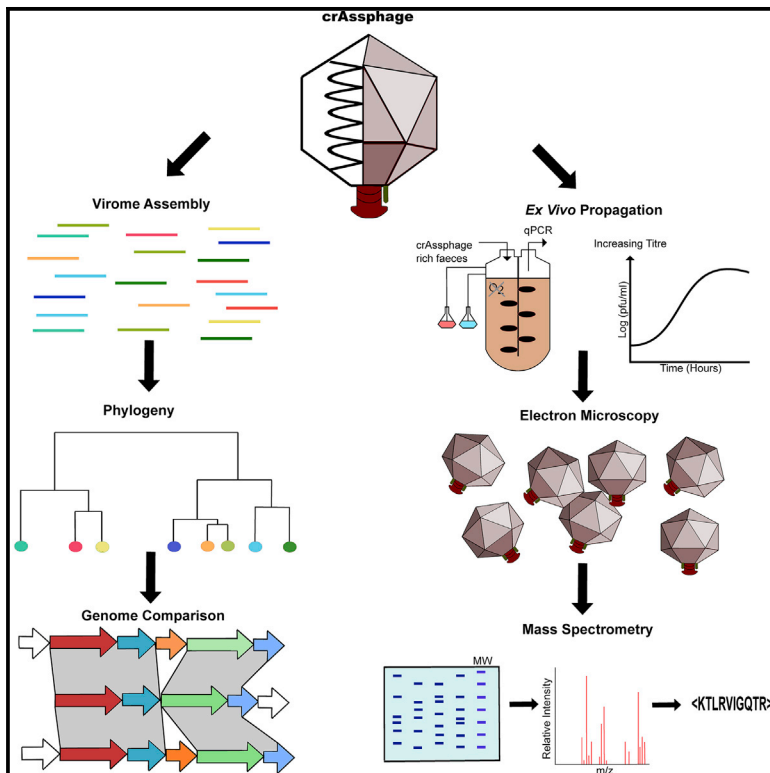


Cell Host & Microbe

Biology and Taxonomy of crAss-like Bacteriophages, the Most Abundant Virus in the Human Gut

Graphical Abstract



Authors

Emma Guerin, Andrey Shkoporov, Stephen R. Stockdale, ..., Enrique Gonzalez-Tortuero, R. Paul Ross, Colin Hill

Correspondence

c.hill@ucc.ie

In Brief

CrAssphage is the most abundant human gut-associated virus. Guerin et al. identify 249 crAss-like phage genomes and classify them into four subfamilies and ten candidate genera that differ among human populations. These *in silico* predictions are combined with *ex vivo* propagations, electron microscopy imaging, and mass spectrometry detection.

Highlights

- Screening of human fecal metagenomic samples reveals 249 crAss-like phage genomes
- The crAss-like phages were classified into 4 subfamilies composed of 10 candidate genera
- A crAss-like phage was propagated in *ex vivo* human fecal fermentations
- Short-tailed phage virions could be visualized by electron microscopy



Biology and Taxonomy of crAss-like Bacteriophages, the Most Abundant Virus in the Human Gut

Emma Guerin,^{1,2,6} Andrey Shkoporov,^{1,6} Stephen R. Stockdale,^{1,3,6} Adam G. Clooney,¹ Feargal J. Ryan,^{1,4} Thomas D.S. Sutton,^{1,2} Lorraine A. Draper,¹ Enrique Gonzalez-Tortuero,^{1,5} R. Paul Ross,^{1,2,3} and Colin Hill^{1,2,7,*}

¹APC Microbiome Ireland, University College Cork, Cork, Ireland

²School of Microbiology, University College Cork, Cork, Ireland

³Teagasc Food Research Centre, Moorepark, Fermoy, Co., Cork, Ireland

⁴Present address: South Australian Health and Medical Research Institute, North Terrace, Adelaide, SA 5000, Australia

⁵Present address: Institute for Genome Sciences, University of Maryland Baltimore School of Medicine, 670 West Baltimore Street, Baltimore, MD 21201, USA

⁶These authors contributed equally

⁷Lead Contact

*Correspondence: c.hill@ucc.ie

<https://doi.org/10.1016/j.chom.2018.10.002>

SUMMARY

CrAssphages represent the most abundant virus in the human gut microbiota, but the lack of available genome sequences for comparison has kept them enigmatic. Recently, sequence-based classification of distantly related crAss-like phages from multiple environments was reported, leading to a proposed familial-level taxonomic group. Here, we assembled the metagenomic sequencing reads from 702 human fecal virome/phageome samples and analyzed 99 complete circular crAss-like phage genomes and 150 contigs ≥ 70 kb. *In silico* comparative genomics and taxonomic analysis enabled a classification scheme of crAss-like phages from human fecal microbiomes into four candidate subfamilies composed of ten candidate genera. Laboratory analysis was performed on fecal samples from an individual harboring seven distinct crAss-like phages. We achieved crAss-like phage propagation in *ex vivo* human fecal fermentations and visualized short-tailed podoviruses by electron microscopy. Mass spectrometry of a crAss-like phage capsid protein could be linked to metagenomic sequencing data, confirming crAss-like phage structural annotations.

INTRODUCTION

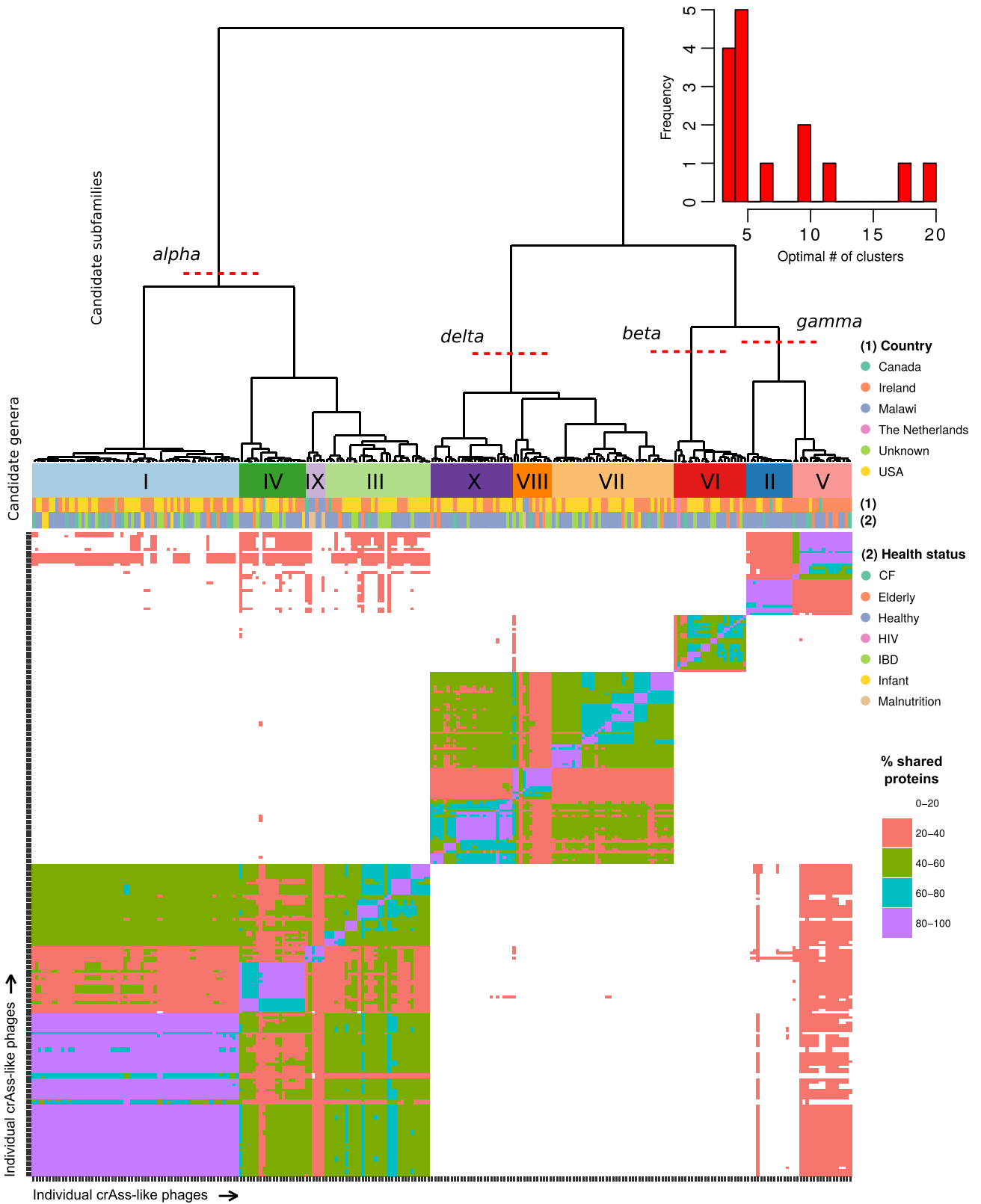
In recent years, increasing numbers of bacteria, archaea, fungi, protists, and viruses residing on and within the human body have been associated with various states of human health and disease, including diet, age, weight, inflammatory bowel disease (IBD), diabetes, and cognition (Minot et al., 2011; Claesson et al., 2012; Cryan and Dinan, 2014; Everard and Cani, 2013; Frank et al., 2011; Norman et al., 2015; Reyes et al., 2010; Tremaroli

and Bäckhed, 2012). A relatively small number of eukaryotic viruses present in the gastrointestinal tract can target human cells; however, much larger and more complex populations of viruses that target bacteria (bacteriophages or phages) are also present. The role of phages in the gut has been a subject of increased interest as initial investigations revealed substantial differences in phage populations between healthy and diseased cohorts (Manrique et al., 2016, 2017; Mills et al., 2013; Norman et al., 2015; Reyes et al., 2015). It is likely that phages have an important role in shaping our gut microbiome, but their precise role remains poorly understood.

In 2014, metagenomic studies of the viral fraction of the human gut microbiota identified a DNA phage, crAssphage, detectable in approximately 50% of individuals from specific human populations and reaching up to 90% of the total viral DNA load in feces of certain individuals (Dutilh et al., 2014). Dutilh and colleagues noted that crAssphage had been overlooked in previous metagenomic studies as the vast majority of its genes do not match known sequences present in databases. Based on host co-occurrence and CRISPR spacer profiling, it was predicted that prototypical crAssphage infects bacteria of the genus *Bacteroides* or other members of the *Bacteroidetes* phylum, an abundant human gut bacteria that is important for the digestion of complex non-dietary carbohydrates.

Originally crAssphage was published as an individual genome following cross-assembly of several metagenomic samples (Dutilh et al., 2014). Analysis by Manrique et al. (2016) of the healthy human gut phageome identified four circular crAssphage genomes and several related incomplete contigs. PCR amplification and sequencing of the crAssphage polymerase gene by Liang et al. (2016) similarly demonstrated diversity among crAssphage-positive fecal samples. Recently, Cineket al. described updated PCR primer sequences for the detection and evaluation of crAssphage diversity, while Stachler et al. developed primers targeting conserved genomic regions to evaluate the abundance of crAssphage as an indicator of human fecal pollution (Cinek et al., 2018; Stachler et al., 2017). Finally, an epidemiological survey of crAssphages conducted by B.E.,





(legend on next page)

Dutilh, R.A. Edwards, C.H., and colleagues (unpublished data) has suggested that crAssphage is associated with humans and primates globally with significant diversity.

A recent study provided the first detailed sequence-based taxonomic categorization of crAss-like phages, proposing a novel familial-level taxonomic group that would include prototypical crAssphage itself (“p-crAssphage”), as well as various related phages, from multiple environments (Yutin et al., 2018). Previous attempts to reconcile sequence-based and classical viral taxonomy have proposed that Podoviridae (viruses with a short-tail morphology) sharing >40% orthologous protein-coding genes should be grouped at the taxonomic rank of genus, while phages sharing only 20%–40% orthologous protein-coding genes should be grouped at the higher taxonomic rank of subfamily (Lavigne et al., 2008). Other reports describe a phage genus as a cohesive group of viruses sharing >50% nucleotide sequence similarity (Adriaenssens and Brister, 2017). As crAssphages are not a single entity, but rather a group of crAss-like phages that share similarity with the originally discovered p-crAssphage at various levels, a comparative analysis of crAss-like phage sequences is required to enable detailed taxonomic characterization.

In this study, we combine several *in silico* and *in vitro* approaches to further explore the diversity of crAss-like phages in the human gut and better understand their biological properties.

RESULTS

Detection of crAss-like Phage Contigs

Following the assembly of 702 human fecal virome/phageome metagenomic samples listed in Table S1, contigs were screened for relatedness to the p-crAssphage. Initially, the polymerase of p-crAssphage (UGP_018, NC_024711.1) was used for crAss-like phage detection due to its use in several studies as a genetic signature to determine diversity of crAss-like phages (García-Aljaro et al., 2017; Liang et al., 2016, 2018). However, we extended our criteria to include partial genomes (≥ 70 kb) that may not have included the polymerase gene in the assembly. Therefore, after an initial detection of crAss-like phages using the polymerase sequence, we identified the most frequently detected crAss-like phage protein in our dataset as the terminase protein, encoded by p-crAssphage UGP_092. This terminase was subsequently used as a second genetic signature for identifying crAss-like phages.

Initially, 239 contigs ≥ 70 kb were detected with similarity to the p-crAssphage polymerase sequence. An additional 59 contigs ≥ 70 kb were subsequently detected with relatedness to the p-crAssphage terminase sequence. Following an initial examination of these contigs, more stringent parameters were implemented. Only those contigs whose polymerase and/or ter-

minase sequence(s) aligned across greater than 350 bp were considered for further analysis as crAss-like phages. This reduced the total number of crAss-like phages to 256. In addition, as several assembled metagenomic samples were from the same person sequenced at multiple time points, redundant contigs were removed. When a contig aligned with 100% identity across the entire length or within a larger contig, the longer contig or that with the highest coverage was retained (see STAR Methods for a detailed description). This resulted in a total of 244 crAss-like contigs, with 144 containing both a polymerase and a terminase, 60 a polymerase only, and 40 a terminase only (Figure S1).

Contigs ≥ 70 kb from the family-level taxonomic analysis of crAss-like phages conducted by Yutin et al. (2018) were analyzed for inclusion in this study. Whole-genome comparisons highlighted that some sequences predicted as related at the familial taxonomic rank shared no significant homology at the nucleotide level (E-value $1e-05$), including one of this study’s predicted 244 crAss-like phages. In addition, contig SRR073438_s_3, while sharing weak nucleotide homology to crAss-like phages, was more similar to larger putative *Caudovirales* phages (~ 170 kb) and was not predicted as a crAss-like phage. Therefore, a total of 249 contigs/genomes were analyzed in this study as crAss-like phages. Metadata were available for the majority of the crAss-like phage-originating fecal samples (Table S2). CrAss-like phages were detected in healthy individuals across a wide age range (including infants 1 year of age and individuals ≥ 65 years of age) and individuals suffering from Crohn’s disease, ulcerative colitis, HIV, cystic fibrosis, kwashiorkor, and marasmus.

Taxonomy of crAss-like Phages

Previously, studies have used the percentage of shared homologous proteins as a means of defining phage taxonomic ranks (Lavigne et al., 2008). Therefore, clusters of phages sharing between 20% and 40% of their protein-coding genes were categorized as related at the subfamily level, while phages sharing >40% protein-coding genes were grouped at the genus level. A heatmap based on the percentages of shared orthologous proteins suggests that crAss-like phages form four candidate subfamilies. The four subfamilies were assigned the nomenclature Alphacrassvirinae (which contains p-crAssphage), Beta-crassvirinae (which contains IAS virus), Gamma-crassvirinae, and Deltacrassvirinae (Figure 1). These subfamilies can be further subdivided into ten candidate genera, with p containing p-crAssphage and candidate genus VI containing the IAS virus. Metadata of all crAss-like phages analyzed in this study, including their categorization into the various taxonomic divisions, is available in Table S2.

An alternative approach for characterizing the encoded proteome of crAss-like phages was performed by visualization of

Figure 1. Determination of crAssphage Candidate Subfamilies and Genera Based on the Percentage of Shared Protein-Encoding Genes

(Upper) The four red lines cut the hierarchical clustering dendrogram of crAss-like phage contigs into the four proposed candidate subfamilies of crAss-like phages. The histogram insert (top right) represents the calculated optimal number of crAss-like phage clusters. The ten optimal crAss-like phage clusters (I–X) represent the putative candidate genera, and are assigned specific colors.

(Lower) Heatmap showing the percentage of shared protein-coding genes between crAss-like phage genomes. CrAss-like phages with 20%–40% shared protein-coding genes are considered related at the subfamily level while phages with >40% similarity are believed to be related at the genus level, consistent with the calculated number of crAss-like phage clusters.

genome clusters using the t-SNE machine-learning algorithm with Euclidean distances of orthologous gene distribution between genomes as an input. Applying the previously determined ten crAss-like phage candidate genera classifications to the t-SNE, two-dimensional ordination demonstrated that some ellipses are tightly clustered (e.g., candidate genus II), suggesting uniformity. Other ellipses are more relative to the number of sequences (e.g., candidate genus III), signifying heterogeneity (Figure 2A). In addition, no single cluster of crAss-like phages is exclusively associated with healthy or diseased individuals ($R^2 = 0.03$ in Adonis/PERMANOVA, $p = 0.006$).

Specific genera of crAss-like phages share similar G+C% nucleotide composition and may share related bacterial hosts, since phage G+C% content can be related to that of their hosts (Edwards et al., 2016; Lucks et al., 2008). Therefore, several groups of crAss-like phages with similar G+C% compositions, such as candidate genera I, IV, VII, IX, and X, are likely to infect closely related bacterial taxa within the human microbiome (Figure 2B). Candidate genus I is the most homogeneous group of crAss-like phages containing p-crAssphage and 30 additional complete circular genomes and 32 linear contigs ≥ 70 kb, which all share a very similar G+C% nucleotide content ($29.12\% \pm 0.14\%$). Candidate genera III and VI display the greatest heterogeneity, with G+C% contents of $29.07\% \pm 3.07\%$ and $35.32\% \pm 2.15\%$, respectively.

Genome Structure of crAss-like Phages

A set of representative complete circular genomes of the ten genera of crAss-like phages ranges in length from 91.3 to 104.5 kb with varying degrees of genomic synteny (Figure 3). A prominent feature of crAss-like phages is two clearly separated genome regions with opposite gene orientation, the smaller region encoding proteins involved in replication, the bigger region coding for proteins involved in transcription and virion assembly. CrAss-like phages encode large open reading frames (ORFs) with sizes up to 18 kb (UGP_052, UGP_053, UGP_052 in the genome of p-crAssphage), possibly coding for fused subunits of an RNA polymerase (Yutin et al., 2018). However, these large ORFs are only observed in candidate genera VII and VIII when a non-standard genetic code is used in ORF prediction (STAR Methods). All of the crAss-like phages of candidate genera I, II, and IV contain no tRNAs, while members of candidate genus VI had large sets of tRNA genes (up to 27; Table S2). Analysis of the crAss-like phage proteome suggests that four proteins are universally conserved across crAss-like phage subfamilies and have monophyletic evolutionary origin: major capsid protein (MCP, UGP_086), terminase (UGP_092), portal protein (UGP_091), and primase (UGP_025; Figure S2; Table S3). Phylogeny of crAss-like phages based on multiple alignment of these four proteins supports that based on percentage of shared proteins and clearly separates the four subfamilies. As clearly shown in Figure 3, blocks of genes responsible for tail morphogenesis are especially variable even between members of the same subfamily, both in the number of genes and sequence of the encoded protein products, highlighting the potentially wide range of bacterial hosts infected by different crAss-like phage genera and strains.

Prevalence of crAss-like Phages in Human Fecal Virome Samples

To obtain insights into relative abundance in various human populations, we aligned quality-filtered reads, representing 512 human fecal samples from the same datasets as used for assembly of crAss-like genomes, with a database of 131 non-redundant crAss-like phage genomic sequences (with $<90\%$ of homology and/or $<90\%$ overlap between them) representing all ten candidate genera.

CrAss-like phage colonization rates varied from 51%–59% in Malawian infants to 98%–100% of healthy individuals of various ages in the Western cohorts (Figure S3A). In total, 77% of all samples were positive for one or more crAss candidate genera. The relative phage abundance ranged from 0% to 95% of total reads per sample (Figure S3B), and depended significantly on the country of residence ($p = 4.2 \times 10^{-9}$ in Kruskal-Wallis test) and age group of the donor ($p = 1.1 \times 10^{-10}$). In $\sim 8\%$ of all virome samples, $>50\%$ of reads aligned to crAss-like phage genomes. The lowest overall crAss-like phage counts were seen in Irish and Malawian infants and in United States adults with IBD (Figure S3A). On a global scale, crAss-like candidate genera I, VIII, and IX seem to be the most prevalent with the highest mean percentage of reads aligned; 5%, 1.7%, and 1.8%, respectively.

The specific composition of crAss-like phages in feces partly separated a cohort of healthy and malnourished infants living in rural areas of Malawi from the healthy and diseased urban Western cohorts (Figure S3C). PERMANOVA analysis suggested that crAss-like phage composition was mostly driven by place of residence ($R^2 = 0.24$, $p = 0.001$) with condition and age group also having significant impact ($R^2 = 0.05$ and 0.01 , respectively, $p = 0.001$). This observation is further supported by a clear difference in the distribution of specific crAss-like candidate genera across different populations (Figure 4). Specifically, candidate genus I, which includes p-crAssphage, is by far the most prevalent type in the Western population regardless of age. At the same time, the same genus was extremely scarce in the Malawian cohort where candidate genera VIII and IX were the most common ($p = 4.7 \times 10^{-2}$ and 4.4×10^{-13} , respectively).

Fecal Fermentations of a Candidate Genus I-Rich Sample

During an ongoing longitudinal study of fecal viromes in healthy adults we identified one individual (subject ID 924), in whom p-crAssphage consistently contributed $>30\%$ of virome metagenomic reads over a 12-month period. Thus, this donor was selected to investigate whether p-crAssphage could be propagated in a batch fecal fermentation system. qPCR detection of a conserved fragment of the p-crAssphage DNA polymerase gene in the viral nucleic acid fractions throughout the fermentation revealed that p-crAssphage was effectively propagated. P-crAssphage was found to increase in titer 89-fold 21 hr into the fermentation (Figure 5A).

Interestingly, a shotgun metagenomic sequencing run, following multiple displacement amplification (MDA) of the viral-enriched DNA from the fermentation supernatants, showed the presence of six other crAss-like phages in the study subject, in addition to a phage highly similar to p-crAssphage (Figure 5B; Table S2). Each of these crAss-like phage contigs were ≥ 70 kb

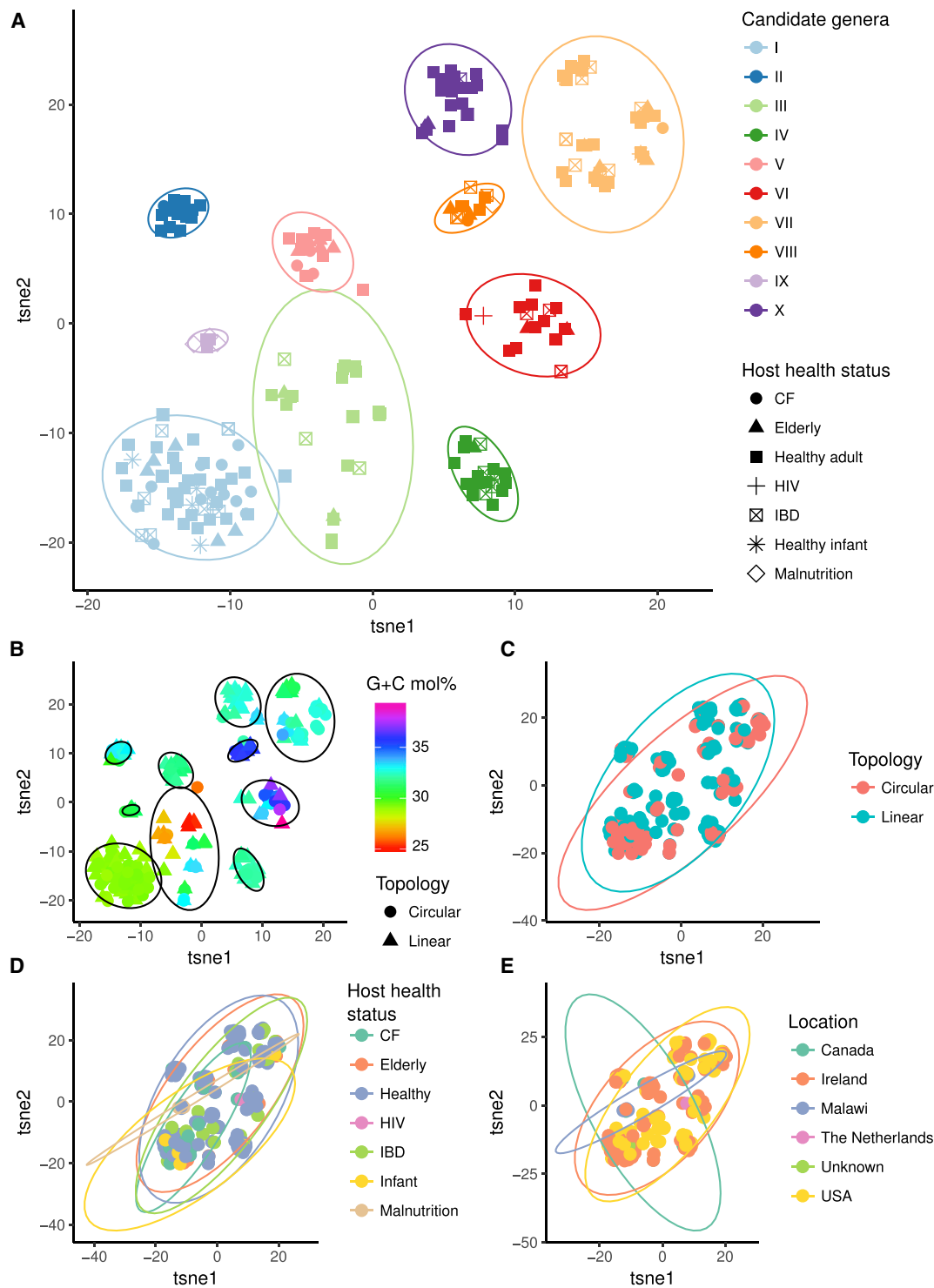


Figure 2. Two-Dimensional Ordination of crAss-like Phages Based on the Abundance of Their Protein-Encoded Orthologous Sequences Was Performed Using t-SNE Machine-Learning Algorithm

(A) CrAss-like phages are colored by candidate genus annotations and shape is determined by the health status of individuals carrying these crAss-like phages. (B) CrAss-like phages are colored by the percentage G + C mol% nucleotide composition of their contig, while shape represents complete (circular) or partial (linear) genomes.

(C) Ellipses highlighting the distribution of complete (circular) or near-complete (linear) crAss-like phage genomes.

(D) Ellipses grouping the health status associated with the individual donating fecal samples which yielded crAss-like phage contigs.

(E) Ellipses highlighting the geographical location of assembled crAss-like phage contigs.

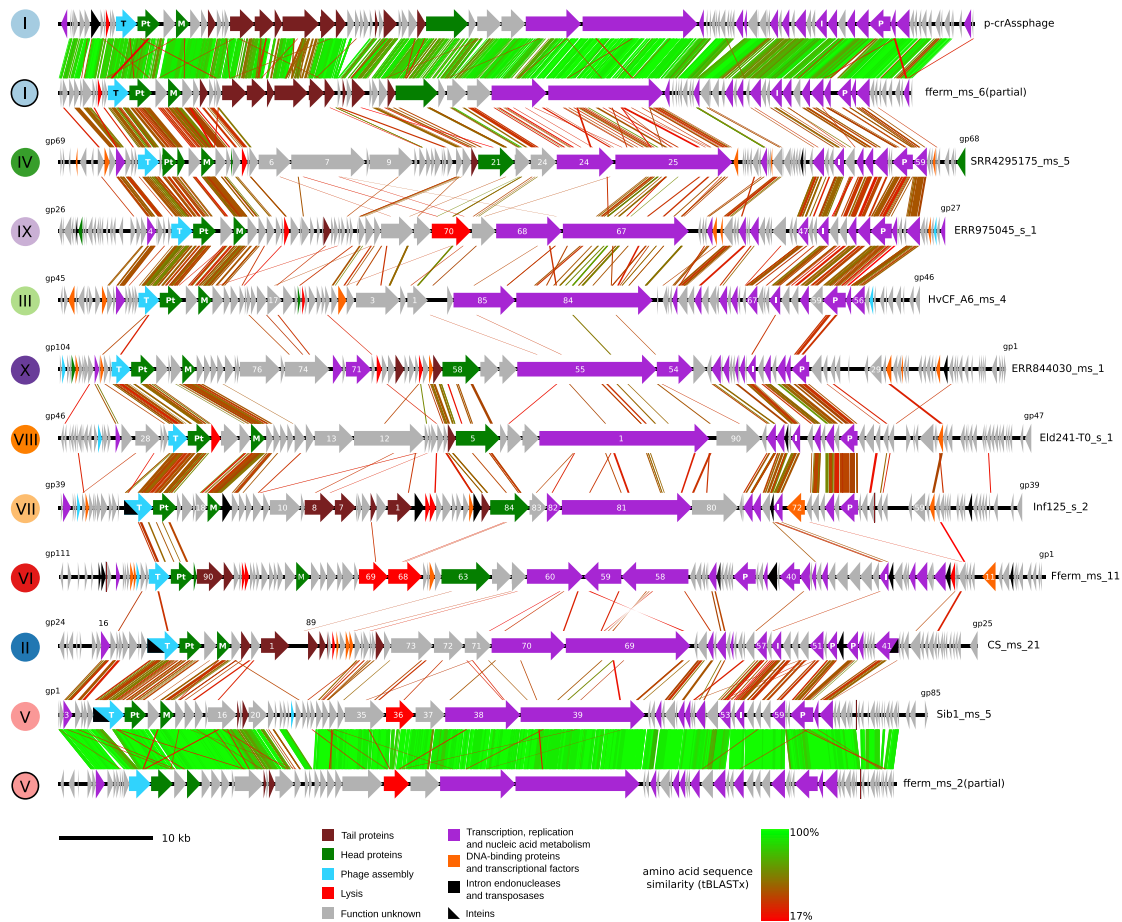


Figure 3. Whole-Genome Comparisons of crAss-like Phages from the Proposed Ten Candidate Genera (Including Ten Complete Circular Genomes Representative of Each Genus)

Partial genomes of Fferm_ms_2 and Fferm_ms_6 of subject ID 924, which are prevalent during the *in vitro* characterizations of crAss-like phages, highlight the inter- and intra-relatedness of the different crAss-like genera. Circular genome maps were permuted in order to standardize for starting coordinate and gene order (gene product [gp] numbers indicate the first and the last gene on a map from left to right). Protein-coding sequences (CDS, arrows) are colored by putative HHPred-predicted function. Numbers inside CDS arrows indicate gp numbers (see Table S3 for detailed information). Regions of tBLASTx homology between phage genomes are highlighted.

and grouped into five candidate genera (Figure 5B). Four of these contigs contributed to $\geq 1\%$ of the reads per sample. The most abundant phage of subject ID 924 detected after sequencing this viral DNA preparation was crAss-like phage Fferm_ms_6 (linear, 90.4kb), a member of candidate genus I and closely related to p-crAssphage. Contig Fferm_ms_2 (linear, 88.8 kb) is the second most abundant in the sample and belongs to candidate genus V. Five additional crAss-like phages were detected in the feces of subject ID 924 (Table S2).

Analysis of the bacterial fraction of the microbiota in the fermentation vessel was performed using compositional 16S rRNA gene sequencing to investigate potential crAss-like phage hosts and to examine the hypothesis that their hosts are of the *Bacteroides* genus or *Bacteroidetes* phylum. The analysis showed a decrease in the relative abundance of *Bacteroides* up to time point 21, after which levels begin to recover. The converse was observed for p-crAssphage propagation for which titers increased up to time point 21 followed by a gradual decrease. Two *Bacteroides* species found to be abundant in

the vessel included *B. dorei* and *B. uniformis*. However, attempts to isolate crAss-like phages on these strains did not yield plaques; therefore, we cannot definitively infer a host for p-crAssphage, but only confirm the presence of bacteria that other studies have suggested as putative hosts (Cinek et al., 2017; Reyes et al., 2013; Figure S4).

Biological Characterization of crAss-like Phages

Transmission electron microscopy (TEM) of a fecal filtrate rich in a candidate genus I crAss-like phage showed a significant presence of short-tailed or non-tailed viral particles with icosahedral or isometric heads (53% with a Podoviridae-like short-tail morphology and 29% of Microviridae or a smaller type of Podoviridae), with lower levels of tailed phages with a Siphoviridae-like morphology (15%; Figure 6A). The large Podoviridae-like icosahedral capsids with short tails could be further classified into two types: type I, with head diameters of ~ 76.5 nm and short tails; and type II, with a similar head size but head-tail collar structures and slightly longer tails (Figure 6B). Sequencing,

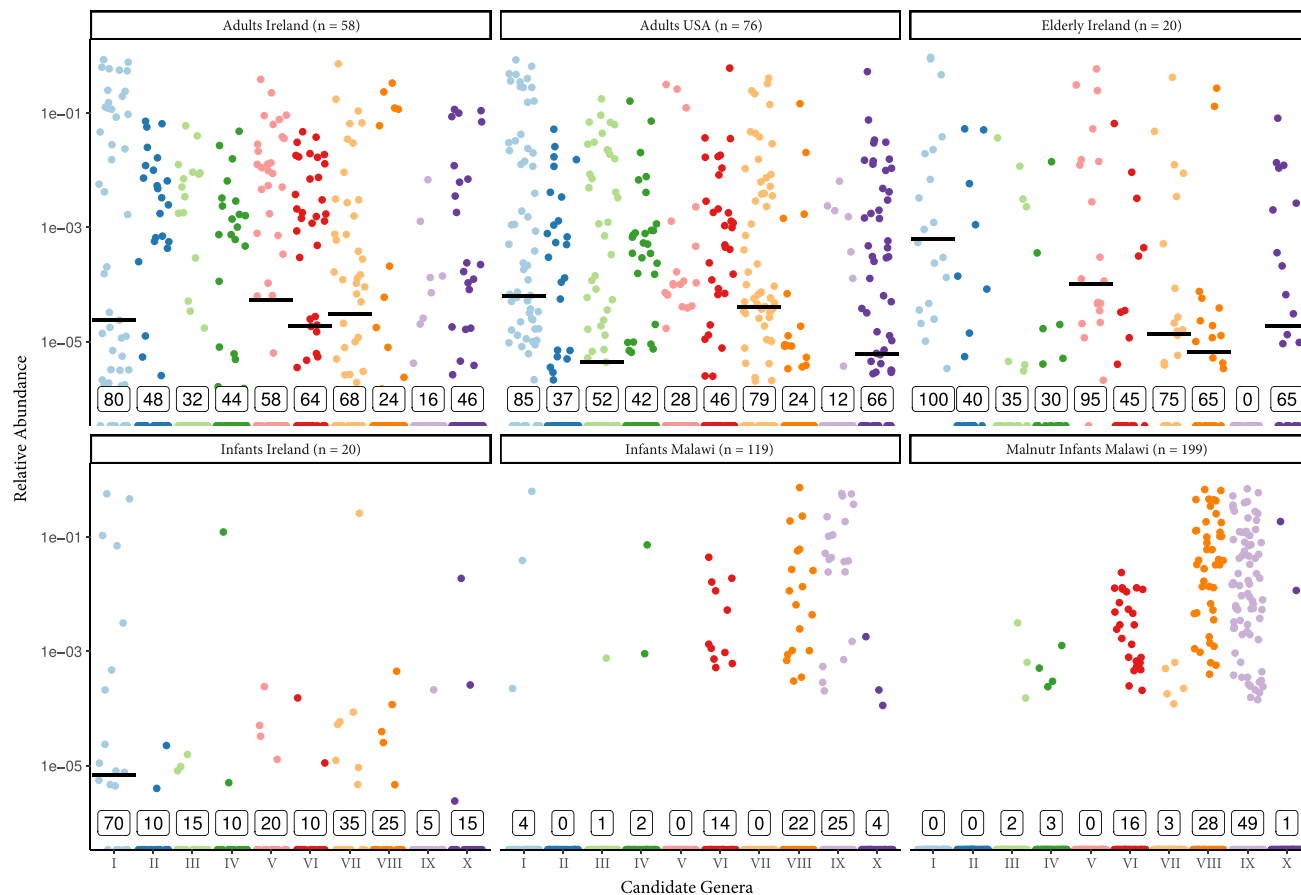


Figure 4. Relative Abundance of the Ten Candidate Genera of crAss-like Phages in Six Different Human Cohorts Based on the Fraction of Metagenomic Reads Aligned

Bars represent median relative abundances, while values within boxes represent percentage of positive samples.

without MDA, of CsCl purified fraction of the same fecal material as used for the TEM showed that approximately 40% of reads aligned to crAss-like genomic contigs (Figure 6C), with the candidate genus V crAss-like phage being the most abundant. Based on the size of the seven crAss-like genomic contigs assembled (88.8–104.6 kb; termed Fferm contigs within Table S2), it is predicted that the predominant podovirus morphology observed corresponds to the crAss-like phages. For comparison, Microviridae phages have genomes 4.4–6.1 kb and icosahedral capsids of approximately 15–30 nm in diameter (Roux et al., 2012; Zhong et al., 2015).

The same CsCl fraction of feces that was subjected to metagenomic sequencing without MDA and TEM visualization was also analyzed by SDS-PAGE followed by identification of major bands using MALDI-TOF mass spectrometry. A major structural protein of crAss-like phage Fferm_ms_2 was detected from a band excised from the ~55 kDa area on an SDS-PAGE gel (Figure 6D). The obtained peptide profile corresponded to a protein of 490 amino acids and 55.4 kDa, with analysis showing the protein as having 37% identity with UGP_086, predicted as the major capsid protein of p-crAssphage (Yutin et al., 2018).

Finally, we attempted to independently establish the size of crAss-like phage virions by passing fecal filtrates through a se-

ries of filters with gradually decreasing pore sizes (Figure S5). Filtration through 0.1- μm pores (equivalent to 100 nm) resulted in partial retention of crAss-like phages while pores of 0.02 μm completely removed crAssphage from the filtrate, as judged by qPCR.

DISCUSSION

The overall objective of this study was to gain an insight into one of the most enigmatic phages discovered to date, crAssphage. This phage is highly abundant in the human microbiome on a global scale; however, it remains poorly understood. One reason why crAssphage has remained such a mystery is due to the lack of available genome sequences for comparison. When crAssphage was assigned a specific nomenclature and uploaded to a public repository by Dutilh et al. (2014), it became a template for other studies.

p-crAssphage was the first identified representative of an expanding group of phages, associated with animal, soil, and oceanic microbiomes (Dutilh et al., 2014; Yutin et al., 2018). While a previous study proposed a sequence-based classification of crAss-like phages at the familial level (Yutin et al., 2018), our *in silico* analysis focuses on human fecal-associated

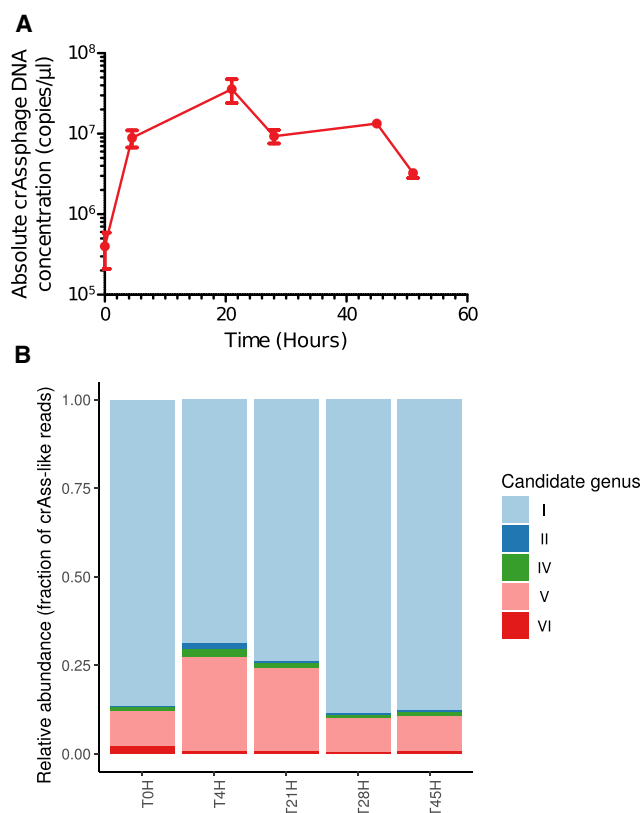


Figure 5. Analysis of crAss-like Phage Dynamics in a Fecal Fermenter

(A) Evidence of a candidate genus I crAss-like phage propagation following *in vitro* fermentations (with standard error of the mean, $n = 3$). The level of propagation was determined by qPCR analysis of viral-enriched DNA, respectively, using primers specific to a segment of the p-crAssphage DNA polymerase gene.

(B) Six additional crAss-like phages that group into five of the candidate genera were identified following sequencing of the same viral-enriched DNA from the fermenter. The relative abundance of each of these crAss-like phages is skewed due to the biased amplification of other components of the viral-enriched DNA fraction that is associated with multiple displacement amplification (MDA).

crAss-like phages. In this study, we present 242 previously unrecognized crAss-like phages from various metagenomic studies. Comparative genomics demonstrates an extensive degree of diversity among these phages, including the identification of four crAss-like phage subfamilies. While the Alphacrassvirinae subfamily currently has the greatest number of sequences of the four subfamilies, future studies looking for additional homologs of Betacrassvirinae, Gammacrassvirinae, and Deltacrassvirinae members will expand and refine these taxonomic ranks. In particular, future fecal virome/phageome studies of humans from diverse geographical locations will likely expand the repertoire of known human-associated crAss-like phages significantly, as the large interpersonal differences in the human virome are likely multiplied by variations of diet and environmental exposure.

Assigning phage taxonomy, in the absence of a universal genetic marker such as 16S rRNA is a difficult and potentially erro-

neous process. While crAss-like phages will likely form their own taxonomic family (Yutin et al., 2018), they possess a dsDNA genome and a Podoviridae-like short-tailed virion. The categorization of crAss-like phages by the percentage of shared proteins identified ten candidate genera, with crAss-like phages in each genera originating from the feces of putatively healthy individuals and people suffering from various metabolic, infectious, and diet- and gut-related disorders.

Several crAss-like phage genera proposed in this study have distinct nucleotide G+C% compositions. The nucleotide composition of obligate parasites, such as phages, can evolve in close association with their host bacterium (Lucks et al., 2008; Mavrich and Hatfull, 2017; Pride et al., 2006; Roux et al., 2015). Following this logic, candidate genera III and VI with diverse G+C% compositions are either heterogeneous groups of crAss-like phages that require further sequences to refine their taxonomic structure, or they are potentially capable of infecting across a broad host range. However, not all phages have a G+C% composition that mirrors their host; therefore, these results must be treated cautiously until further investigated (Henry et al., 2015).

Quantitative analysis of the crAss-like phage content in several cohorts revealed that, in agreement with previous studies, the vast majority of fecal viral metagenomic samples contained varied amounts of crAssphage DNA. P-crAssphage (candidate genus I) is by far most predominant type in Western populations, co-existing with other crAss-like phages in the majority of samples. By contrast, in the cohort of malnourished and healthy Malawian infants (Reyes et al., 2015; Smith et al., 2013), other candidate genera such as VI, VIII, and IX are more abundant. It is well known that non-Western rural populations, which mostly consume a high-fiber, low-fat, and low-animal-protein diet are predominantly associated with high *Prevotella*/low *Bacteroides* type of gut microbiota (known as enterotype II [Arumugam et al., 2011]), as opposed to *Bacteroides*/*Clostridia*-dominated microbiota (enterotype I) in urban populations consuming a Western diet (Filippo et al., 2010; Gorvitovskaia et al., 2016). Indeed, our analysis of the Reyes et al. (2015) 16S rRNA gene sequencing data confirmed high prevalence of *Prevotella* in Malawian samples (Figure S6). Therefore, one can hypothesize that crAss-like phages of candidate genera VIII and IX might be associated with *Prevotella* or other members of the order *Bacteroidales*.

The *in vitro* analysis of samples obtained from subject ID 924 was particularly intriguing. By mapping metagenomic sequencing reads against p-crAssphage, it was initially thought that this donor only carried the prototypical crAssphage at levels exceeding 30% of total viral reads for a 1-year period. Subsequent mining for phages related to p-crAssphage using metagenomic sequencing at later time points, with and without MDA, resulted in six additional crAss-like phages being simultaneously detected in this donor. It is possible that many additional crAss-like phage genomes could be present within the metagenomic datasets examined in this study, but they were not included in our analysis because of the inclusion criteria chosen or even the choice of assembly program.

In total, subject ID 924 consistently carried seven crAss-like phages, which resolved in our taxonomic analysis into five candidate genera. Three belonged to candidate genus VI, supporting the notion this is a heterogeneous group and not simply

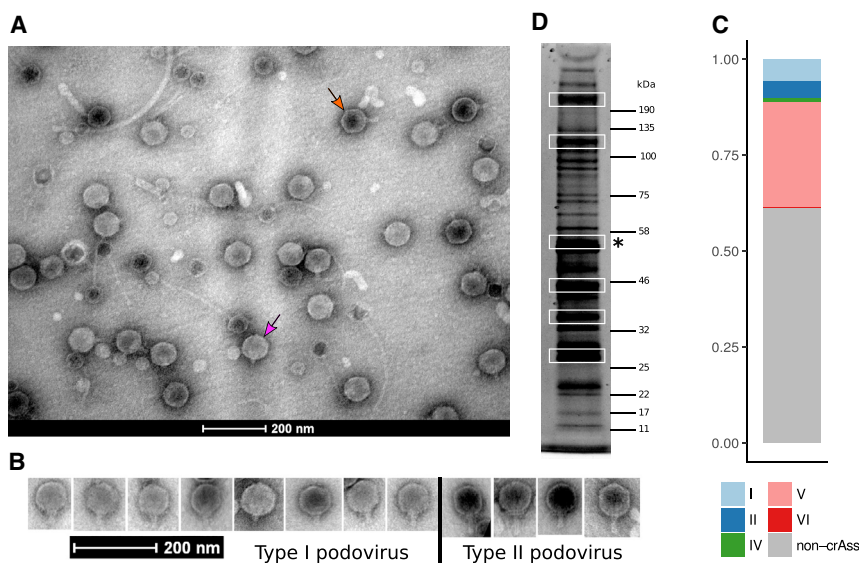


Figure 6. CrAss-like Phage Morphology Was Examined Using a CsCl Fraction Purified from a crAss-like Phage-Rich Filtrate of Donor Subject ID 924

(A) Analysis of the fraction through TEM was performed, and is largely dominated by podoviruses (53%), microviruses (29%), siphoviruses (15%), and other phage morphologies (3%).

(B) Further examination of the observed podovirus virions identifies two variants with differing tail morphologies, highlighted in (A) with a pink and orange arrow, respectively. Both variants have head diameters of ~ 76.5 nm.

(C) Sequencing of the CsCl purified viral fraction, without MDA, showed that approximately 40% of the reads aligned to crAss-like phages.

(D) SDS-PAGE gel of the CsCl fraction, highlighting six bands that were excised and analyzed by mass spectrometry. The major capsid protein of crAss-like phage Fferm_ms_2 was detected from the ~ 55 kDa band (the white box highlighted by the asterisk).

composed of broad host range infecting phages. It is possible that there are more than seven crAss-like phages within subject ID 924. However, it is most probable that only a single representative of each candidate crAss-like phage genus (with the exception of the heterogeneous candidate genus VI) could assemble correctly, with two or more highly identical phages amalgamating their SNPs into a single-consensus representative sequence. This is a known feature of the chosen assembly program, where microdiversity is lost at the expense of assembling longer contigs/low-coverage contigs within complex samples (Vollmers et al., 2017).

This study demonstrates the proliferation of crAss-like phages in a fecal fermenter, providing evidence of a phage similar to p-crAssphage propagating under laboratory conditions. Furthermore, following our ability to propagate fecal crAss-like phages, we conducted TEM imaging of these phages using fecal filtrate from sample ID 924 prepared prior to fermentation. The most abundant fecal viruses had short tails and were Podoviridae-like in structure. This agrees with the predictions made by Yutin et al. (2018) following their detailed genome annotation of two crAss-like phages. Interestingly, however, our TEM images suggest the presence of two types of virions with short non-contractile tails (Figure 6B). Presumably, the more abundant type I virions with shorter tails may belong to members of candidate genus V, found as the most abundant crAss-like phage group sequenced in the same sample from subject ID 924 (Figure 6C). But without isolating these phages in pure culture, it is not possible to accurately assign which tail structure corresponds to which specific crAss-like phage subfamily or genera.

This work provides multiple levels of *in vitro* evidence confirming that crAss-like phages have a short-tailed podovirus structure. Experimentally, this is shown using the same CsCl fraction purified from a crAssphage-rich fecal sample of a healthy human donor. A qPCR performed on this fraction using candidate genus I-specific primers showed that approximately 1×10^9 copies per microliter were present. To examine the size of the crAss-like phage virions *in vitro*, we passed fecal filtrates from the same

sample through a number of filters with decreasing pore sizes (ranging from 450 to 20 nm). We found that p-crAssphage could no longer be detected after filtration through a pore size of 20 nm, but was partially retained by filter sizes of 100 nm. This supports the size prediction of crAss-like phage virions observed in TEM. Sequencing, without MDA, of the same CsCl fraction that was visualized by TEM confirmed that almost 40% of the reads aligned to crAss-like phages, consistent with the percentage of short-tailed Podoviridae-like phage virions present in the sample. An examination of the protein content in the sample visualized by TEM detected the predominant major capsid protein of the Fferm_ms_2 crAss-like phage. The capsid protein was found to have similarity to other crAss-like phages of candidate genus V, as well as a moderate degree of similarity to p-crAssphage (candidate genus I).

Identifying a means of propagating crAss-like phages is of particular importance in expanding our knowledge on crAss-like phages. However, the primers applied in the qPCR analyses of viral nucleic acids were not suitable for targeting crAss-like phages associated with the various subfamilies and candidate genera other than p-crAssphage. With the availability of more crAss-like phage sequences, broad- and narrow-spectrum qPCR assays can be subsequently designed and applied to the analysis of these phages, which will be an important part of future work.

It is clear that human gut-associated crAss-like phages are not a single entity, but rather a group of diverse viruses sharing genomic traits, which target diverse bacterial taxa of the human microbiome. Previously, a member of the *Bacteroides* genus was hypothesized as being the host for crAssphage (Dutilh et al., 2014). In a study prior to the discovery of crAssphage (Reyes et al., 2013), a 95.9-kb contig corresponding to a putative virus ϕ HSC05 was shown to be stably engrafted after transplantation of human fecal virus fraction into germ-free mice colonized with an artificial defined community of 15 bacterial species. The retrospective analysis of contigs from that study conducted by ourselves showed that the ϕ HSC05 contig was 91.73% identical

by its nucleotide sequence to p-crAssphage. The artificial bacterial community, among others, included: *Bacteroides thetaiotaomicron* (2 strains), *B. caccae*, *B. ovatus*, *B. vulgatus*, *B. cellulosilyticus*, and *B. uniformis*. We suggest that one of these strains is more likely to have served as a host for this crAss-like phage propagation than the remaining eight strains of Gram-positive anaerobic bacteria. Since crAssphage had not been described at the time the article was published, this very interesting observation obviously could not have been made at that time. Recently, a crAss-like phage infecting *B. intestinalis* has been isolated in axenic culture (Shkoporov et al., 2018), allowing a preliminary investigation of its host range, replication strategy, virion morphology, and potential ecological impact on the human gastrointestinal microbiota.

With more divergent sequences, we could assume that different members of the *Bacteroides* genus, or even *Bacteroidetes* phylum for example, may serve as hosts for different crAss-like phages. One host that has been hypothesized for prototypical crAss-like phages is *B. dorei* (Cinek et al., 2017). This was inferred following the analysis of a dataset generated from infants and toddlers with islet autoimmunity. It was shown that crAssphage was only present when *B. dorei* also was detected within the samples. This was not true for other *Bacteroides* members tested, including *B. vulgatus*, which is highly related to *B. dorei*. This correlation is compelling; however, it should be noted that there was no confirmation that crAssphage has any role in causing bacteriome alterations that lead to islet autoimmunity. Interestingly, one of the key *Bacteroides* species detected from our fecal fermentation 16S rRNA analysis was *B. dorei*. Although our results cannot confirm this as a possible host of a crAss-like phage, this phage-host pair as well as the *Bacteroides* discussed above merit further investigation.

CrAss-like phages have also been defined as a part of the core human gut phageome (Manrique et al., 2016). This emphasizes the importance of identifying hosts for diverse crAss-like phages belonging to different candidate genera proposed in this study. The ability to propagate crAss-like phages *in vitro* will prove a key step in gaining an insight into their biological significance, including the possible role they play in shaping the bacterial composition of the human gut microbiome. This could be in a positive or negative manner, in the context of various disease states, such as IBD, cancer, and obesity among others. Thus far, only a few studies have attempted to correlate crAss-like phages with a gastrointestinal disorder (Cinek et al., 2017; Liang et al., 2016; Norman et al., 2015).

In conclusion, our results expand the repertoire of known crAss-like phages significantly, providing a path toward the identification of further crAss-like phages and their hosts. This will lead to a better understanding of their role, if any, in human health and disease. Our work also provides an interesting insight into the diversity of these human gut-associated phages in various populations. In addition, we also demonstrate that these enigmatic phages can be efficiently propagated *in vitro* in a mixed culture and present TEM images of crAss-like phages, giving an insight into their morphology. CrAss-like phages appear to be universally present in human populations, including those with various disease states. Due to the specificity of phage-host interactions, the diversity of crAss-like phages sug-

gests that they infect multiple diverse bacteria of the human gastrointestinal microbiota. However, more studies will be required to determine the biological significance and role of crAss-like phages in the human gut and to determine whether its presence positively or negatively affects human gastrointestinal health.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- EXPERIMENT MODEL AND SUBJECT DETAILS
 - CrAss-like Phage Rich Faeces
- METHOD DETAILS
 - Metagenomic Datasets and Contig Assemblies
 - Detection and Curation of crAss-like Phages
 - Identification of crAss-like Phage Orthologous Proteins and Clusters
 - Genomic Comparisons of crAss-like Phages
 - Alignment of Virome Metagenomic Reads to crAss-like Contigs
 - Recruitment of a crAssphage Faecal Donor and Faecal Fermentation
 - Extraction of Viral Nucleic Acids and Sequencing Library Preparation
 - P-crAssphage PCR Detection
 - Electron Microscopy and Detection of crAss-like Phage Proteins
 - 16S rRNA Gene Library Preparations
- QUANTIFICATION AND STATISTICAL ANALYSIS
 - Alignment of Virome Metagenomic Reads to crAss-like Contigs
 - Faecal Fermentations
- DATA AND SOFTWARE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental Information includes six figures, four tables, and one data file and can be found with this article online at <https://doi.org/10.1016/j.chom.2018.10.002>.

ACKNOWLEDGMENTS

This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under grant numbers SFI/12/RC/2273, SFI/15/ERC/3189, and SFI/14/SP APC/B3032, and a research grant from Janssen Biotech.

AUTHOR CONTRIBUTIONS

E.G. and S.R.S. performed the laboratory and bioinformatics work, respectively. A.S. assisted in both the laboratory and bioinformatics analyses. A.G.C. performed the 16S analysis. F.J.R., T.D.S.S., L.A.D., and E.G.-T. assisted in the design, implementation, and interpretation of experiments. E.G., A.S., and S.R.S. wrote the paper and generated the figures. A.G.C., F.J.R., T.D.S.S., L.A.D., and E.G.-T. reviewed drafts of the manuscript and provided constructive criticism for its improvement. R.P.R. and C.H. secured the funding and wrote the paper. All authors contributed to the analysis of the data.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: April 16, 2018

Revised: July 2, 2018

Accepted: September 17, 2018

Published: October 25, 2018

REFERENCES

- Adriaenssens, E., and Brister, J.R. (2017). How to name and classify your phage: an informal guide. *Viruses* 9, <https://doi.org/10.3390/v9040070>.
- Allard, G., Ryan, F.J., Jeffery, I.B., and Claesson, M.J. (2015). SPINGO: a rapid species-classifier for microbial amplicon sequences. *BMC Bioinformatics* 16, 324.
- Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402.
- Anderson, M.J. (2001). A new method for non-parametric multivariate analysis of variance. *Aust. Ecol.* 26, 32–46.
- Arumugam, M., Raes, J., Pelletier, E., Le Paslier, D., Yamada, T., Mende, D.R., Fernandes, G.R., Tap, J., Bruls, T., Batto, J.-M., et al. (2011). Enterotypes of the humangutmicrobiome. *Nature* 473, 174–180.
- Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Pribelski, A.D., et al. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* 19, 455–477.
- Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120.
- Callahan, B.J., McMurdie, P.J., Rosen, M.J., Han, A.W., Johnson, A.J.A., and Holmes, S.P. (2016). DADA2: high-resolution sample inference from Illumina amplicon data. *Nat. Methods* 13, 581–583.
- Cinek, O., Kramna, L., Lin, J., Oikarinen, S., Kolarova, K., Ilonen, J., Simell, O., Veijola, R., Autio, R., and Hyöty, H. (2017). Imbalance of bacteriome profiles within the Finnish Diabetes Prediction and Prevention study: parallel use of 16S profiling and virome sequencing in stool samples from children with islet autoimmunity and matched controls. *Pediatr. Diabetes* 18, 588–598.
- Cinek, O., Mazankova, K., Kramna, L., Odeh, R., Alassaf, A., Ibeke, M.U., Ahmadov, G., Mekki, H., Abdullah, M.A., Elmahi, B.M.E., et al. (2018). Quantitative CrAssphage real-time PCR assay derived from data of multiple geographically distant populations. *J. Med. Virol.* 90, 767–771.
- Claesson, M.J., Jeffery, I.B., Conde, S., Power, S.E., O'Connor, E.M., Cusack, S., Harris, H.M.B., Coakley, M., Lakshminarayanan, B., O'Sullivan, O., et al. (2012). Gut microbiota composition correlates with diet and health in the elderly. *Nature* 488, 178–184.
- Cryan, J.F., and Dinan, T.G. (2014). Mind-altering microorganisms: the impact of the gut microbiota on brain and behaviour. *Nat. Rev. Neurosci.* 13, 701–712.
- Duncan, S.H., Hold, G.L., Harmsen, H.J.M., Stewart, C.S., and Flint, H.J. (2002). Growth requirements and fermentation products of *Fusobacterium prausnitzii*, and a proposal to reclassify it as *Faecalibacterium prausnitzii* gen. nov., comb. nov. *Int. J. Syst. Evol. Microbiol.* 52, 2141–2146.
- Dutilh, B.E., Cassman, N., McNair, K., Sanchez, S.E., Silva, G.G.Z., Boling, L., Barr, J.J., Speth, D.R., Seguritan, V., Aziz, R.K., et al. (2014). A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. *Nat. Commun.* 5, 5498.
- Edgar, R.C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26, 2460–2461.
- Edwards, R.A., McNair, K., Faust, K., Raes, J., and Dutilh, B.E. (2016). Computational approaches to predict bacteriophage–host relationships. *FEMS Microbiol. Rev.* 40, 258–272.
- Everard, A., and Cani, P.D. (2013). Diabetes, obesity and gut microbiota. *Best Pract. Res. Clin. Gastroenterol.* 27, 73–83.
- Filippo, C.D., Cavalieri, D., Paola, M.D., Ramazzotti, M., Poulet, J.B., Massart, S., Collini, S., Pieraccini, G., and Lionetti, P. (2010). Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa. *Proc. Natl. Acad. Sci. U S A* 107, 14691–14696.
- Frank, D.N., Robertson, C.E., Hamm, C.M., Kpadeh, Z., Zhang, T., Chen, H., Zhu, W., Sartor, R.B., Boedeker, E.C., Harpaz, N., et al. (2011). Disease phenotype and genotype are associated with shifts in intestinal-associated microbiota in inflammatory bowel diseases. *Inflamm. Bowel Dis.* 17, 179–184.
- García-Aljaro, C., Ballesté, E., Muniesa, M., and Jofre, J. (2017). Determination of crAssphage in water samples and applicability for tracking human faecal pollution. *Microb. Biotechnol.* 10, 1775–1780.
- Garneau, J.R., Depardieu, F., Fortier, L.-C., Bikard, D., and Monot, M. (2017). PhageTerm: a tool for fast and accurate determination of phage termini and packaging mechanism using next-generation sequencing data. *Sci. Rep.* 7, 8292.
- Gorvitovskaia, A., Holmes, S.P., and Huse, S.M. (2016). Interpreting Prevotella and Bacteroides as biomarkers of diet and lifestyle. *Microbiome* 4, 15.
- Henry, M., Bobay, L.-M., Chevallereau, A., Sausseureau, E., Ceysens, P.-J., and Debarieux, L. (2015). The search for therapeutic bacteriophages uncovers one new subfamily and two new genera of pseudomonas-infecting myoviridae. *PLoS One* 10, e0117163.
- Hyatt, D., Chen, G.-L., LoCascio, P.F., Land, M.L., Larimer, F.W., and Hauser, L.J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 11, 119.
- Klindworth, A., Pruesse, E., Schweer, T., Peplies, J., Quast, C., Horn, M., and Glöckner, F.O. (2013). Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Res.* 41, e1.
- Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9, 357–359.
- Laslett, D., and Canback, B. (2004). ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Res.* 32, 11–16.
- Lavigne, R., Seto, D., Mahadevan, P., Ackermann, H.-W., and Kropinski, A.M. (2008). Unifying classical and molecular taxonomic classification: analysis of the Podoviridae using BLASTP-based tools. *Res. Microbiol.* 159, 406–414.
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* 25, 2078–2079.
- Li, L., Stoeckert, C.J., and Roos, D.S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* 13, 2178–2189.
- Liang, Y., Jin, X., Huang, Y., and Chen, S. (2018). Development and application of a real-time polymerase chain reaction assay for detection of a novel gut bacteriophage (crAssphage). *J. Med. Virol.* 90, 464–468.
- Liang, Y.Y., Zhang, W., Tong, Y.G., and Chen, S.P. (2016). crAssphage is not associated with diarrhoea and has high genetic diversity. *Epidemiol. Infect.* 144, 3549–3553.
- Lucks, J.B., Nelson, D.R., Kudla, G.R., and Plotkin, J.B. (2008). Genome landscapes and bacteriophage codon usage. *PLoS Comput. Biol.* 4, e1000001.
- van der Maaten, L., and Hinton, G. (2008). Visualizing data using t-SNE. *J. Mach. Learn. Res.* 9, 2579–2605.
- Manrique, P., Bolduc, B., Walk, S.T., van der Oost, J., de Vos, W.M., and Young, M.J. (2016). Healthy human gut phageome. *Proc. Natl. Acad. Sci. U S A* 113, 10400–10405.
- Manrique, P., Dills, M., and Young, M.J. (2017). The human gut phage community and its implications for health and disease. *Viruses* 9, 141.
- Marshall, O.J. (2004). PerlPrimer: cross-platform, graphical primer design for standard, bisulphite and real-time PCR. *Bioinformatics* 20, 2471–2472.
- Mavrich, T.N., and Hatfull, G.F. (2017). Bacteriophage evolution differs by host, lifestyle and genome. *Nat. Microbiol.* 2, 17112.
- Mills, S., Shanahan, F., Stanton, C., Hill, C., Coffey, A., and Ross, R.P. (2013). Movers and shakers: influence of bacteriophages in shaping the mammalian gut microbiota. *Gut Microbes* 4, 4–16.

- Milne, I., Bayer, M., Cardle, L., Shaw, P., Stephen, G., Wright, F., and Marshall, D. (2010). Tablet—next generation sequence assembly visualization. *Bioinformatics* 26, 401–402.
- Minot, S., Sinha, R., Chen, J., Li, H., Keilbaugh, S.A., Wu, G.D., Lewis, J.D., and Bushman, F.D. (2011). The human gut virome: inter-individual variation and dynamic response to diet. *Genome Res.* 21, 1616–1625.
- Norman, J.M., Handley, S.A., Baldrige, M.T., Droit, L., Liu, C.Y., Keller, B.C., Kambal, A., Monaco, C.L., Zhao, G., Fleshner, P., et al. (2015). Disease-specific alterations in the enteric virome in inflammatory bowel disease. *Cell* 160, 447–460.
- Nurk, S., Meleshko, D., Korobeynikov, A., and Pevzner, P.A. (2017). metaSPAdes: a new versatile metagenomic assembler. *Genome Res.* 27, 824–834.
- O'Donnell, M.M., Rea, M.C., O'Sullivan, Ó., Flynn, C., Jones, B., McQuaid, A., Shanahan, F., and Ross, R.P. (2016). Preparation of a standardised faecal slurry for ex-vivo microbiota studies which reduces inter-individual donor bias. *J. Microbiol. Methods* 129, 109–116.
- Pride, D.T., Wassenaar, T.M., Ghose, C., and Blaser, M.J. (2006). Evidence of host-virus co-evolution in tetranucleotide usage patterns of bacteriophages and eukaryotic viruses. *BMC Genomics* 7, 8.
- Reyes, A., Haynes, M., Hanson, N., Angly, F.E., Heath, A.C., Rohwer, F., and Gordon, J.I. (2010). Viruses in the fecal microbiota of monozygotic twins and their mothers. *Nature* 466, 334–338.
- Reyes, A., Wu, M., McNulty, N.P., Rohwer, F.L., and Gordon, J.I. (2013). Gnotobiotic mouse model of phage-bacterial host dynamics in the human gut. *Proc. Natl. Acad. Sci. U S A.* 110, 20236–20241.
- Reyes, A., Blanton, L.V., Cao, S., Zhao, G., Manary, M., Trehan, I., Smith, M.I., Wang, D., Virgin, H.W., Rohwer, F., et al. (2015). Gut DNA viromes of Malawian twins discordant for severe acute malnutrition. *Proc. Natl. Acad. Sci. U S A.* 112, 11941–11946.
- Roux, S., Krupovic, M., Poulet, A., Debroas, D., and Enault, F. (2012). Evolution and diversity of the Microviridae viral family through a collection of 81 new complete genomes assembled from virome reads. *PLoS One* 7, e40418.
- Roux, S., Hallam, S.J., Woyke, T., and Sullivan, M.B. (2015). Viral dark matter and virus–host interactions resolved from publicly available microbial genomes. *Elife* 4, e08490.
- Schloss, P.D., Westcott, S.L., Ryabin, T., Hall, J.R., Hartmann, M., Hollister, E.B., Lesniewski, R.A., Oakley, B.B., Parks, D.H., and Robinson, C.J. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.* 75, 7537–7541.
- Shkoporov, A., Khokhlova, E.V., Fitzgerald, C.B., Stockdale, S.R., Draper, L.A., Ross, R.P., and Hill, C. (2018). Φ CrAss001, a member of the most abundant bacteriophage family in the human gut, infects *Bacteroides*. *bioRxiv*. <https://doi.org/10.1101/354837>.
- Smith, M.I., Yatsunenko, T., Manary, M.J., Trehan, I., Mkakosya, R., Cheng, J., Kau, A.L., Rich, S.S., Concannon, P., Mychaleckyj, J.C., et al. (2013). Gut microbiomes of Malawian twin pairs discordant for kwashiorkor. *Science* 339, 548–554.
- Söding, J., Biegert, A., and Lupas, A.N. (2005). The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res.* 33, W244–W248.
- Stachler, E., Kelty, C., Sivaganesan, M., Li, X., Bibby, K., and Shanks, O.C. (2017). Quantitative CrAssphage PCR assays for human fecal pollution measurement. *Environ. Sci. Technol.* 51, 9146–9154.
- Sullivan, M.J., Petty, N.K., and Beatson, S.A. (2011). Easyfig: a genome comparison visualizer. *Bioinformatics* 27, 1009–1010.
- Tremaroli, V., and Bäckhed, F. (2012). Functional interactions between the gut microbiota and host metabolism. *Nature* 489, 242.
- Vollmers, J., Wiegand, S., and Kaster, A.-K. (2017). Comparing and evaluating metagenome assembly tools from a microbiologist's perspective—not only size matters! *PLoS One* 12, e0169662.
- Ward, J.H.J. (1963). Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.* 58, 236–244.
- Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis* (Springer-Verlag).
- Yutin, N., Makarova, K.S., Gussow, A.B., Krupovic, M., Segall, A., Edwards, R.A., and Koonin, E.V. (2018). Discovery of an expansive bacteriophage family that includes the most abundant viruses from the human gut. *Nat. Microbiol.* 3, 38–46.
- Zhong, X., Guidoni, B., Jacas, L., and Jacquet, S. (2015). Structure and diversity of ssDNA Microviridae viruses in two peri-alpine lakes (Annecy and Bourget, France). *Res. Microbiol.* 166, 644–654.
- Zimmermann, L., Stephens, A., Nam, S.-Z., Rau, D., Kübler, J., Lozajic, M., Gabler, F., Söding, J., Lupas, A.N., and Alva, V. (2017). A completely reimplemented MPI bioinformatics toolkit with a new HHpred server at its core. *J. Mol. Biol.* 430, 2237–2243.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological Samples		
crAssphage rich faeces – donor subject ID 924	This study	APC055 924
Chemicals, Peptides, and Recombinant Proteins		
Phosphate buffered saline	Sigma-Aldrich	Cat# P4417
L-Cysteine hydrochloride monohydrate	Sigma-Aldrich	Cat# 168149
Resazurin sodium salt	Sigma-Aldrich	Cat# R7017
YCFA broth	(Duncan et al., 2002)	N/A
D-(+)-Glucose (≥ 99.5% (GC))	Sigma-Aldrich	Cat# G8270
Soluble starch	Sigma-Aldrich	Cat# S9765
D-(+)-Cellobiose	Sigma-Aldrich	Cat# 22150
D-(+)-Maltose monohydrate	Sigma-Aldrich	Cat# M5885
Hydrochloric acid	Sigma-Aldrich	Cat# H1758
Sodium chloride	Sigma-Aldrich	Cat# S7653
Sodium hydroxide	Sigma-Aldrich	Cat# 221465
Polyethylene glycol 8000 (PEG 8000)	Sigma-Aldrich	Cat# P2139
Trizma base	Sigma-Aldrich	Cat# T6066
Magnesium sulfate heptahydrate	Sigma-Aldrich	Cat# 230391
Magnesium chloride hexahydrate	Sigma-Aldrich	Cat# M9272
Calcium chloride	Sigma-Aldrich	Cat# 793639
DNase (TURBO)	Biosciences	Cat# AM2239
RNase1	Fisher Scientific	Cat# 10568930
Proteinase K from Tritirachium album	Sigma-Aldrich	Cat# 2308
Guanidine thiocyanate solution	Sigma-Aldrich	Cat# 50983
Sodium citrate tribasic dehydrate (for molecular biology >99%)	Sigma-Aldrich	Cat# C8532
N-Lauroylsarcosine sodium salt	Sigma-Aldrich	Cat# 5125
2-Mercaptoethanol	Sigma-Aldrich	Cat# 6250
Chloroform, contains approximately 0.75% ethanol as preservative, for molecular biology, ≥ 99%	Fisher Scientific	Cat# 10727024
Phenol/chloroform/isoamyl alcohol, 25:24:1 mixture, pH 6.7/8.0, ≥ 99.0%	Fisher Scientific	Cat# 10306413
Caesium Chloride	Sigma-Aldrich	Cat# C4036
NuPAGE MOPS SDS Running Buffer (20X)	Thermo Fisher	Cat# NP0001
Critical Commercial Assays		
DNeasy Blood and Tissue kit	Qiagen	Cat# 69506
Qubit dsDNA HS Assay Kit	Biosciences	Cat# Q32851
SuperScript IV First-Strand Synthesis System	Thermo Fisher	Cat# 18091200
Illustra GenomiPhi V2 DNA Amplification Kit	GE Healthcare Life Sciences	Cat# 25-6600-31
TOPO TA Cloning Kit for Sequencing	Thermo Fisher	Cat# K4575J10
SensiFAST SYBR No-ROX One-Step Kit	Bioline	Cat# BIO-72005
QIAamp Fast DNA Stool Mini Kit	Qiagen	Cat# 51604
Phusion High-Fidelity DNA Polymerase	Fisher Scientific	Cat# F-531L
TruSeq RNA Library Preparation Kit v2	Illumina	Cat# RS-122-2001

(Continued on next page)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Nextera XT Index Kit v2 Set D	Illumina	Cat# FC-131-2004
HiSeq	GATC	N/A
MiSeq	GATC	N/A
Qubit dsDNA BR Assay Kit	Biosciences	Cat# Q32850
Bolt 4-12% Bis-Tris Plus Gels, 10 well	Thermo Fisher	Cat# NW04120BOX
Oligonucleotides		
pCrAss-DNAPol-Fwd5 qPCR; 5'-GCCTATTGTTGCTCAAGCTATTGAA-3'	This study	N/A
pCrAss-DNAPol-Rev5 qPCR; 5'-ACAACAGAACCAGCTGCCAT-3'	This study	N/A
pCrAss-DNAPol-Fwd6 qPCR; 5'-AGTGGTCTTGCTCCNGAACAAATGG-3'	This study	N/A
pCrAss-DNAPol-Rev6 qPCR; 5'-AACCTCCAGTTGCAACAGTATAAGT-3'	This study	N/A
MiSeq341F; 5'-TCGTCCGCGAGCGTCAGATGTGTATAAGAGACAGCCTACGGGNGGCWGCAG-3'	(Klindworth et al., 2013)	N/A
MiSeq805R; 5'-GTCTCGTGGGCTCGGA GATGTGTATAAGAGACAGGACTACHVGGGTATCTAATCC-3'	(Klindworth et al., 2013)	N/A
Recombinant DNA		
pCR2.1::pCrAssDNApol5	This study	N/A
Software and Algorithms		
Trimmomatic (v0.32)	(Bolger et al., 2014)	http://www.usadellab.org/cms/?page=trimmomatic
SPAdes (v3.6.2)	(Bankevich et al., 2012)	https://github.com/ablab/spades/tree/spades_3.6.2
metaSPAdes (v3.10.0)	(Nurk et al., 2017)	https://omictools.com/metaspades-tool
BLAST (v2.2.28+)	(Altschul et al., 1997)	ftp://ftp.ncbi.nlm.nih.gov/blast/executables/
HHPred Web Server	(Zimmermann et al., 2017)	https://toolkit.tuebingen.mpg.de/#/tools/hhpred
Prodigal (v2.6.3)	(Hyatt et al., 2010)	https://github.com/hyattprodigal
OrthoMCL (v2.0)	(Li et al., 2003)	https://github.com/stajichlab/OrthoMCL
R Statistical Computing Software	The R Foundation	https://www.r-project.org/
R Package NbClust (v3.0)	N/A	https://github.com/cran/NbClust
Ward.D2 algorithm in R	(Ward, 1963)	N/A
tsne (v0.1-3) for R	(van der Maaten and Hinton, 2008)	https://cran.r-project.org/package=tsne
R package ggplot2 (v2.2.1)	(Wickham, 2009)	http://ggplot2.org/
Easyfig (v2.2.2)	(Sullivan et al., 2011)	https://github.com/mjsull/Easyfig
ARAGORN (v1.2.36)	(Laslett and Canback, 2004)	http://mbio-serv2.mbioekol.lu.se/ARAGORN
PhageTerm	(Garneau et al., 2017)	https://sourceforge.net/projects/phageterm/
Tablet (v1.17.08.17)	(Milne et al., 2010)	https://ics.hutton.ac.uk/tablet/
Bowtie2 (v2.3.0)	(Langmead and Salzberg, 2012)	https://sourceforge.net/projects/bowtie-bio/files/bowtie2/2.3.0/
Samtools (v0.1.19)	(Li et al., 2009)	http://samtools.sourceforge.net/
R Package Vegan (v2.4.3)	N/A	https://cran.r-project.org/web/packages/vegan/
PerlPrimer v1.1.21	N/A	http://perlprimer.sourceforge.net/
DADA2 (v1.6.0)	(Callahan et al., 2016)	https://github.com/benjjneb/dada2

(Continued on next page)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
USEACH (v8.1)	(Edgar, 2010)	https://www.drive5.com/usearch/download.html
mothur (v1.34.4)	(Schloss et al., 2009)	https://www.mothur.org/
SPINGO	(Allard et al., 2015)	https://github.com/GuyAllard/SPINGO

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and reagents should be directed to the Lead Contact, Colin Hill (c.hill@ucc.ie).

EXPERIMENT MODEL AND SUBJECT DETAILS**CrAss-like Phage Rich Faeces**

Ethics for the collection of faecal samples from consenting donor subject ID 924, according to study protocol APC055, were approved by the Cork Research Ethics Committee. The samples were collected (without fixative or preservative) in the volunteer's home and transported to the research facility at ambient temperature, avoiding exposure to heat. In general, all samples were processed into frozen standard inoculum immediately. The donor, denoted as subject ID 924, is a healthy female. Later metadata showed the subject suffered from gastritis and is vitamin B12 deficient. Recruitment of this individual was based on the consistent presence of crAss-like phages the faecal samples over a 12 month period.

METHOD DETAILS**Metagenomic Datasets and Contig Assemblies**

Sequencing reads from publicly available metagenomic datasets were downloaded from NCBI Sequence Read Archive (SRA) database. All published and unpublished metagenomic datasets that yielded crAss-like phage contigs, the DNA preparation protocol, the sequencing technology, the assembly program, and information related to contig nomenclature, are briefly described in [Table S1](#). All reads were processed using Trimmomatic v0.32 (Bolger et al., 2014) to remove adaptor sequences and to trim reads when the Phred quality score dropped below 30 for a 4bp sliding window. Trimmed reads were assembled using either SPAdes v3.6.2 (Bankevich et al., 2012) or metaSPAdes v3.10.0 (Nurk et al., 2017). Contigs from the assembly of 702 metagenomic samples were assigned a specific nomenclature, representing: [1] study/sample description, [2] SPAdes or metaSPAdes assembly, and [3] numerical rank of largest-to-smallest assembled contigs. The full list of contigs assembled in this study and available associated metadata are detailed in [Table S2](#).

Detection and Curation of crAss-like Phages

The detection of crAss-like phage contigs was performed as follows. The amino acid polymerase sequence of prototypical crAssphage (p-crAssphage; UGP_018, NC_024711.1) was queried using BLAST v2.2.28+ (Altschul et al., 1997) against a translated nucleotide database consisting of assembled metagenome contig sequences. The most conserved orthologous protein group detected in our initial putative crAss-like phage screening included p-crAssphage protein UGP_092, which was annotated through the HHPred homology and structural prediction web server (Söding et al., 2005; Zimmermann et al., 2017) as a phage terminase. This was then used as a second genetic signature of crAss-like phages and used in an additional BLAST search. All putative crAss-like phages selected for analysis met the following criteria: [1] a BLASTp hit against either p-crAssphage polymerase or terminase with an E-value less than 1E-05, [2] a BLAST query alignment length ≥ 350 bp, and [3] a minimum contig length of 70kb (representing near-complete crAss-like phage contigs).

Identification of crAss-like Phage Orthologous Proteins and Clusters

The encoded proteins of crAss-like phages were predicted using Prodigal v2.6.3 (Hyatt et al., 2010). Orthologous proteins shared between crAss-like phages were detected using OrthoMCL v2.0 using default parameters (Li et al., 2003). The presence/absence of orthologous proteins between crAss-like phages was initially converted into a binary count matrix where the percentage of shared orthologous proteins was calculated ([Figure 1B](#)). The optimum number of phage clusters was calculated using the percentage of shared homologous proteins using the NbClust v3.0 package for R. Hierarchical clustering was performed on the count matrix of percentage shared crAss-like phage orthologous proteins using Ward's minimum variance method ['Ward.D2' algorithm in R (Ward, 1963)]. The resulting dendrogram was cut at $k = 10$ based on the estimation of the number of crAss-like phage clusters ([Figure 1A](#)).

As a verification of the 10 predicted crAss-like phage clusters, the original abundance matrix of crAss-like phage orthologous proteins was used to calculate Euclidean distances between sequences. These distance variations were calculated using the t-SNE machine learning algorithm ['tsne' v0.1-3 for R; (van der Maaten and Hinton, 2008)] and plotted using ggplot v2.2.1 ([Figure 2](#)).

Genomic Comparisons of crAss-like Phages

Complete circular genomes were annotated automatically using VIGA (<https://github.com/EGTortuero/viga>, genetic code table 11 or 15) and then manually using HHPred suit (Zimmermann et al., 2017) against the following databases: PDB_mm_CIF70_28_July, Pfam-A_v31.0, NCBI_CD_v3.16, TIGRFAMs_v15.0 (Table S3 for details). Genome comparison image was generated with Easyfig v2.2.2 (Sullivan et al., 2011), using tBLASTx algorithm with the following parameters: e-value cutoff 0.001, length filter 30 (Figure 3). The presence of crAss-like phage tRNA-encoding sequences were detected using ARAGORN v1.2.36 (Laslett and Canback, 2004).

Conserved protein sequences were aligned using MUSCLE v3.8.31 and approximately-maximum-likelihood phylogenetic trees were generated using FastTree v2.1.7 with default parameters (Figure S2).

Alignment of Virome Metagenomic Reads to crAss-like Contigs

The quality filtered reads from 512 human faecal viromes (as subset of 702 viromes selected based on availability of sufficient meta-data) were then aligned to the set of 131 nonredundant crAss-like phage genomes (with <90% identity and/or <90% overlap between them) using Bowtie2 v2.3.0 (Langmead and Salzberg, 2012) using the end-to-end alignment mode. A count table of reads aligned to contigs was generated with Samtools v0.1.19, which was then imported into R v3.3.1 for statistical analysis.

Recruitment of a crAssphage Faecal Donor and Faecal Fermentation

Human faecal viromes from a number of ongoing studies sequenced using Illumina HiSeq and MiSeq platforms were screened for crAss-like phages by aligning the obtained sequencing reads against prototypical crAssphage NC_024711.1 using Bowtie2 v2.3.0. One individual (subject ID 924) was found to carry crAssphage consistently at levels exceeding 30% of the total number of reads over a one year period. A frozen standard inoculum (FSI) sample was processed as described by (O'Donnell et al., 2016) with the following modification: the sample was resuspended in 1X phosphate buffered saline (37 mM NaCl, 2.7 mM KCl, 8 mM Na₂HPO₄, and 2 mM KH₂PO₄), 0.05% (w/v) L-cysteine (Sigma Aldrich, Ireland) and (1 mg/L) resazurin (Sigma Aldrich, Ireland). The crAssphage-rich FSI was inoculated into 400 ml YCFA-GSCM broth in a 500 ml fermenter vessel at 5% (v/v). Fermentation media was prepared exactly as described by Duncan et al. (2002) with the addition of glucose (2 g/L), soluble starch (2 g/L), cellobiose (2 g/L) and maltose (2 g/L). Fermentation was performed in batch format at approximately 37°C for 51 hours. Dissolved oxygen was sustained at <0.1% by constantly sparging the vessel with anaerobic gas mix (80% (v/v) N₂, 10% (v/v) CO₂, 10% (v/v) H₂) and stirring at 200 rpm. Both 2M NaOH and HCl solutions were used to maintain pH at ~7. Samples were collected at the following time points; 0, 4, 21, 28, 45 and 51 hours. Collected samples were centrifuged at 4,700 rpm at +4°C for 10 minutes. The resulting supernatants were filtered once through a 0.45 µm pore syringe filter and stored at +4°C. Resultant pellets were stored at -80°C.

Extraction of Viral Nucleic Acids and Sequencing Library Preparation

Total virome extractions were performed on 0.45 µm pore filtered fermentation supernatants. Solid NaCl and polyethylene glycol 8000 were added to the filtrates to give a final concentration of 0.5M and 10% (w/v), respectively. After overnight incubation at +4°C samples were centrifuged at 4,700 rpm and +4°C for 20 minutes. The pellets were then resuspended in 400µl of SM buffer (1M Tris-HCl pH 7.5, 5M NaCl, 1M MgSO₄) and briefly vortexed with an equal volume of chloroform. This mixture was then centrifuged at 2,500g for 5 minutes using a standard desktop centrifuge. The resultant aqueous phase was then transferred into an Eppendorf to which 40µl DNase buffer (10mM CaCl₂ and 50mM MgCl₂) and 8U and 4U TURBO DNase (Ambion/Thermo Fisher Scientific) and RNase I (Thermo Fisher Scientific) were added, respectively. This was incubated at 37°C for 1 hour followed by an enzyme inactivation step at 70°C for 10 minutes. This was followed by the addition of 2µl proteinase K and 10% SDS and further incubation at 56°C for 20 minutes. Lastly, 100µl phage lysis buffer (4.5 M guanidiniumisothiocyanate, 44 mM sodium citrate pH 7.0, 0.88% sarkosyl, 0.72% 2-mercaptoethanol) was added to lyse the viral particles. The final incubation was carried out at 65°C for 10 minutes. The resulting lysates were lightly vortexed with an equal volume of phenol/chloroform/isoamyl alcohol 25:24:1 (Fisher Scientific) and were centrifuged at room temperature for 5 minutes at 8,000g. This was again repeated with the resulting aqueous phase. Following the second extraction, the aqueous phase was passed through a DNeasy Blood and Tissue Kit (Qiagen) for final viral nucleic acid purification. The wash steps were each repeated twice and the final elution was carried out in 50µl elution buffer. Viral DNA quantification was carried out with the Qubit HS DNA Assay Kit (Invitrogen/Thermo Fisher Scientific) in a Qubit 3.0 Fluorometer (Life Technologies). The viral nucleic acids were then subjected to reverse transcription using SuperScript IV Reverse Transcriptase (RT) kit (Invitrogen/Thermo Fisher Scientific). The protocol was carried out exactly as described in the manufacturer's protocol for random hexamer primers. Following this, 1µl of the reversed transcribed viral DNA was subjected to GenomiPhi V2 (GE Healthcare) Multiple Displacement Amplification (MDA). Finally, MDA and non-MDA viral DNA was prepared for sequencing using TruSeq DNA Library Preparation Kit (Illumina, Ireland). All steps were performed as per the manufacturer's instructions. Prepared libraries were sequenced on an Illumina HiSeq platform (Illumina, San Diego, California) with 2x300bp paired-end chemistry at GATC Biotech AG, Germany. Reads were filtered, trimmed and assembled into contigs as described above. A count matrix was created by aligning quality-filtered reads back to contigs using Bowtie2 and Samtools.

P-crAssphage PCR Detection

Oligonucleotide primer pairs were designed based on the p-crAssphage DNA polymerase sequence UGP_018 (Dutilh et al., 2014) using PerlPrimer software (Marshall, 2004). Primer sequences are as follows: pCrAss-DNAPol-Fwd5 5'-GCCTATTGTTGCTCAAGC

TATTGAA-3' and pCrAss-DNApol-Rev5 5'-ACAACAGAACCAGCTGCCAT-3'. PCR products were cloned into pCR2.1-TOPO TA vector (Thermo Fisher Scientific) and obtained plasmids at known concentrations were used to establish calibration curves through serial ten-fold dilutions. This plasmid was denoted as pCR2.1::pCrAssDNApol5. Subsequently, qPCR were run in 15 μ l reaction volumes using SensiFAST SYBR No-ROX mastermix (Bioline) and LightCycler 480 thermocycler with the following conditions: initial denaturation at 95°C for 5 minutes, then 35 cycles of 94°C for 20 seconds, 55°C for 20 seconds and 72°C for 20 seconds, with a final extension at 72°C for 7 minutes. All samples were run in triplicate and the standard error was determined following calculation of DNA concentration based on the above standard curve.

Electron Microscopy and Detection of crAss-like Phage Proteins

A virus-enriched fraction of the crAssphage positive faecal sample, collected from subject ID 924, was prepared for electron microscopy imaging as follows. A 1:20 suspension (w/v) of faeces was prepared in SM buffer followed by vigorous vortexing until homogenised. The homogenised sample was chilled on ice for 5 minutes prior to centrifugation twice at 4,700 rpm for 10 minutes at +4°C. The resulting supernatant was then filtered twice through a 0.45 μ m pore syringe filters. The filtrate was ultra-centrifuged at 120,000g for 3 hours using a F65L-6x13.5 rotor (Thermo Scientific). The resulting pellets were resuspended in 5 ml SM buffer. The viral suspensions were ultracentrifuged again by overlaying them onto a caesium chloride (CsCl) step gradient of 5M and 3M, followed by centrifugation at 105,000g for 2.5 hours. A band of viral particles visible under side illumination was collected and buffer-exchanged using 3 sequential rounds of 10-fold diluting and concentrating to the original volume by ultra-filtration using AmiconCentrifugal Filter Units 10,000 MWCO (Merck). The purified fraction was then analysed by qPCR for the presence of crAssphage as described above. Following this, 5 μ l aliquots of the viral fraction were applied to Formvar/Carbon 200 Mesh, Cu grids (Electron Microscopy Sciences) with subsequent removal of excess sample by blotting. Grids were then negatively contrasted with 0.5% (w/v) uranyl acetate and examined at UCD Conway Imaging Core Facility (University College Dublin, Dublin, Ireland) by transmission electron microscope. The faecal viral fraction from subject ID 924 was further concentrated using Amicon Ultra-0.5 Centrifugal Filter Unit with 3 kDa MWCO membrane (Merck, Ireland). This concentrated fraction was loaded onto a premade Bolt 4-12% Bis-Tris Plus reducing SDS-PAGE gel (Invitrogen) and separated at 200 V for 30 minutes using 1X NuPAGE MOPS SDS Running Buffer. Six brightest bands with approximate molecular weights of 28, 35, 45, 55, 120 and 200 kDa were excised and subjected to MALDI-TOF/TOF (Bruker ultraflex III) protein identification following in-gel trypsinization, at Metabolomics & Proteomics Technology Facility (University of York, York, UK).

16S rRNA Gene Library Preparations

Total DNA was extracted from the pellets formed following centrifugation of fermentation samples. This was carried out using the QIAamp Fast DNA Stool Mini Kit (Qiagen, Hilden, Germany). All steps were carried out as per the manufacturer's protocol with the addition of a bead-beating step to aid total DNA extraction from the bacterial cells. Approximately 200mg of each pellet was placed in a 2ml screw-cap tube containing a mixture of one 3.5 mm glass bead, a 200 μ l scoop of 1mm zirconium beads and a 200 μ l scoop of 0.1mm zirconium beads (ThistleScientific) with 1ml of InhibitEX Buffer. Bead-beating was carried out three times for 30 seconds using the FastPrep-24 benchtop homogeniser (MP Biomedicals). Between each bead-beating the samples were cooled on ice for 30 seconds. The samples were then lysed at 95°C for 5 minutes. All other steps were carried out as per the manufacturer's protocol. Following extraction of total bacterial DNA, the hypervariable regions of V3 and V4 16S ribosomal RNA genes were amplified from 15ng of the DNA using Phusion High-Fidelity PCR Master Mix (Thermo Fisher Scientific) and 0.2 μ M of each of the following primers, containing Illumina-compatible overhang adapter sequences: MiSeq341F: 5'-TCGTCGGCAG CGTCAGATGT GATAAGAGA CAGCCTACGG GNGGCWGCAG-3' and MiSeq805R 5'-GTCTCGTGGG CTCGGAGATG TGTATAAGAG ACAGGAC TAC HVGGGTATCT AATCC-3' (56) The PCR program was run as follows: 98°C for 30 seconds, 25 cycles of 98°C for 10 seconds, 55°C for 15 seconds and 72°C for 20 seconds, with a final extension of 72°C for 5 minutes. The amplicons were then purified using AgencourtAMPure XP magnetic beads (Beckman-Coulter) followed by a second PCR to attach dual Illumina Nextera indices using the Nextera XT index kit v2 (Illumina). Purification was performed once again and the libraries were quantified using a Qubit dsDNA HS Assay Kit. The libraries were then pooled in equimolar concentration and sent for sequencing on an Illumina MiSeq platform (Illumina, San Diego, California) at GATC Biotech AG, Germany. The quality of the raw reads were assessed with FastQC (v1.15) and initial quality filtering was performed using Trimmomatic v0.36. Filtered reads were imported into R (v3.4.3) for analysis with DADA2 v1.6.0. (Callahan et al., 2016) Further quality filtering and trimming (maxN of 0 and a maxEE of 2) was carried out on both the forward and reverse reads with only retention in cases of pairs being of sufficient high quality. Error correction was performed on forward and reverse reads separately and following this, reads were merged. The resulting unique Ribosomal Variant Sequences (RSVs) were subjected to further chimera filtering using USEARCH v8.1 (55) with the Chimera-Slayer gold database v20110519. The retained, high quality, chimera-free, RSVs were classified with the RDP-classifier in mothur v1.34.4 (Schloss et al., 2009) against the RDP database v11.4 (phylum to genus) and SPINGO (Allard et al., 2015) for species assignment. Plots were generated using the R package ggplot2 v2.2.1.

QUANTIFICATION AND STATISTICAL ANALYSIS

Alignment of Virome Metagenomic Reads to crAss-like Contigs

The count table generated from Samtools v0.1.19 was then imported into R v3.3.1 for statistical analysis. The β -diversity of crAss-like viral populations in human cohorts was visualized using PCoA plot based on Spearman rank distances ($D = 1 - \rho$, where ρ is

Spearman rank correlation coefficient of relative abundance of different crAss-like contigs between samples). Statistical analysis was performed using permutational multivariate analysis of variance (PERMANOVA) implemented in Vegan v2.4.3 package for R (Anderson, 2001) and non-parametric Kruskal-Wallis test.

Faecal Fermentations

Statistical analysis was performed on the qPCR data acquired for the technical triplicates set-up for each fermentation time point. This was done to ensure minimal variation and error. The analysis was carried using GraphPad Prism v7.0 software using the standard error of the mean (SEM). Based on these bars we can have high confidence in the precision of the mean p-crAssphage copies/ μ l calculated for each time point.

DATA AND SOFTWARE AVAILABILITY

The data and R scripts required to reproduce the analyses of this study are provided as [Data S1](#). The 249 crAss-like phage contigs analysed in this study are also provided as [Supplemental Information](#) within [Data S1](#).