across two crAssphage families A) *Steigviridae* family comprising five isolates spanning across three genera B) *Intestiviridae* family comprising two isolates from two genera. The arrows represent genes, with their direction indicating the gene direction, and their color indicating cluster group with the grey-colored arrows representing unique genes that didn't form any clusters. The functional color coding for A and B are different. Finally, the links connecting the genes are color-coded based on sequence similarity, ranging from grey (30%) to black (100%). New isolates included in this study are represented with * next to their name, and the tail proteins that were shared between the three isolates from this study are highlighted with a red box.

Structural proteins playing a role in host specificity

To test if the common genes across the three new species in this study play a role in host specificity, we performed evolutionary analyses comparing 1,887 genes across the 14 crAssphage isolates from this study along with four reference genomes. Overall, 1,766 of them were categorized into 383 orthologous groups (Table S6), and the remaining 121 genes were singletons. 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 5 orthogroups (10 genes) specific to *Jahgtovirus* genera. Testing for host-specificity within each of these groups only identified two orthogroups–OG000000 (including Bc01 protein gp23) and OG000008 (including Bc01 protein gp22)–encoding a hypothetical gene and tail spike protein within the structural block, which only included genes from crAssphage isolates that infect *B. cellulosilyticus* WH2 (Table 2). These two orthogroups also represent the two tail proteins that were common across the three new isolates in this study (Figure 3).

Calculating the number of synonymous (d_S) and non-synonymous (d_N) mutations occurring within the orthogroups reflected d_N/d_S <1, suggesting purifying selection. Averaging all the sequence pairs, we used the codon-based z-test to identify genes under selection and found that OG000000 followed the null hypothesis (z-score=0.56, p-value=0.29) suggesting that the gene is under neutral evolution, while OG000008 rejected the null hypothesis (z-score=0.56, p-value<0.001), suggesting that the gene is under purifying selection.

Table 2: Inferred selection of the three orthogroups with genes from all crAssphage isolates infecting the host *B. cellulosilyticus*. Orthogroups under purifying selection are represented with an asterisk (*)

Prediction of the phage receptor binding domain

The predicted structure of all 103 proteins from *K. winsdale* (Bc01) were generated using Colabfold (35) (Protein structures available at doi.org/10.25451/flinders.21946034). Each of those models was individually docked against all 3,223 predictions from the *B. cellulosilyticus* WH2 proteome available in the AlphaFold database using hdock-lite. The PDB files for all the structural models and a summary of the docking scores are provided in the supplemental material (Table S7). The strongest predicted docking interactions were between *K. winsdale* protein RNA polymerase subunit (gp47) with the bacterial transmembrane EamA-family transporter (UniProt ID: A0A0P0GB77, hdock-score =-711), tail spike (gp22) with TonB-dependent receptors (UniProtID: A0A0P0GGA2, hdock-score =-700), and transmembrane protein (gp44) with outer membrane protein (UniProt ID: A0A0N7IEG6, hdock-score =-672). Although 34 of the phage proteins bind to bacterial proteins named TonB receptors, they each had unique UniProtID. The bacterial host, *B. cellulosilyticus* WH2 contains six TonB homologs and 112 TonB-dependent transporters, similar to other *Bacteroides* genomes. These transporters are typically used by *Bacteroides* to take up starches (36), but we propose, based on structural modelling, that one or more of the TonB-dependent receptors, are utilized by crAssphage to penetrate through the cell membrane.

Figure 4: A) 3D structure of tail spike protein (gp22) from *K. winsdale* (Bc01) visualized using PyMOL, B) 3D structure of tail spike protein (gp22) interaction (black) with TonB-dependent receptor (yellow) from the bacterial host, *B. cellulosilyticus WH2.*

Discussion

The role of crAssphage in the human gut is enigmatic, partly because so few crAssphages have been cultured (17). Here, we present three novel crAssphages spanning three genera but infecting one *B. cellulosilyticus* WH2, which suggests that these phages may not have co-evolved with their bacterial host, suggesting other factors driving to crAssphage evolution. Functional characterization of these novel species and comparison to other reference genomes revealed synteny across crAssphage isolates, with a higher range of gene similarity observed within phage belonging to the same taxa. Across the three novel species, we identified a single structural gene encoding tail spike protein that was found in all three species and was under purifying selection. Protein docking predicted that the tail spike protein binds to the TonB-dependent transporter on the bacterial surface, which is a known receptor for phage sensitivity, suggesting the tail spike protein plays a role in host specificity.

We isolated and classified the three novel crAssphages to family and genus level classification (Figure 1), with two species in the *Steigviridae* family, that we named *Kehishuvirus winsdale* (Bc01), and *Kolpuevirus frurule* (Bc03). *K. winsdale* (Bc01) is closely related to the first crAssphage isolate, *K. primarius* (crAss001), while *K. frurule* (Bc03) is the first pure culture isolate from its genus. Finally, the third species belongs to a novel genus and species that we named *Rudgehvirus redwords* (Bc11). Whole genomic comparison revealed synteny across all the crAssphage genomes with gene order conserved, despite differences in gene similarity (Figure 2, Figure 3) as observed Crassvirales order (16). As expected, isolates from the same family-level classification have more gene similarity than across families (Figure 3). Within a family, conserved genes include structural genes (MCP, terminase large subunit, portal, Integration Host Factor (IHF), tail tubular and stabilization protein), replication machinery genes (DnaG primase, DnaB helicases, SNF2 family, AAA+ superfamily and lysis genes), and transcription genes (RNA polymerase subunits, nuclease of the PDDEXK family) (Figure 3, Table S3, S4 and S5) (42) as observed in other crAssphages (16). The study found that the crAssphage isolates within the *Steigviridae* family encoded DNA polymerase A, while those within the *Intestiviridae* family encoded DNA polymerase B. This observation is consistent with the DNA polymerase switching previously reported in crAssphages(16). While some crAssphage isolates have been found to have reassigned stop codons (30,33), none of the novel isolates in this study reassigned stop codons by measuring coding density and the absence of suppressor tRNA (Table S2).

On the other hand, comparing the morphology of the three species showed differences in capsid sizes relative to the reference crAssphage isolates. The capsid diameter was larger in size, between 90 to 97 nm for the three species in this study (Figure 2) compared to the reference genomes that were estimated to be 77 nm (22,24). We confirmed this difference was not an artefact of how the capsid diameter was measured by repeating the measurements for *K. primarius* (crAss001) and *J. secundus* (crAss002) (Figure S4), and all the phages were negatively stained with uranyl acetate. In tailed phages, the genome length is expected to increase as the cube of the capsid diameter, because DNA is packaged at a constant density of about 0.5bp/nm^2 (37,38). However, the two closely related genomes *K. primarius* (crAss001) and *K. winsdale* (Bc01) only varied by 1,838 bp while *K. winsdale* capsid was 22% larger in diameter (94 nm versus 77 nm), displaying an internal volume that could accommodate 80,000 bp more than *K. primarius* (crAss001). The difference in the capsid size could be attributed to the variations in the scaffold protein, as observed in *Staphylococcus aureus* phages where the same major capsid protein assembled within different scaffolding proteins generates different capsid sizes (39,40). Two possible mechanisms might be responsible for filing the additional interval volume of *K. winsdale* (Bc01). First, this species could follow a headful packaging mechanism storing more DNA than just the viral genome length. This packaging mechanism was predicted in *J. secundus* (crAss002) (24); nonetheless, the absence of direct terminal repeats (Table 1) is uncommon in tailed phages using the headful mechanism (41). Second, the presence of internal proteins in the capsid could occupy the remaining volume, as shown in the capsid of *K. primarius* (crAss001) genome (26).

Isolation of multiple crAssphages infecting the same host revealed that the tail spike protein is conserved among crAssphages infecting the same host (Figure 3). Tail spike proteins are important in tailed bacteriophages for binding to specific membrane receptors on the bacteria (43). We also found evidence of purifying selection acting on the tail spike protein, indicating its essential role in host specificity (Table 2). Using structural modelling, we predicted that the tail spike protein interacts with the TonB-dependent receptor on the bacterial surface, a gene that has been shown to play a critical role in phage infection of Bacteroidota and other species (Porter et al., 2020;

Shkoporov, Khokhlova, et al., 2021). The interaction between crAssphage and TonB-dependent receptors along with capsular polysaccharides has been shown to play a role in the long-term persistence observed in these phages (25). TonB-dependent receptors are involved in the uptake of various nutrients, including iron, which is an essential element for many bacteria. By targeting these transporters, the crAssphage is able to hijack the host's nutrient acquisition machinery such as capsular polysaccharide biosynthesis for its own benefit, allowing it to persist and replicate within the host over a longer period of time (25). Further, another study has shown that crAssphage RNA polymerase and tail proteins were undergoing positive selection by comparing genomes assembled from parent and infant fecal samples(44). However, in that study they didn't characterize the crAssphage taxa or host, therefore the difference in the selection pressure may be likely due to these external evolutionary pressures. We also suggest that tail spike proteins can be used to cluster crAssphage genomes to groups that potentially infect the same bacterial host.

Overall, in this study we identified and characterized three novel crAssphage species isolates from wastewater and infecting the bacterial host *Bacteroides cellulosilyticus WH2.* Our findings expand the known crAssphage isolates and provide insight into the role of the tail proteins in host specificity and infection. Outcomes of these efforts resulted in a unique model of three distinct crAssphages infecting the same strain which can be now utilized experimentally to study one of the dominant members of the gut microbiome, 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 1L Nalgene bottles. Upon collection, the bottles were placed on ice and in the dark for processing. An aliquot of 30 ml influent was put into a sterile 50 ml centrifuge tube and centrifuged at 5,000 RCF for 5 min to pellet the debris. The supernatant was decanted and passed through a 0.22 µm Sterivex filter. The filtrate was used as a phage source and stored between 2 to 8℃.

Host bacteria cultivation

The bacterial strain, *B. cellulosilyticus* WH2 (45) was used as the bacterial host and was received as glycerol stocks from Washington University, St. Louis. Bacteria were grown in brain-heart infusion media supplemented with 2mM $MgSO_{4,}$ and 10mM $MgCl_2$ we denote as BHISMg. Culture plates were supplemented with 1.5% agar and incubated at 37℃ for 48 hours under anaerobic conditions with 5% H_2, 5% $CO_{2,}$ and 90% N_2. Following incubation, a sterile loop was used to transfer an isolated colony into a 12 hr deoxygenated BHISMg broth. Following anaerobic incubation at 37℃ for 24 hours the liquid cultures were further subcultured into another BHISMg broth and incubated overnight.

Plaque assays

Plaque assays were performed with minor modifications described below. Before beginning the plaque assays, BHISMg plates were deoxygenated for 12 hrs in the anaerobic chamber and pre-warmed before use. For top agar plates, cooled molten BHISMg with 0.7 % agar was inoculated with 500 µl of bacteria, and between 2µl and 50 µl of processed phage influent was overlayed onto BHISMg plates. The plates were cooled for 15 minutes before incubating at 37°C for up to five 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 three hours of incubation. Post incubation, tubes were centrifuged at 4500 *x* g for 5 min, and the supernatant was collected and concentrated using a 50,000 kDa MWCO Vivaspin ultrafiltration unit (Sartorius). Phage lysate was stored at 4°C.

Titering enumeration

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

Viral DNA extraction and sequencing

DNA extractions were performed using the Phage DNA isolation kit (Norgen) as per manufacturer instructions. In short, 1 ml of phage lysate was DNase I treated, lysed, and treated with Proteinase K. The sample was added to a spin column and washed three times. DNA was eluted in 75 µl of the elution buffer. The second elution recommended by the kit was not performed. The DNA obtained was quantified using a Qubit 1x dsDNA high-sensitivity assay kit (Invitrogen, Life Technologies). For Oxford Nanopore MinION sequencing the library preparation and sequencing were performed using Oxford Nanopore Rapid Barcoding Sequencing Kit (SQK-RBK0004). 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 as described in (46). Eight of the samples were sequenced on both Nanopore and Illumina sequencing platforms (Bc01 to Bc08), four were sequenced only on Nanopore platforms (Bc09 to Bc12), and the remaining four were sequenced only on Illumina platforms (Bc13 to Bc16). 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. The reads were then processed with Filtlong v.0.2.20 (47) to remove reads less than 1,000 bp in length and exclude 5% of the lowest-quality reads. Similarly, Illumina sequences were processed with prinseq++ (48), filtering reads less than 60 bp in length, reads with quality scores less than 25, and exact duplicates.

Genome assembly

The steps for genome assembly are available as a pipeline based on Snakemake using Snaketool (49) and available on GitHub (50).

Nanopore reads were assembled using Unicycler (51) and Flye (52), while Illumina reads were assembled using MEGAHIT (53). These assemblers were selected as they provide assembly graphs of the contigs assembled, which would be utilized for completing fragmented genome assemblies, (54–58). To assess the quality of the assembly, the resulting contigs were processed with ViralVerify to detect viral contigs (59), read coverage was calculated using CoverM (60), and whether the contigs were fragmented using assembly graph information. The assembly graph includes details of connecting unitigs (high-quality contigs) that represent the longest non-branching paths joined together to form contigs.

Unitigs that were greater than 90 kb, identified as viral, with the highest read coverage, and described as complete (CheckV) were selected from each assembly. For each genome, one representative unitig (higher quality contigs) was selected (circular, longest contig with high read coverage) per sample, as the complete phage assembly. In the end, the assemblies were polished with high-coverage Illumina reads using Polca to reduce sequencing-related errors (61).

Taxonomic and functional annotation

The isolates in this study were processed with CrassUS (62). CrassUS was developed specifically for annotating genomes from the Crassvirales order, incorporating a focused database including known crAssphage genomes. This program generates a table of taxa annotations, functional annotations, the presence of direct terminal repeats (DTR), and average nucleotide identities of the most similar reference genomes. Taxonomic annotations from CrassUS were used, as they follow the ICTV (27,29) Crassvirales order demarcation criteria to determine taxonomy. The three conserved genes across crAssphages including major capsid protein (MCP), portal protein, and terminase large subunit (*terL)* were used to build the phylogenetic trees. The phylogenetic trees were plotted in iTol (63). The predicted genes and their arrangement across species were visualized using clinker plots (64), after re-circularizing the genes to start at the *terL* in order to examine synteny across the phage genomes assembled. Finally, phages may encode their own tRNA genes escaping the host translation machinery, and these were predicted with tRNA-scanSE (65).

Transmission electron microscopy imaging

Bacteroides phages were grown using the phage overlay method described above. Phage lysates were diluted 1:10, and 5 µL of the diluted phage lysate was applied to a plasma-cleaned grid for two minutes at room temperature. The grids were formvar and carbon coated 200 mesh grids and they were plasma cleaned using the Gatan (Solarus) Advanced plasma system for 30 sec 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 minute. 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 120kV at a magnification of

49,000x and the images were recorded on an AMT Nanosprint 15 digital camera using software v7.0.1.

Phage measurements were calculated using the ImageJ software (66). The capsid diameter was measured by obtaining the diameter of the circle circumscribing the capsid, such that the more distant vertices of the projected capsid contacted the circle (Figure S3). For the tail, the length 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 were not calculated (Figure S3). Average measurements from 5 phages were calculated and reported. The TEM image was further edited for publication using the GNU Image Manipulation Program (GIMP) (67).

Evolutionary analyses

The 14 crAssphage isolated and sequenced in this study and the four reference pure culture isolates, *K. primarius* (crAss001), *W. bangladeshii* (DAC15), *Wulfhauvirus isolate* (DAC17), and *J. secundus* (crAss002) were assessed together for this analysis. Orthologous genes were identified from genes predicted from the above 18 genomes, using Orthofinder (68). Orthogroups that included genes that were only present in phages from the host—*B. cellulosilyticus* WH2—were assessed further for signatures of host specificity.

These orthogroups were aligned using Muscle (69) codon-based multiple sequence alignment in MEGA11 (70). To test for codon-based positive selection, we calculated the probability of rejecting the null hypothesis of strict neutrality ($d_N = d_S$), and in favour of the alternate hypothesis ($d_N > d_S$). The d_N/d_S values were calculated from the MSA using MEGA v.11.0 (71), with the Li-Wu-Luo method (72). The variance of the difference was computed using bootstraps, set to 100 replicates.

Predicting proteins 3D structure and docking

The 3D structures of the proteins from *K. winsdale* (Bc01), *K. frurule* (Bc03), and *R. redwords* (Bc11) were predicted using Colabfold version 1.4.0 (35) on the Gadi server at the National Computational Infrastructure (NCI). 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 (73). All protein pairs were docked using hdock-lite version 1.1 (74) on the Gadi server. The results from hdock were sorted based on the binding score in the output file to identify the highest-quality binding predictions for each phage protein. The 3D structure of the proteins were visualized using PyMOL.

Data availability

The samples were submitted to Sequence Read Archive within the project, PRJNA737576. *B. cellulosilyticus WH2*, *K. winsdale* (Bc01), *K.frurule* (Bc03), and *R. redwords* (Bc11) are all available on Genbank with accessions QQ198717 (Bc01), QQ198718 (Bc03), and QQ198719 (Bc11), and we are working on making them available through ATCC. The 3D protein structures for the three crAssphage genomes are available to download at doi.org/10.25451/flinders.21946034.

Author Contributions

BNP performed bioinformatics analysis, wrote the paper, analyzed the genomes. AV, CS, CC, WP, NJ SG collected samples, isolated and cultured phages. MFM, CS, MA, KP, LD sequenced phages. CL, SG took TEM images. KM, MR, PD, SRG, VM, GB, AL, SH performed bioinformatics analysis. AMS, EAD, RAE conceived the project, performed the bioinformatics, and wrote the paper with input from all authors.

Acknowledgments

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.

Funding

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.

References

1. Hou K, Wu Z-X, Chen X-Y, Wang J-Q, Zhang D, Xiao C, et al. Microbiota in health and diseases. Signal Transduct Target Ther [Internet]. 2022 Apr 23;7(1):135. Available from: http://dx.doi.org/10.1038/s41392-022-00974-4

2. Integrative HMP (iHMP) Research Network Consortium. The Integrative Human Microbiome Project. Nature [Internet]. 2019 May;569(7758):641–8. Available from: http://dx.doi.org/10.1038/s41586-019-1238-8

3. Shamash M, Maurice CF. Phages in the infant gut: a framework for virome development during early life. ISME J [Internet]. 2022 Feb;16(2):323–30. Available

from: http://dx.doi.org/10.1038/s41396-021-01090-x

4. Hugenholtz P, Goebel BM, Pace NR. Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. J Bacteriol [Internet]. 1998 Sep;180(18):4765–74. Available from: http://dx.doi.org/10.1128/JB.180.18.4765-4774.1998

5. Pace NR, Stahl DA, Lane DJ, Olsen GJ. The Analysis of Natural Microbial Populations by Ribosomal RNA Sequences. In: Marshall KC, editor. Advances in Microbial Ecology [Internet]. Boston, MA: Springer US; 1986. p. 1–55. Available from: https://doi.org/10.1007/978-1-4757-0611-6_1

6. Inglis LK, Edwards RA. How Metagenomics Has Transformed Our Understanding of Bacteriophages in Microbiome Research. Microorganisms [Internet]. 2022 Aug 19;10(8). Available from: http://dx.doi.org/10.3390/microorganisms10081671

7. Roach M, Beecroft S, Mihindukulasuriya KA, Wang L, Lima LFO, Dinsdale EA, et al. Hecatomb: An End-to-End Research Platform for Viral Metagenomics [Internet]. bioRxiv. 2022 [cited 2022 Nov 22]. p. 2022.05.15.492003. Available from: https://www.biorxiv.org/content/biorxiv/early/2022/05/16/2022.05.15.492003

8. Hesse RD, Roach M, Kerr EN, Papudeshi B, Lima LFO, Goodman AZ, et al. Phage Diving: An Exploration of the Carcharhinid Shark Epidermal Virome. Viruses [Internet]. 2022 Sep 5;14(9). Available from: http://dx.doi.org/10.3390/v14091969

9. Anthenelli M, Jasien E, Edwards R, Bailey B, Felts B, Katira P, et al. Phage and bacteria diversification through a prophage acquisition ratchet [Internet]. bioRxiv. 2020 [cited 2022 Jul 19]. p. 2020.04.08.028340. Available from: https://www.biorxiv.org/content/10.1101/2020.04.08.028340v1

10. Knowles B, Silveira CB, Bailey BA, Barott K, Cantu VA, Cobián-Güemes AG, et al. Lytic to temperate switching of viral communities. Nature [Internet]. 2016 Mar 24;531(7595):466–70. Available from: http://dx.doi.org/10.1038/nature17193

11. Chevallereau A, Pons BJ, van Houte S, Westra ER. Interactions between bacterial and phage communities in natural environments. Nat Rev Microbiol [Internet]. 2022 Jan;20(1):49–62. Available from: http://dx.doi.org/10.1038/s41579-021-00602-y

12. Silveira CB, Luque A, Rohwer F. The landscape of lysogeny across microbial community density, diversity and energetics. Environ Microbiol [Internet]. 2021 Aug;23(8):4098–111. Available from: http://dx.doi.org/10.1111/1462-2920.15640

13. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, et al. A human gut microbial gene catalogue established by metagenomic sequencing. Nature [Internet]. 2010 Mar 4;464(7285):59–65. Available from: http://dx.doi.org/10.1038/nature08821

14. HMP Consortium. Structure, function and diversity of the healthy human

microbiome. Nature [Internet]. 2012 Jun 13 [cited 2022 Jul 25];486(7402):207–14. Available from: https://www.nature.com/articles/nature11234

15. Pargin E, Roach M, Skye A, Edwards R, Giles S. The human gut virome: Composition, colonisation, interactions, and impacts on human health [Internet]. 2022. Available from: http://dx.doi.org/10.31219/osf.io/s9px2

16. Yutin N, Benler S, Shmakov SA, Wolf YI, Tolstoy I, Rayko M, et al. Analysis of metagenome-assembled viral genomes from the human gut reveals diverse putative CrAss-like phages with unique genomic features. Nat Commun [Internet]. 2021 Feb 16;12(1):1044. Available from: http://dx.doi.org/10.1038/s41467-021-21350-w

17. Edwards RA, Vega AA, Norman HM, Ohaeri M, Levi K, Dinsdale EA, et al. Global phylogeography and ancient evolution of the widespread human gut virus crAssphage. Nat Microbiol [Internet]. 2019 Oct;4(10):1727–36. Available from: http://dx.doi.org/10.1038/s41564-019-0494-6

18. Rossi A, Treu L, Toppo S, Zschach H, Campanaro S, Dutilh BE. Evolutionary Study of the Crassphage Virus at Gene Level. Viruses [Internet]. 2020 Sep 17;12(9). Available from: http://dx.doi.org/10.3390/v12091035

19. Norman JM, Handley SA, Baldridge MT, Droit L, Liu CY, Keller BC, et al. Disease-specific alterations in the enteric virome in inflammatory bowel disease. Cell [Internet]. 2015 Jan 29;160(3):447–60. Available from: http://dx.doi.org/10.1016/j.cell.2015.01.002

20. Dutilh BE, Cassman N, McNair K, Sanchez SE, Silva GGZ, Boling L, et al. A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. Nat Commun [Internet]. 2014 Jul 24;5:4498. Available from: http://dx.doi.org/10.1038/ncomms5498

21. Guerin E, Shkoporov A, Stockdale SR, Clooney AG, Ryan FJ, Sutton TDS, et al. Biology and Taxonomy of crAss-like Bacteriophages, the Most Abundant Virus in the Human Gut. Cell Host Microbe [Internet]. 2018 Nov 14;24(5):653-664.e6. Available from: http://dx.doi.org/10.1016/j.chom.2018.10.002

22. Shkoporov AN, Khokhlova EV, Fitzgerald CB, Stockdale SR, Draper LA, Ross RP, et al. ΦCrAss001 represents the most abundant bacteriophage family in the human gut and infects Bacteroides intestinalis. Nat Commun [Internet]. 2018 Nov 14;9(1):4781. Available from: http://dx.doi.org/10.1038/s41467-018-07225-7

23. Hryckowian AJ, Merrill BD, Porter NT, Van Treuren W, Nelson EJ, Garlena RA, et al. Bacteroides thetaiotaomicron-Infecting Bacteriophage Isolates Inform Sequence-Based Host Range Predictions. Cell Host Microbe [Internet]. 2020 Sep 9;28(3):371-379.e5. Available from: http://dx.doi.org/10.1016/j.chom.2020.06.011

24. Guerin E, Shkoporov AN, Stockdale SR, Comas JC, Khokhlova EV, Clooney AG, et

al. Isolation and characterisation of ΦcrAss002, a crAss-like phage from the human gut that infects Bacteroides xylanisolvens. Microbiome [Internet]. 2021 Apr 12;9(1):89. Available from: http://dx.doi.org/10.1186/s40168-021-01036-7

25. Shkoporov AN, Khokhlova EV, Stephens N, Hueston C, Seymour S, Hryckowian AJ, et al. Long-term persistence of crAss-like phage crAss001 is associated with phase variation in Bacteroides intestinalis. BMC Biol [Internet]. 2021 Aug 18;19(1):163. Available from: http://dx.doi.org/10.1186/s12915-021-01084-3

26. Antson A, Bayfield O, Shkoporov A, Yutin N, Khokhlova E, Smith J, et al. Structural atlas of the most abundant human gut virus [Internet]. Research Square. 2022. Available from: https://www.researchsquare.com/article/rs-1898492/v1

27. Shkoporov AN, Stockdale SR, Adriaenssens EM, Yutin N, Koonin EV, Dutilh BE, et al. Create one new order (Crassvirales) including four new families, ten new subfamilies, 42 new genera and 73 new species (Caudoviricetes) [Internet]. 2021. Available from: https://ictv.global/ictv/proposals/2021.022B.R.Crassvirales.zip

28. Dutilh BE, Varsani A, Tong Y, Simmonds P, Sabanadzovic S, Rubino L, et al. Perspective on taxonomic classification of uncultivated viruses. Curr Opin Virol [Internet]. 2021 Dec;51:207–15. Available from: http://dx.doi.org/10.1016/j.coviro.2021.10.011

29. Walker PJ, Siddell SG, Lefkowitz EJ, Mushegian AR, Adriaenssens EM, Alfenas-Zerbini P, et al. Recent changes to virus taxonomy ratified by the International Committee on Taxonomy of Viruses (2022). Arch Virol [Internet]. 2022 Nov;167(11):2429–40. Available from: http://dx.doi.org/10.1007/s00705-022-05516-5

30. Borges AL, Lou YC, Sachdeva R, Al-Shayeb B, Jaffe AL, Lei S, et al. Stop codon recoding is widespread in diverse phage lineages and has the potential to regulate translation of late stage and lytic genes [Internet]. bioRxiv. 2021 [cited 2022 Jul 23]. p. 2021.08.26.457843. Available from: https://www.biorxiv.org/content/biorxiv/early/2021/08/26/2021.08.26.457843

31. Ivanova NN, Schwientek P, Tripp HJ, Rinke C, Pati A, Huntemann M, et al. Stop codon reassignments in the wild. Science [Internet]. 2014 May 23;344(6186):909–13. Available from: http://dx.doi.org/10.1126/science.1250691

32. Crisci MA, Chen L-X, Devoto AE, Borges AL, Bordin N, Sachdeva R, et al. Closely related Lak megaphages replicate in the microbiomes of diverse animals. iScience [Internet]. 2021 Aug 20;24(8):102875. Available from: http://dx.doi.org/10.1016/j.isci.2021.102875

33. Peters SL, Borges AL, Giannone RJ, Morowitz MJ, Banfield JF, Hettich RL. Experimental validation that human microbiome phages use alternative genetic coding. Nat Commun [Internet]. 2022 Sep 29;13(1):5710. Available from: http://dx.doi.org/10.1038/s41467-022-32979-6

34. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol [Internet]. 1990 Oct 5;215(3):403–10. Available from: http://dx.doi.org/10.1016/S0022-2836(05)80360-2

35. Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. ColabFold: making protein folding accessible to all. Nat Methods [Internet]. 2022 Jun;19(6):679–82. Available from: http://dx.doi.org/10.1038/s41592-022-01488-1

36. Pollet RM, Martin LM, Koropatkin NM. TonB-dependent transporters in the Bacteroidetes: Unique domain structures and potential functions. Mol Microbiol [Internet]. 2021 Mar;115(3):490–501. Available from: http://dx.doi.org/10.1111/mmi.14683

37. Luque A, Benler S, Lee DY, Brown C, White S. The Missing Tailed Phages: Prediction of Small Capsid Candidates. Microorganisms [Internet]. 2020 Dec 8;8(12). Available from: http://dx.doi.org/10.3390/microorganisms8121944

38. Lee DY, Bartels C, McNair K, Edwards RA, Swairjo MA, Luque A. Predicting the capsid architecture of phages from metagenomic data. Comput Struct Biotechnol J [Internet]. 2022 Jan 5;20:721–32. Available from: http://dx.doi.org/10.1016/j.csbj.2021.12.032

39. Spilman MS, Damle PK, Dearborn AD, Rodenburg CM, Chang JR, Wall EA, et al. Assembly of bacteriophage 80α capsids in a Staphylococcus aureus expression system. Virology [Internet]. 2012 Dec 20;434(2):242–50. Available from: http://dx.doi.org/10.1016/j.virol.2012.08.031

40. Dearborn AD, Wall EA, Kizziah JL, Klenow L, Parker LK, Manning KA, et al. Competing scaffolding proteins determine capsid size during mobilization of Staphylococcus aureus pathogenicity islands. Elife [Internet]. 2017 Oct 6;6. Available from: http://dx.doi.org/10.7554/eLife.30822

41. Casjens SR, Gilcrease EB. Determining DNA packaging strategy by analysis of the termini of the chromosomes in tailed-bacteriophage virions. Methods Mol Biol [Internet]. 2009;502:91–111. Available from: http://dx.doi.org/10.1007/978-1-60327-565-1_7

42. Weinheimer AR, Aylward FO. A distinct lineage of Caudovirales that encodes a deeply branching multi-subunit RNA polymerase. Nat Commun [Internet]. 2020 Sep 9;11(1):4506. Available from: http://dx.doi.org/10.1038/s41467-020-18281-3

43. Nobrega FL, Vlot M, de Jonge PA, Dreesens LL, Beaumont HJE, Lavigne R, et al. Targeting mechanisms of tailed bacteriophages. Nat Rev Microbiol [Internet]. 2018 Dec;16(12):760–73. Available from: http://dx.doi.org/10.1038/s41579-018-0070-8

44. Brown BP, Chopera D, Havyarimana E, Wendoh J, Jaumdally S, Nyangahu DD, et al. crAssphage genomes identified in fecal samples of an adult and infants with evidence of positive genomic selective pressure within tail protein genes. Virus Res

[Internet]. 2021 Jan 15;292:198219. Available from: http://dx.doi.org/10.1016/j.virusres.2020.198219

45. McNulty NP, Wu M, Erickson AR, Pan C, Erickson BK, Martens EC, 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 [Internet]. 2013 Aug 20;11(8):e1001637. Available from: http://dx.doi.org/10.1371/journal.pbio.1001637

46. Kim AH, Armah G, Dennis F, Wang L, Rodgers R, Droit L, et al. Enteric virome negatively affects seroconversion following oral rotavirus vaccination in a longitudinally sampled cohort of Ghanaian infants. Cell Host Microbe [Internet]. 2022 Jan 12;30(1):110-123.e5. Available from: http://dx.doi.org/10.1016/j.chom.2021.12.002

47. Wick RR. Filtlong: Tool for filtering long reads by quality [Internet]. 2018. Available from: https://github.com/rrwick/Filtlong/

48. Cantu VA, Sadural J, Edwards R. PRINSEQ++, a multi-threaded tool for fast and efficient quality control and preprocessing of sequencing datasets [Internet]. PeerJ Preprints; 2019 Feb [cited 2022 Jul 24]. Report No.: e27553v1. Available from: https://peerj.com/preprints/27553v1/

49. Roach M, Pierce-Ward NT, Suchecki R, Mallawaarachchi V, Papudeshi B, Handley SA, et al. Ten simple rules and a template for creating workflows-as-applications [Internet]. 2022. Available from: http://dx.doi.org/10.31219/osf.io/8w5j3

50. Papudeshi B, Mallawaarachchi V, Roach M, Edwards R. spae: Phage genome assembly and annotation [Internet]. 2022. Available from: https://github.com/linsalrob/spae

51. Wick RR, Judd LM, Gorrie CL, Holt KE. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. PLoS Comput Biol [Internet]. 2017 Jun;13(6):e1005595. Available from: http://dx.doi.org/10.1371/journal.pcbi.1005595

52. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol [Internet]. 2019 May;37(5):540–6. Available from: http://dx.doi.org/10.1038/s41587-019-0072-8

53. Li D, Luo R, Liu C-M, Leung C-M, Ting H-F, Sadakane K, et al. MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. Methods [Internet]. 2016 Jun 1;102:3–11. Available from: http://dx.doi.org/10.1016/j.ymeth.2016.02.020

54. Wick RR, Schultz MB, Zobel J, Holt KE. Bandage: interactive visualization of de novo genome assemblies. Bioinformatics [Internet]. 2015 Oct 15;31(20):3350–2. Available from: http://dx.doi.org/10.1093/bioinformatics/btv383

55. Bruce T, De Wardt R, Papudeshi B, Robinson B, Cuevas DA, Aguinaldo K, et al. The intersection of genotype and phenotype: Informing ecotype development.

56. Mallawaarachchi V, Wickramarachchi A, Lin Y. GraphBin: refined binning of metagenomic contigs using assembly graphs. Bioinformatics [Internet]. 2020 Jun 1;36(11):3307–13. Available from: http://dx.doi.org/10.1093/bioinformatics/btaa180

57. Mallawaarachchi VG, Lin Y. Metacoag: Binning metagenomic contigs via composition, coverage and assembly graphs. Res Comput Mol Biol [Internet]. 2022; Available from: https://link.springer.com/chapter/10.1007/978-3-031-04749-7_5

58. Mallawaarachchi VG, Wickramarachchi AS, Lin Y. GraphBin2: refined and overlapped binning of metagenomic contigs using assembly graphs. WASA [Internet]. 2020; Available from: https://drops.dagstuhl.de/opus/volltexte/2020/12797/

59. Raiko M. viralVerify: viral contig verification tool [Internet]. 2021. Available from: https://github.com/ablab/viralVerify

60. Woodcroft BJ. CoverM:DNA read coverage and relative abundance calculator [Internet]. 2021. Available from: https://github.com/wwood/CoverM

61. Zimin AV, Salzberg SL. The genome polishing tool POLCA makes fast and accurate corrections in genome assemblies. PLoS Comput Biol [Internet]. 2020 Jun;16(6):e1007981. Available from: http://dx.doi.org/10.1371/journal.pcbi.1007981

62. Carrillo D. CrassUS - Crassvirales Uncovering Software [Internet]. 2022. Available from: https://github.com/dcarrillox/CrassUS

63. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic Acids Res [Internet]. 2019 Jul 2;47(W1):W256–9. Available from: http://dx.doi.org/10.1093/nar/gkz239

64. Gilchrist CLM, Chooi Y-H. Clinker & clustermap.js: Automatic generation of gene cluster comparison figures. Bioinformatics [Internet]. 2021 Jan 18; Available from: http://dx.doi.org/10.1093/bioinformatics/btab007

65. Chan PP, Lowe TM. tRNAscan-SE: Searching for tRNA Genes in Genomic Sequences. Methods Mol Biol [Internet]. 2019;1962:1–14. Available from: http://dx.doi.org/10.1007/978-1-4939-9173-0_1

66. Schneider CA, Rasband WS, Eliceiri KW. NIH Image to ImageJ: 25 years of image analysis. Nat Methods [Internet]. 2012 Jul;9(7):671–5. Available from: http://dx.doi.org/10.1038/nmeth.2089

67. The GIMP Development Team. GIMP [Internet]. 2019. Available from: https://www.gimp.org

68. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol [Internet]. 2019 Nov 14;20(1):238. Available from: http://dx.doi.org/10.1186/s13059-019-1832-y

69. Edgar RC. High-accuracy alignment ensembles enable unbiased assessments of sequence homology and phylogeny [Internet]. bioRxiv. 2022 [cited 2022 Aug 30]. p. 2021.06.20.449169. Available from: https://www.biorxiv.org/content/10.1101/2021.06.20.449169v2

70. Stecher G, Tamura K, Kumar S. Molecular Evolutionary Genetics Analysis (MEGA) for macOS. Mol Biol Evol [Internet]. 2020 Apr 1;37(4):1237–9. Available from: http://dx.doi.org/10.1093/molbev/msz312

71. Tamura K, Stecher G, Kumar S. MEGA11: Molecular Evolutionary Genetics Analysis Version 11. Mol Biol Evol [Internet]. 2021 Jun 25;38(7):3022–7. Available from: http://dx.doi.org/10.1093/molbev/msab120

72. 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 [Internet]. 1985 Mar;2(2):150–74. Available from: http://dx.doi.org/10.1093/oxfordjournals.molbev.a040343

73. Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res [Internet]. 2022 Jan 7;50(D1):D439–44. Available from: http://dx.doi.org/10.1093/nar/gkab1061

74. Yan Y, Tao H, He J, Huang S-Y. The HDOCK server for integrated protein-protein docking. Nat Protoc [Internet]. 2020 May;15(5):1829–52. Available from: http://dx.doi.org/10.1038/s41596-020-0312-x

Supplementary Figures

Figure S1: Genome coverage of the sequenced reads across samples. A) Genome coverage in Nanopore sequenced reads, B) Genome coverage in Illumina sequenced reads. The samples with * in red were not sequenced on the platform, and * in black didn't produce a complete phage contig

Figure S2: Showing the taxa classification of the three novel species remains consistent across the three conserved proteins A) Major capsid protein (MCP), B) portal protein C) terminase large subunit (*terL*). The outgroup across all three trees set to *Cellulophaga phage phi13:2.* The placement of the three novel species are highlighted on the tree, Bc01 belonging to *Kehishuvirus* genera (light blue), Bc03 belonging to *Kolpuevirus* genera (purple), and Bc11 belonging to a novel genus named *Rudgehvirus* (brown).

Figure S3**:** TEM phage measurements were taken for A) Capsid diameter, by drawing a circle around the polygon with the edges within the circle. The diameter of this circle was measured and represented as the capsid diameter. B) For tail length, a line was drawn from the base of the capsid to the visible edge of the tail fibers. This was repeated over five phages of the same sample and an average with standard deviation was calculated across all of them

Figure S4: TEM measurements of capsids of A) *K. primarius* (crAss001) image from (Shkoporov et al. 2018) measured to be 81 ± 2 nm and B) *J. secundus* (crAss002) images from (Guerin et al. 2021) measured to be 75 ± 3 nm using ImageJ software. The scale bar on both figures represents 100nm.

Supplementary Tables

Table S1: Phage genome assembly overview

Table S2: Coding density and search for tRNA suppressors within the three Bc genomes

Table S3: *Kehishuvirus winsdale* (Bc01) functional annotation from crassus

Table S4: *Kolpuevirus frurule* (Bc03) functional annotation from crassus

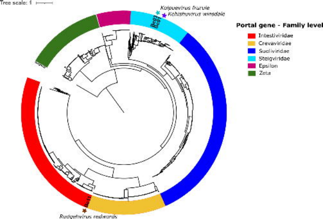
Table S5: *Rudgehvirus redwords* (Bc11) functional annotation from crassus

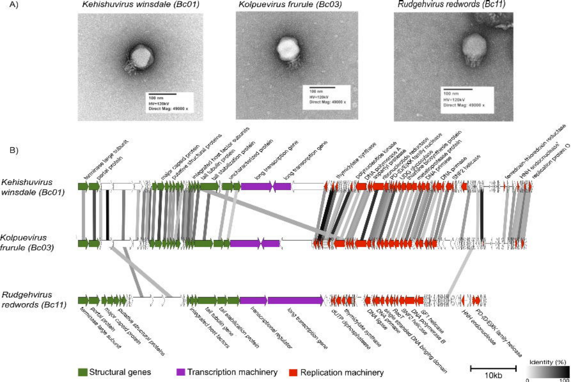
Table S6: Orthologous groups identified across the 18 crAssphage isolates, highlighted the two orthogroups that are present within the 14 crAssphage isolates from this study, infecting the same bacterial host.

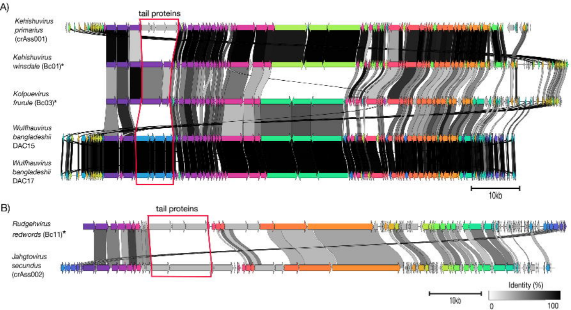
Table S7: Summary of the Kehishuvirus winsdale (Bc01) proteins interacting with *Bacteroides cellulosilyticus WH2* proteins using hdock-lite.

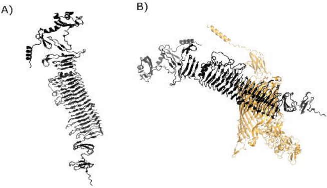
Genome	***length (bp)***	***GC %***	***Coding density***	***Number of CDS***	***Unknown function***	***tRNA***	***DTR***	***Taxa***
Bc01	100,841	35.09	91.84	104	58	24	False	*Kehishuvirus winsdale*
Bc03	99,523	33.00	92.06	108	63	5	False	*Kolpuevirus frurule*
Bc11	90,575	29.15	87.45	84	48	0	False	*Rudgehvirus redwords*

ORTHOGROUP	Gene Function	Number Of Sequences	$D_{n/}D_S$	P Value	Z Statistic
OG0000000	Structural protein (gp23)	20	0.96	0.29	0.56
OG0000008*	Tail spike protein (gp22)	14	0.26	<0.001	9.15


Tree scale: 1
Kolpuevirus frurule
Kehishuvirus winsdale
Bc03
Bc01
Bc11
Rudgehvirus redwords
Portal gene - Family level
🟥 Intestiviridae
🟨 Crevaviridae
🟦 Suoliviridae
🟦 Steigviridae
🟪 Epsilon
🟩 Zeta


A)
Kehishuvirus winsdale (Bc01)
Kolpuevirus frurule (Bc03)
Rudgehvirus redwords (Bc11)
100 nm
HV=120kV
Direct Mag: 49000 x
B)
Kehishuvirus winsdale (Bc01)
Kolpuevirus frurule (Bc03)
Rudgehvirus redwords (Bc11)
terminase large subunit
portal protein
major capsid protein
putative structural proteins
integrated host factor subunits
tail tubulin protein
tail stabilization protein
uncharacterized protein
long transcription gene
long transcription gene
thymidylate synthase
polynucleotide kinase
DNA polymerase A
aspartyl protease
ribonucleotide reductase
PD-(D/E)XK family nuclease
UDG glycosylase
thiamine biosynthesis protein
metalloprotease protein
DNA primase
DNA primase
SNF2 helicase
ferredoxin-thioredoxin reductase
HNH endonuclease
replication protein O
terminase large subunit
portal protein
major capsid protein
putative structural proteins
integrated host factors
tail tubulin gene
tail stabilization protein
transcriptional regulator
long transcription gene
dUTP diphosphatase
thymidylate synthase
DNA ligase
DNA primase
single stranded DNA binging domain
RecT
SNF2 helicase
DNA polymerase B
SF1 helicase
HNH endonuclease
PD-(D/E)XK family helicase
🟩 Structural genes
🟪 Transcription machinery
🟥 Replication machinery
10kb
Identity (%)
0
100


A)
tail proteins
Kehishuvirus primarius (crAss001)
Kehishuvirus winsdale (Bc01)*
Kolpuevirus frurule (Bc03)*
Wulfhauvirus bangladeshii DAC15
Wulfhauvirus bangladeshii DAC17
10kb
B)
tail proteins
Rudgehvirus redwords (Bc11)*
Jahgtovirus secundus (crAss002)
10kb
Identity (%)
0
100


A)
B)