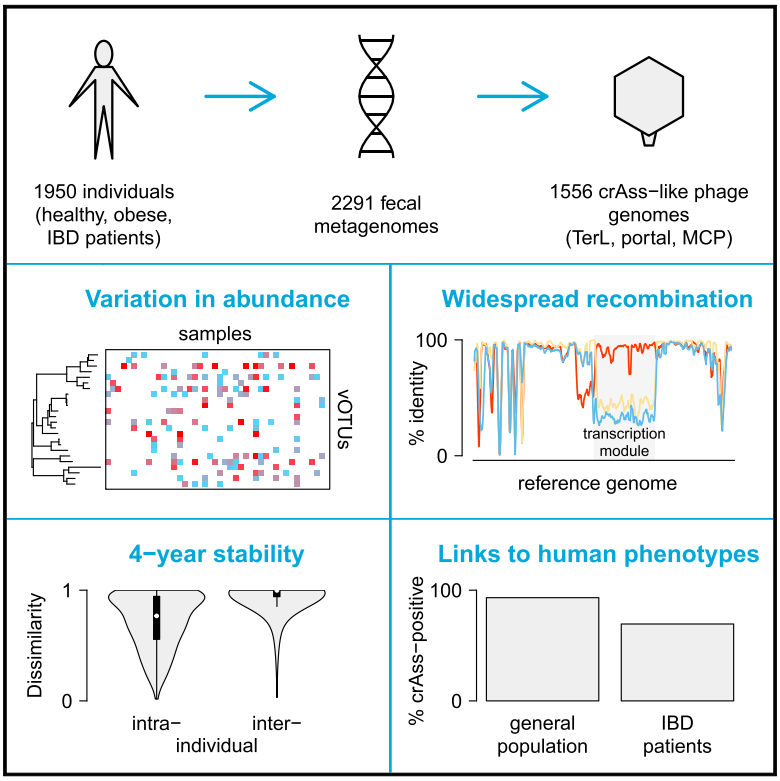
Cell Reports

Article

Discovery, diversity, and functional associations of crAss-like phages in human gut metagenomes from four Dutch cohorts

Graphical abstract



Highlights

- Detection of 125 crAss-like species-level clusters absent from reference databases
- Closely related crAss-like phages may possess divergent transcription gene modules
- crAss-like phageome of the human gut remains relatively stable after 4 years
- Gut crAss-like phages are depleted in inflammatory bowel disease patients

Authors

Anastasia Gulyaeva, Sanzhima Garmaeva, Renate A.A.A. Ruigrok, ..., Arnau Vich Vila, Alexander Kurilshikov, Alexandra Zhernakova

Correspondence

a.gulyaeva@umcg.nl (A.G.), a.zhernakova@umcg.nl (A.Z.)

In brief

Gulyaeva et al. analyze gut metagenomes of 1,950 individuals from general population and patient cohorts with a specific focus on crAss-like phages, including their genomic features, diversity, and ecology. Gut crAss-like phageome is shown to be relatively stable after a 4-year period and to be depleted in patients with inflammatory bowel disease.

Gulyaeva et al., 2022, Cell Reports *38*, 110204
January 11, 2022 © 2021 The Authors.
https://doi.org/10.1016/j.celrep.2021.110204

CellPress

Cell Reports

CellPress
OPEN ACCESS

Article

Discovery, diversity, and functional associations of crAss-like phages in human gut metagenomes from four Dutch cohorts

Anastasia Gulyaeva,[1,*] Sanzhima Garmaeva,[1,5] Renate A.A.A. Ruigrok,[1,2,5] Daoming Wang,[1,5] Niels P. Riksen,[3] Mihai G. Netea,[3] Cisca Wijmenga,[1] Rinse K. Weersma,[1,2] Jingyuan Fu,[1,4] Arnau Vich Vila,[1,2] Alexander Kurilshikov,[1] and Alexandra Zhernakova[1,6,*]

[1]Department of Genetics, University of Groningen, University Medical Center Groningen, Groningen 9713GZ, the Netherlands
[2]Department of Gastroenterology and Hepatology, University Medical Center Groningen, Groningen 9713GZ, the Netherlands
[3]Department of Internal Medicine, Radboud University Medical Center, Nijmegen 6525GA, the Netherlands
[4]Department of Pediatrics, University of Groningen, University Medical Center Groningen, Groningen 9713GZ, the Netherlands
[5]These authors contributed equally
[6]Lead contact
*Correspondence: a.gulyaeva@umcg.nl (A.G.), a.zhernakova@umcg.nl (A.Z.)
https://doi.org/10.1016/j.celrep.2021.110204

SUMMARY

The crAss-like phages are a diverse group of related viruses that includes some of the most abundant viruses of the human gut. To explore their diversity and functional role in human population and clinical cohorts, we analyze gut metagenomic data collected from 1,950 individuals from the Netherlands. We identify 1,556 crAss-like phage genomes, including 125 species-level and 32 genus-level clusters absent from the reference databases used. Analysis of their genomic features shows that closely related crAss-like phages can possess strikingly divergent regions responsible for transcription, presumably acquired through recombination. Prediction of crAss-like phage hosts points primarily to bacteria of the phylum Bacteroidetes, consistent with previous reports. Finally, we explore the temporal stability of crAss-like phages over a 4-year period and identify associations between the abundance of crAss-like phages and several human phenotypes, including depletion of crAss-like phages in inflammatory bowel disease patients.

INTRODUCTION

Our planet harbors an enormous diversity of viruses. There are millions of distinct species of cellular organisms, and all can likely be infected by viruses. The diversity of hosts, high mutation rate, and propensity to exchange genetic material with hosts and other viruses co-infecting a host cell all contribute to the unparalleled diversity of virus genomes. Remarkably, viruses employ a limited number of key conserved proteins responsible for genome replication and virion formation, allowing for comprehension and classification of the global virome (Koonin et al., 2020).

The advent of metagenomics led to a rapid expansion of the known viral sequence space, although it is still far from saturation. Due to the medical importance of the human gut ecosystem, its virome has become a major focus of metagenomics research, resulting in the discovery of multiple novel viruses, including the crAss-like phages. The first representative of these phages, the crAssphage *sensu stricto*, was discovered in 2014 and is believed to be one of the most abundant viruses in the human gut (Dutilh et al., 2014). This initial discovery was followed by the identification of five groups of crAss-like phages associated with the human gut—alpha/gamma, beta, delta, epsilon, and zeta—as well as diverse crAss-like phages associated with other environments (Guerin et al., 2018; Yutin et al., 2018, 2021).

The crAss-like phages are tailed phages with up to 192 kb circular (or terminally redundant) double-stranded DNA genomes. According to phylogenetic analysis of their conserved proteins, crAss-like phages form a monophyletic clade related to the prokaryotic virus order *Caudovirales*. There are three major functional regions that can be distinguished in the crAss-like phage genome: a replication gene module, a transcription gene module, and a capsid gene module (Koonin and Yutin, 2020; Yutin et al., 2018). Transcription of the three modules is initiated early (by a virion-packaged RNA polymerase [RNAP]), in the middle of, and late in the infection, respectively (Drobysheva et al., 2021; Holmfeldt et al., 2013; Shkoporov et al., 2018, 2021). The genomes of various crAss-like phages bear unique features, including reassignment of the stop codons TAG and TGA for amino acids and a high density of introns and inteins (Yutin et al., 2021).

Numerous crAss-like phages have been predicted to infect bacteria of the phylum Bacteroidetes (Dutilh et al., 2014; Yutin et al., 2021), and accordingly, several crAss-like phages have been isolated in cultures of Bacteroidetes hosts (Guerin et al., 2021; Hryckowian et al., 2020; Shkoporov et al., 2018), enriched

This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Table 1. Characteristics of the four cohorts used in the study

Cohort	Participant characteristics	Number of participants	Male (%)	Female (%)	Age (y)[a]	BMI (kg/m^2)
LLD	individuals from the general population	1,135	41.8	58.2	45.0 ± 13.6	25.3 ± 4.2
LLD follow-up	LLD cohort participants 4 years after initial LLD cohort data collection	338	44.4	55.6	51.7 ± 11.7	25.6 ± 4.0
300OB	individuals with BMI >27 kg/m^2	298	55.7	44.3	67.1 ± 5.4	30.7 ± 3.5
IBD	individuals with IBD	455[b]	40.7	59.3	42.8 ± 14.8	25.5 ± 5.0

[a]Data in the last two columns are provided as mean ± standard deviation.
[b]In total, 520 metagenomic samples from the inflammatory bowel disease (IBD) cohort were screened for the presence of crAss-like phage genomes, but data from 62 individuals with stoma and ileoanal pouches, one duplicated sample, and two samples without metadata were excluded from the statistical analysis, bringing the number of individuals to 455.

in the presence of Bacteroidetes hosts (Fitzgerald et al., 2021), or linked to Bacteroidetes hosts using a chromosome conformation capture approach (Marbouty et al., 2021), although there is also evidence that some crAss-like phages may infect hosts from different phyla (Yutin et al., 2021). Apparently, crAss-like phages may coexist with their hosts relatively peacefully, because the same crAss-like phages were often recurrently detected in metagenomic samples of longitudinal study participants (Edwards et al., 2019; Shkoporov et al., 2019; Siranosian et al., 2020), and four crAss-like phages were shown to persist in the cultures of their hosts for a long time without a significant reduction in the host population (Guerin et al., 2021; Shkoporov et al., 2018, 2021). Interestingly, hosts were shown to develop resistance to crAss-like phage infection via the phase variation mechanism: inversions in the bacterial genome loci responsible for cell surface coating biosynthesis were linked to altered cell surface architecture and a reduced ability of the bacteria to absorb phage (Shkoporov et al., 2021). The crAss-like phages are globally distributed in the human population, although the abundance of specific crAss-like phage taxa may vary depending on the geographic location (Camarillo-Guerrero et al., 2021; Edwards et al., 2019; Guerin et al., 2018). The abundance of certain crAss-like phages in the human gut metagenome has been shown to correlate with several health conditions, including advanced stages of HIV infection (Oude Munnink et al., 2014), inflammatory bowel disease (IBD) (Clooney et al., 2019), Parkinson's disease, atherosclerosis, liver cirrhosis, and ankylosing spondylitis (Tisza and Buck, 2021).

To further our understanding of the diversity, abundance, and functional role of the crAss-like phages in the human gut microbial community, we identified and analyzed crAss-like phages present in 2,291 metagenomic samples from four cohorts collected in the Netherlands. These include 1,135 baseline and 338 follow-up samples from the population cohorts Lifelines-DEEP (LLD) and LLD follow-up (Chen et al., 2021; Tigchelaar et al., 2015; Zhernakova et al., 2016), 298 samples from the 300-Obesity cohort (300OB) (Kurilshikov et al., 2019; Ter Horst et al., 2020), and 520 samples from a cohort of patients with IBD (Imhann et al., 2019; Vich Vila et al., 2018). Analysis of these samples led to identification of 1,556 crAss-like phage genomes, including 125 genomes prototyping species absent from the employed reference databases. We predicted hosts of crAss-like phages and found associations between the abundance of crAss-like phages and structural variations in bacterial genomes. Finally, we characterized the temporal dynamics and stability of crAss-like phages in longitudinal samples and explored links between the abundance of crAss-like phages and human metabolic, environmental, and health phenotypes.

RESULTS

Assembly and identification of crAss-like phage genomes

To identify crAss-like phage genomes in our gut metagenomic data, we assembled contigs for 2,291 metagenomic samples from four cohorts: 1,135 LLD, 338 LLD follow-up, 298 300OB, and 520 IBD samples (Table 1). Contigs of 10 kb or longer were retained for further analysis. The number of these contigs ranged from 0 to 3,036 per sample and was proportional to the number of sequencing reads available for each sample (Figures S1A and S1B).

To identify crAss-like phage genomes, we searched for contigs encoding three conserved capsid and genome-packaging proteins of crAss-like phages: terminase large subunit (TerL), portal protein, and major capsid protein (MCP). To enable a high-sensitivity search and account for the stop codon reassignments observed in some crAss-like phages (Yutin et al., 2021), we used both the protein sequences predicted under the standard bacterial genetic code and the full-length contig translations in six frames as search subjects. A contig was considered as a potential crAss-like phage genome fragment if either of the two types of search subjects was hit (HMMER, $E < 0.001$) by profiles (Yutin et al., 2021) of the three crAss-like phage proteins.

To verify the validity of our approach, we first applied it to four publicly available control datasets: a database of diverse manually annotated viral genomes (Viral RefSeq 202) (Brister et al., 2015), a database of 249 crAss-like phage genomes (DB249) (Guerin et al., 2018), a database of 673 crAss-like phage genomes (DB673) (Yutin et al., 2021), and a database of 146 nucleotide sequences (DB146) recognized as crAss-like by the NCBI Taxonomy resource (Schoch et al., 2020). As expected, only a few Viral RefSeq genomes were recognized as crAss-like phages (Table S1): three established crAss-like phages (Dutilh et al., 2014; Oude Munnink et al., 2014; Shkoporov et al., 2018), five bacteriophages of the fish pathogen *Flavobacterium psychrophilum* (phylum Bacteroidetes), and seven bacteriophages of the aquatic bacterium *Cellulophaga baltica* (phylum Bacteroidetes). The latter 12 bacteriophages were classified as belonging to the family *Podoviridae* at the time of discovery (Castillo and Middelboe, 2016; Holmfeldt et al., 2013). Also as

expected, almost all entries of the three crAss-specific databases (DB249, DB673, and DB146) were recognized as crAss-like phages. The only exceptions (Table S1) were 13 unrecognized entries from DB249 and DB673, all represented by seemingly partial genomes lacking the complete capsid gene module, and three unrecognized entries of DB146 representing two marine phages, *Cellulophaga* phage phi38:1 and phi40:1 (Holmfeldt et al., 2013), which likely form a lineage divergent from known crAss-like phages.

Following successful validation of our crAss-like phage detection approach in the control datasets, we applied it to the contigs of 10 kb or longer, which were assembled for each of the individual metagenomic samples from our cohorts. The crAss-like contigs were detected in 43% of LLD, 58% of LLD follow-up, 61% of 300OB, and 29% of IBD cohort samples, and their number ranged from one to six per sample. The correlation between the number of crAss-like contigs detected and the number of sequencing reads per sample was relatively weak ($r_{Spearman}$ = 0.182, p = 1.47 × 10^{-18}; Figure S1C). In total, we detected 1,556 crAss-like contigs. Notably, the sets of crAss-like contigs identified based on hits to individual protein queries and on hits to contig translations in six frames were almost identical: 1,553 of the 1,556 crAss-like contigs detected were detected by both methods. The lengths of the crAss-like contigs ranged from 10 to 196 kb, with a mean length of 59 kb. A considerable fraction of these contigs (255, 16%) had a 5′ end sequence of 10 or more nucleotides that perfectly matched the 3′ end sequence, tentatively indicating that these are fully sequenced circular or terminally redundant linear genomes (Table S2).

In order to assign the identified contigs to species-level virus operational taxonomic units (vOTUs), we pooled together the 1,556 newly detected crAss-like contigs with the DB249, DB673, and DB146 genomes and clustered under the 95% average nucleotide identity (ANI) over the 85% alignment fraction (AF) threshold (Roux et al., 2019). As a result, we delineated 378 vOTUs (Table S2). 151 of these vOTUs were composed entirely of crAss-like phage genomes from the databases, 102 included both genomes identified in this study and genomes from the databases, and 125 were composed entirely of the crAss-like phage genomes identified in this study (Table S2). 217 vOTUs contained tentatively complete genomes, which were selected to represent these vOTUs in subsequent analyses. The remaining 161 vOTUs were represented by the longest available genome fragments.

Next, we explored the genome organization of the 125 vOTUs composed entirely of the crAss-like phage genomes identified in this study. Their representative genomes and genome fragments possess characteristics similar to those of the previously discovered crAss-like phages, including the identifiable replication, transcription, and capsid gene modules, and the presence and order of the key conserved genes (Data S1). Notably, of the 125 vOTU representatives, 49 and 2 apparently utilize the TAG and the TGA stop codon reassignment for an amino acid, respectively (Data S1). Interestingly, all complete and nearly complete crAss-like phage genomes representing vOTUs possess a similar GC skew pattern, which was first reported for the ΦCrAss001 phage (Shkoporov et al., 2018): each strand of the genome is characterized by high GC skew values in its coding region and low GC skew values in the rest of the genome (Data S1). A similar but slightly less pronounced pattern was observed for the AT skew in most crAss-like phages outside of the alpha/gamma group (Data S1).

The 378 genomes representing species-level vOTUs were further clustered into 182 genus-level clusters under the 50% AF threshold (Adriaenssens and Brister, 2017). 90 of these clusters were composed entirely of crAss-like phage genomes from the databases, 60 included both genomes identified in this study and genomes from the databases, and 32 were composed entirely of crAss-like phage genome sequences identified in this study (Table S2). On TerL-based and portal-based phylogenetic trees, all of the vOTUs containing crAss-like phages identified in this study fell into the five known groups associated with the human gut: alpha/gamma, beta, delta, epsilon, and zeta (Figure 1; Table S2; Data S2). The names used to designate genus-level clusters in this study are composed of the name of the group to which a cluster belongs and a number, e.g., delta27.

Closely related crAss-like phages can possess highly divergent genome regions responsible for transcription

Following the delineation of vOTUs, we competitively mapped the sequencing reads from each sample to the 378 crAss-like phage genomes representing vOTUs. When analyzing the coverage of crAss-like phage genomes by sequencing reads, we noticed a common anomaly: the transcription gene module often received either zero coverage or, conversely, higher coverage than the rest of the genome in a fraction of analyzed samples (Figures 2A and 2B; Data S1 and S3). Although this anomaly was associated with multiple diverse crAss-like phage genomes, we selected one random genome, OLNE01000081, belonging to the delta27 genus-level cluster, to analyze the anomaly in-depth. The region of OLNE01000081 encoding RNAP and a protein homologous to VP02740, an uncharacterized transcription-related protein (Yutin et al., 2021), was not covered by reads in a fraction of samples, while in other samples, coverage of this region was comparable with the rest of the genome (Figures 2A and 2B; Data S3).

To investigate the cause of this coverage anomaly, we compared the OLNE01000081 genome with complete genomes belonging to the same vOTU (Figure 2C) and with complete and nearly complete genomes belonging to other delta27 vOTUs (Figure 2D). Although genomes belonging to the same vOTU were almost identical to the OLNE01000081 genome in all regions except the region encoding capsid proteins, where islands of high conservation are interspersed with low conservation stretches (Figure 2C), genomes belonging to other delta27 vOTUs demonstrated a notable divergence from OLNE01000081 in some regions and very strong similarity in others. Most strikingly, nucleotide identity between the OLNE01000081 genome and genomes from all other delta27 vOTUs (except NL_crAss001078) abruptly dropped to ~30% at the 5′ end of the RNAP gene and was abruptly restored to ~100% at the 3′ end of the VP02740 gene (Figure 2D). This pattern involves exactly the same region as the coverage anomaly described above and is consistent with the substitution of the genome region in question via recombination (Lole et al., 1999).

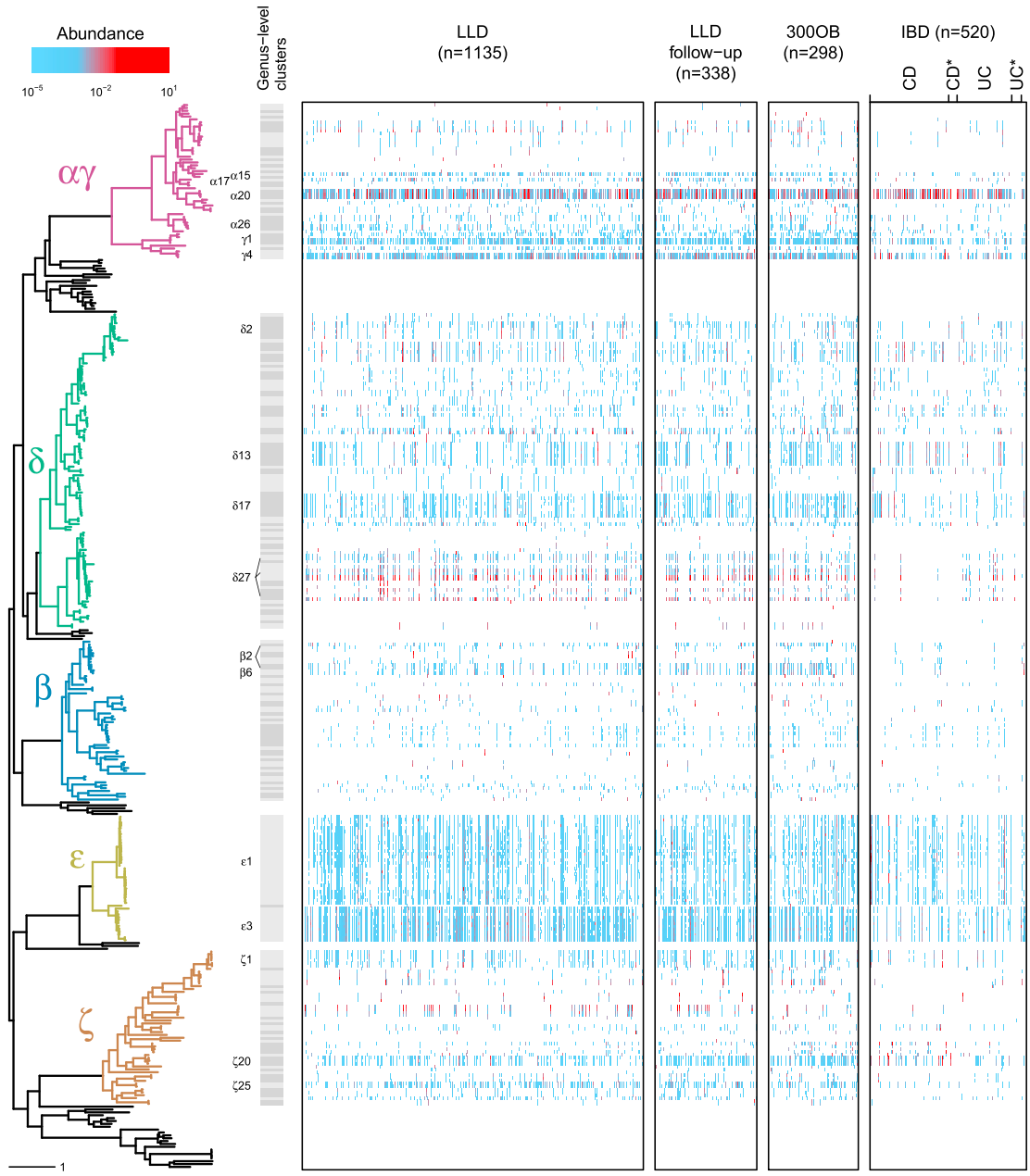


Figure 1. Abundance of crAss-like phages in the four cohorts

A TerL-based phylogenetic tree is juxtaposed with four abundance matrices representing the four cohorts. The tree is midpoint pseudo-rooted, and tree tips correspond to crAss-like phages representing vOTUs. Seven vOTUs represented by partial genomes lacking a TerL gene are not shown. The five clades of crAss-like phages associated with the human gut are designated by the color of the branches. Genus-level clusters of the crAss-like phages belonging to the five groups associated with the human gut are depicted by the gray bars next to the tree, with names indicated for genus-level clusters present in >10% of samples in any cohort. See Data S2 for the detailed annotation of the tree. Rows of the matrices correspond to crAss-like phage vOTUs and are ordered according to their position on the phylogenetic tree. Columns of the matrices correspond to samples and are ordered according to the age of the individuals (LLD and LLD follow-up cohorts), body mass index (BMI) (300OB cohort), and diagnosis (IBD cohort). The color of a matrix element indicates estimated vOTU abundance in a sample. Asterisk (*) next to the diagnosis abbreviation indicates that the individual has a stoma. CD, Crohn's disease; UC, ulcerative colitis.

A phylogenetic tree reconstructed based on the transcription gene module shows that there are four distinct types of this region within the delta27 genus-level cluster (Figure 3). The origin of the type present in OLNE01000081 remains unknown, because no close (>20% gene length coverage by BLASTN hits with E < 0.001) non-delta27 homologs of either RNAP or VP02740 genes were detected in crAss-like phage genomes and GenBank entries (Sayers et al., 2021).

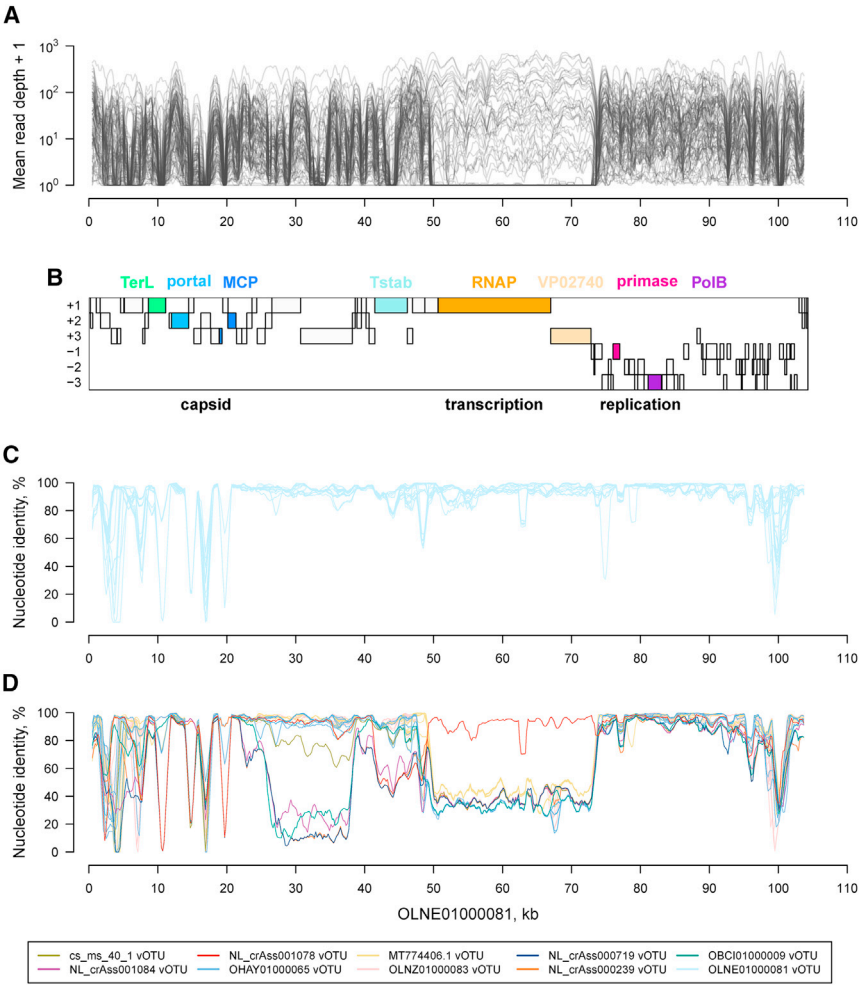


Figure 2. OLNE01000081 genome properties

(A) Depth of the OLNE01000081 genome coverage by sequencing reads from the LLD cohort. Each transparent gray line corresponds to a sample and represents mean depth in a 1,001-nt sliding window.

(B) OLNE01000081 genome organization. A genome is presented as a black rectangular contour, and three forward and three reverse frames of the genome are indicated. Open reading frames (ORFs) translated under the TAG stop codon reassignment for glutamine are presented as bars. Products of the ORFs shown in color are indicated above the genome map: MCP, major capsid protein; PolB, DNA polymerase family B; RNAP, RNA polymerase; TerL, terminase large subunit; Tstab, tail stabilization protein; VP02740, uncharacterized transcription-related protein (Yutin et al., 2021). Functional regions of the genome are indicated below the genome map.

(C) Nucleotide identity between the OLNE01000081 and other tentatively complete genomes belonging to the same vOTU.

(D) Nucleotide identity between the OLNE01000081 genome and (nearly) complete genomes belonging to other vOTUs of the delta27 genus-level cluster. Data for genomes belonging to different vOTUs are distinguished by color (see inset). Nucleotide identity presented in (C) and (D) was calculated using a 1,001-nt sliding window.

The fact that besides the delta group, the coverage anomaly associated with the transcription gene module was also observed in alpha/gamma, beta, and zeta groups indicates that the presence of highly divergent transcription gene modules in closely related crAss-like phages may be a common phenomenon. We further confirmed the link between the coverage anomaly and nucleotide sequence divergence in the genus-level clusters alpha6, beta8, and zeta9 belonging to the three latter groups (Data S4).

Prevalence and temporal stability of crAss-like phages

Competitive mapping of sequencing reads to crAss-like phage genomes representing vOTUs allowed us to estimate the abundance of crAss-like phages in metagenomic samples. The percentage of filtered and quality-trimmed sequencing reads mapping to the crAss-like phage genomes varied from 0% to 20.7% per sample, with a mean of 0.4% in each of the four cohorts (Figure S2).

A vOTU was considered to be detected in a sample if >10% of the length of its representative sequence was covered by reads. Based on this criterion, all but 17 vOTUs belonging to the five groups associated with the human gut were detected in at least one sample in one of the four cohorts, and no vOTUs outside of these five groups were detected in any sample (Figure 1). The correlation between the number of detected vOTUs and the number of sequencing reads per sample was relatively weak ($r_{Spearman}$ = 0.213, p = 6.28 × 10^{-25}; Figure S1D).

Almost universally, there was a correlation between detection of vOTUs belonging to the same genus-level cluster (Figure 1). This pattern is likely explained by misalignment of reads to reference genomes representing vOTUs closely related to, but different from, the vOTU of their origin. Because it complicates the vOTU-level analysis of the abundance of crAss-like phages, we decided to conduct all the subsequent analyses at the genus and higher taxonomic levels.

A total of 118 of the 132 crAss-like phage genus-level clusters from groups associated with the human gut were detected in at least one sample, and 1,998 of the 2,291 samples contained at least one crAss-like phage. On average, 4.3, 4.7, 5.3, and 1.9 crAss-like phage genus-level clusters were detected in LLD, LLD follow-up, 300OB, and IBD cohort samples, respectively. The most abundant clusters were alpha20, which contains crAssphage *sensu stricto* (Dutilh et al., 2014), and delta27 (Figure S3). The most frequently detected clusters were alpha20, gamma1, gamma4, epsilon1, and epsilon3 (detected in >20% of samples in at least one cohort; Figure S3).

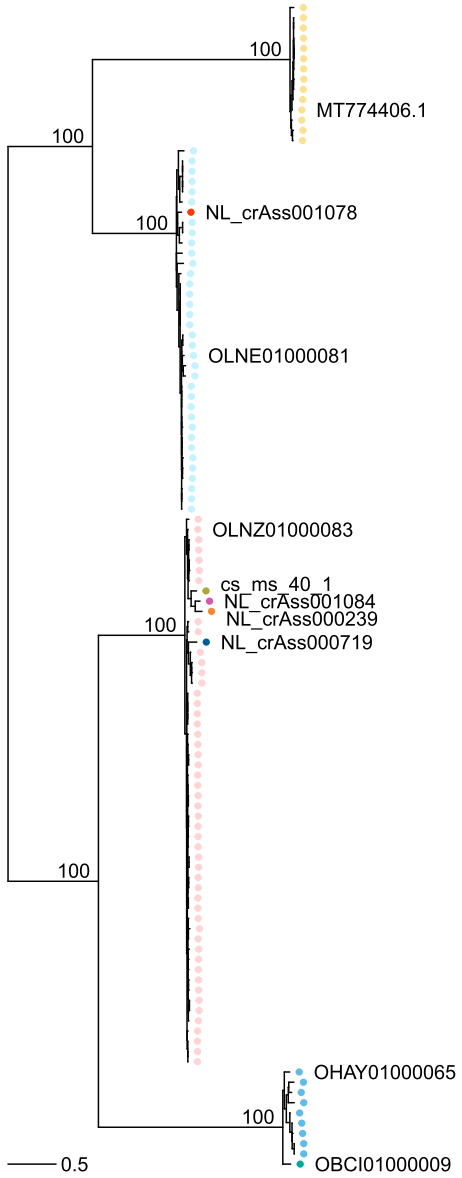


Figure 3. Phylogenetic tree based on the transcription gene module of the delta27 crAss-like phages

A total of 114 delta27 crAss-like phages with (nearly) complete genome sequences are included. The vOTU of each tree tip is indicated by color. Tree tips corresponding to genomes representing vOTUs are labeled. Bootstrap support is specified for long branches.

To study the temporal stability of the crAss-like phages, we used the data from the LLD (Zhernakova et al., 2016) and LLD follow-up (Chen et al., 2021) cohorts: samples collected from the same 338 individuals at two time points approximately 4 years apart. First, we assessed the stability of individual genus-level clusters of crAss-like phages. Only clusters detected in >10% of LLD or LLD follow-up samples were considered. For all of these clusters, the number of individuals positive for a cluster at both time points was considerably higher than what would be expected if the crAss-like phage composition at the two time points was completely independent (Figure 4A). The overall composition of the crAss-like fraction of the virome was also relatively stable. The dissimilarity between samples collected from the same individual at the two time points, quantified using Bray-Curtis (BC) dissimilarity measure, was smaller than the dissimilarity between pairs of samples collected from different individuals at the same time point (mean intra-individual BC dissimilarity, 0.72; mean inter-individual BC dissimilarity in LLD cohort, 0.94; Wilcoxon signed-rank test, p = 2.06e−90) (Figure 4B). Thus, we observed that, on average, people are more similar to themselves over time than they are to unrelated individuals at the same time point.

Potential hosts of crAss-like phages and relation to bacterial structural variants

Next, we predicted the potential hosts of the crAss-like phages using two approaches: we analyzed matches between microbial CRISPR spacers and crAss-like phage genomes, and we explored the co-abundance of crAss-like phages and gut microbes in the samples from the four cohorts.

Sequence similarity of CRISPR spacers in microbial genomes to regions of crAss-like phage genomes allowed us to link crAss-like phages from 32 genus-level clusters to 10 bacterial genera (Figure 5; Table S3). Many of these links were previously reported by Yutin et al. (2021). The majority (98%) of the links identified bacteria from the genera *Bacteroides*, *Prevotella*, *Porphyromonas*, and *Parabacteroides*, all belonging to the order Bacteroidales, as potential hosts, consistent with the prediction made by the discoverers of the crAssphage *sensu stricto* (Dutilh et al., 2014). However, we also identified bacteria from phyla Firmicutes and Fusobacteria as potential hosts of a small number of crAss-like phages. There does not seem to be a one-to-one correspondence between groups of crAss-like phages and bacterial genera, because the alpha, delta, and epsilon groups of crAss-like phages were each linked to three out of the four Bacteroidales genera mentioned above (Figure 5).

To detect correlation between the abundance of crAss-like phages and microbes, which may be a sign of a phage-host relationship, we calculated the Spearman correlation coefficient for pairs of phage and microbe taxa in each cohort, followed by meta-analysis of the results from three independent cohorts: LLD, 300OB, and IBD cohorts (Table S3). The meta-analysis identified 19 pairs of taxa characterized by an absolute value of the correlation coefficient >0.2. All of these correlations were positive, and 16 of them involved bacteria from the order Bacteroidales, while the remaining three were with bacteria from the phyla Firmicutes and Actinobacteria. The phage-microbe pair with the highest absolute value of the correlation coefficient in each cohort and meta-analysis (r_{meta} = 0.447, p = 1.29 × 10^{-78}) was crAss-like phage genus-level cluster delta27 and bacterial species *Prevotella copri* (Figure S4; Table S3). This observation, together with the fact that a number of delta27 crAss-like phages were linked to *P. copri* based on CRISPR analysis (Figure 5; Table S3), identifies *P. copri* as a likely host of delta27 crAss-like phages. Overall, the results of the two methods of host prediction were consistent, both with each other and with previous studies, and identified members

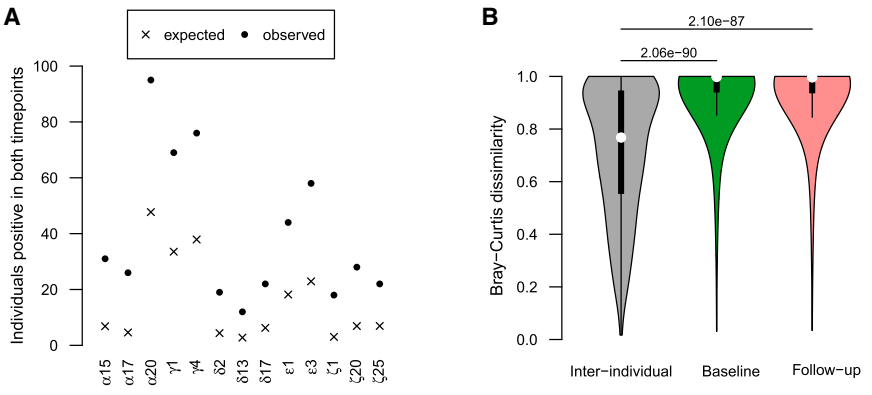


Figure 4. Comparison of the crAss-like phage composition in LLD and LLD follow-up longitudinal samples

(A) Expected and observed numbers of individuals positive for genus-level clusters of crAss-like phages at both time points. The 13 clusters detected in >10% of LLD or LLD follow-up samples were considered. The number of individuals expected to be positive for a cluster at both time points if the crAss-like phage composition at the two time points was completely independent was calculated as 338 × n × m, where 338 is the number of individuals under consideration, and n and m are the fractions of individuals positive for the cluster at the first and second time point, respectively.

(B) Intra- and inter-individual dissimilarity of the crAss-like phage assemblage. A distribution of the Bray-Curtis dissimilarities, quantifying the dissimilarity of the composition between samples, is shown for pairs of samples collected from the same individual at the baseline and follow-up time points (gray) and for pairs of samples collected from different individuals at the same time point (green and red). p values of the one-sided Wilcoxon signed-rank tests comparing the distributions are indicated above the violin plots.

of the phylum Bacteroidetes as the major hosts of the crAss-like phages associated with the human gut.

Next, we explored associations between the abundance of crAss-like phages and structural variants (SVs) in bacterial genomes (Wang et al., 2021; Zeevi et al., 2019). Based on the collection of samples from the three independent cohorts, LLD, 300OB, and IBD, we detected 7,999 deletion SVs (dSVs, genome segments absent in 25%–75% samples) and 3,559 variable SVs (vSVs, genome segments absent in <25% of samples). We analyzed associations between the detected SVs and the abundance of crAss-like phages in each of the three cohorts separately and integrated the results via meta-analysis and heterogeneity analysis. In total, we identified 22 replicable significant SV associations (19 vSV associations and 3 dSV associations) with the relative abundance of crAss-like phages at the genus and higher taxonomic levels (Table S4). The crAss-like phage taxonomic clusters involved in these associations included the entire crAss-like phage assemblage, groups beta and delta, and genus-level clusters alpha15, gamma1, gamma4, delta27, and epsilon1. All the bacteria involved in these associations belonged to the phyla Actinobacteria and Firmicutes. The SVs in these associations ranged from 1 to 14 kb, and according to proGenomes database (Mende et al., 2017) annotation, encoded functionally diverse proteins, including DNA primases and components of the adenosine triphosphate-binding cassette (ABC)-type transport system (Table S4).

crAss-like phage associations with human phenotypes

To identify whether crAss-like taxonomic clusters are associated with intrinsic, dietary, and clinical phenotypes in a general population context, we analyzed 1,135 metagenomic samples and 207 phenotypes from the population-based LLD cohort. The analysis was conducted for 30 taxonomic clusters of crAss-like phages detected in >5% of LLD samples (crAss-like phage assemblage, five groups, 24 genus-level clusters) and utilized five different methods that demonstrated overall agreement (Table S5). Associations were considered significant at a false discovery rate (FDR) < 0.05.

When Spearman correlation analysis was applied to the crAss-like phage presence-absence data, the fecal level of the secretory protein chromogranin A (CgA), which is a precursor to peptides with regulatory, antimicrobial, and antifungal activities (Bartolomucci et al., 2011), was negatively correlated with the presence of crAss-like phage assemblage, groups alpha/gamma and beta, and genus-level clusters gamma1 and gamma4, while the fecal level of the antimicrobial peptide human beta-defensin-2 (HBD-2) was positively correlated with the presence of genus-level cluster gamma1. This analysis also linked crAss-like phages to three dietary phenotypes: frequency of coffee intake was positively correlated with the presence of crAss-like phage assemblage and group beta, meat intake was positively correlated with the presence of group epsilon and genus-level cluster epsilon1, and the amount of carbohydrates in the diet was positively correlated with the presence of genus-level cluster alpha20. In addition, the presence of the alpha/gamma group was negatively associated with irritable bowel syndrome (IBS) diagnosis, stomach ulcer diagnosis, and usage of laxatives (Figure 6; Table S5).

In contrast, when Spearman correlation analysis was applied to the crAss-like phage abundance data, only five associations were statistically significant (FDR < 0.05): the three associations with CgA level described above and two of the dietary phenotype associations described above (Table S5). Unadjusted logistic regression and logistic regression adjusted for the age and sex of cohort participants yielded almost identical results: six and seven associations described above (with the level of CgA and health conditions) were recognized as statistically significant, respectively (FDR < 0.05; Table S5). Interestingly, when the logistic regression was further adjusted for the log-transformed abundance of the bacterial genera *Bacteroides*, *Prevotella*, *Porphyromonas*, and *Parabacteroides*, which were most frequently identified as potential hosts of crAss-like phages based on CRISPR spacer matches (Figure 5), there were no significant associations (FDR < 0.05), and all of the associations described above were only nominally significant.

Overall, crAss-like phage taxonomic clusters were associated with a limited number of human phenotypes, and the strongest, most consistent between analyses were associations with the

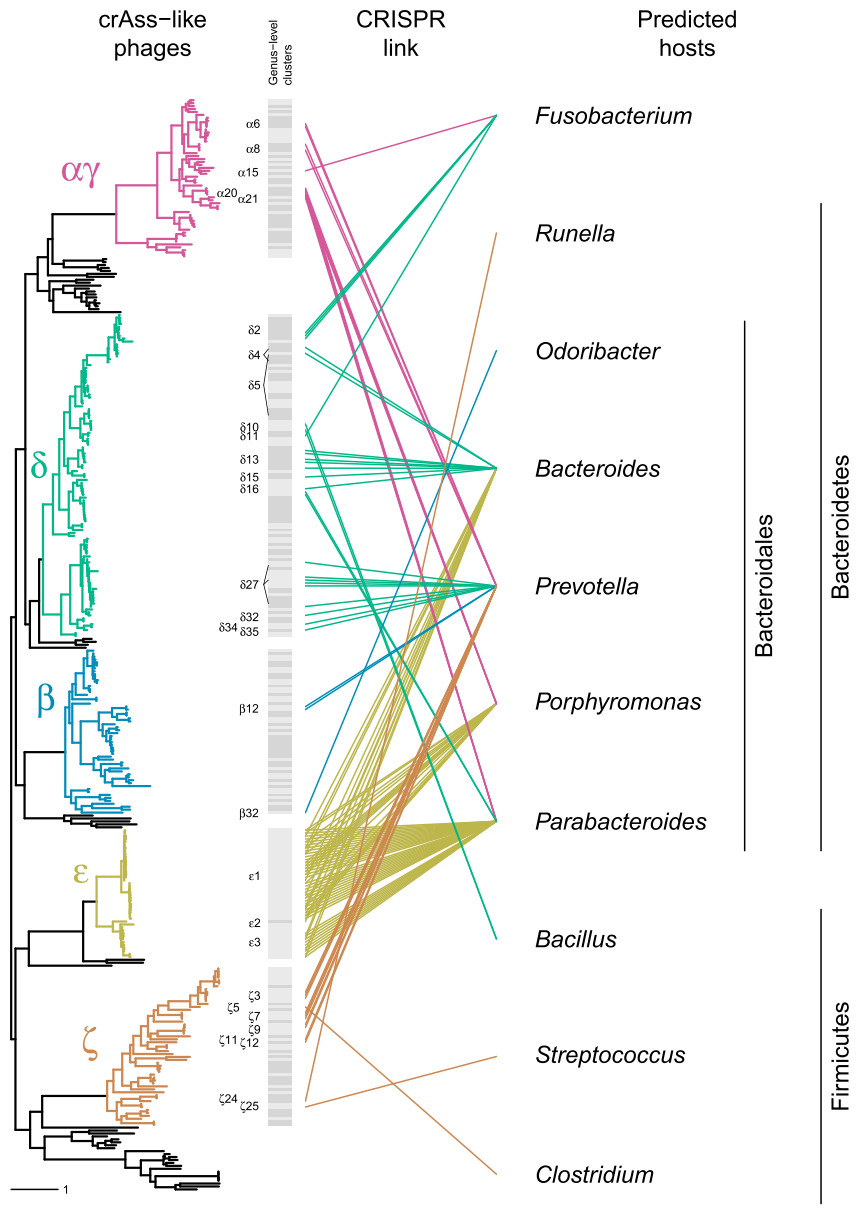


Figure 5. Links between crAss-like phages and bacterial genera based on CRISPR spacer matches

For details about the phylogenetic tree, see the legend of Figure 1.

level of CgA. This finding is in line with the massive effect of CgA on the gut bacteriome reported earlier for this cohort (Zhernakova et al., 2016). The associations were partially independent of the human age and sex but were influenced by the abundance of the potential host genera.

crAss-like phages in the context of IBD and obesity

Next, we aimed to explore the crAss-like phage composition in two different disease contexts: the metabolically altered environment represented by a cohort of overweight or obese individuals (300OB cohort; BMI > 27 kg/m^2) (Kurilshikov et al., 2019; Ter Horst et al., 2020) and the inflammatory environment represented by a cohort of patients with IBD (IBD cohort) (Imhann et al., 2019; Vich Vila et al., 2018). We studied the potential effect of the two disease contexts on the crAss-like phage composition in the gut by comparing the prevalence of crAss-like phage assemblage, five groups and 24 genus-level clusters, which were detected in >5% samples of the population-based LLD cohort, between the respective cohort and LLD. The analyses were performed using logistic regression, correcting for age and sex, as well as for carbohydrates and meat consumption, frequency of coffee consumption, usage of laxatives, and CgA and HBD-2 levels where applicable (see STAR Methods). The prevalence of crAss-like phage assemblage, five groups, and 19 genus-level clusters was significantly lower in IBD samples compared with LLD samples (FDR < 0.05; Table S6). All but two of these associations remained significant when the regression was further adjusted for the log-transformed abundance of the four potential host genera: *Bacteroides*, *Prevotella*, *Porphyromonas*, and *Parabacteroides* (FDR < 0.05; Table S6). In contrast, we observed no significant difference in the prevalence of any of the crAss-like phage taxonomic clusters between LLD and 300OB cohorts (FDR > 0.05; Table S6).

Previous reports have shown that the fecal bacterial composition differs in IBD patients with Crohn's disease (CD) versus ulcerative colitis (UC), as well as in IBD patients with varying disease locations (Franzosa et al., 2019; Vich Vila et al., 2018). To investigate whether these differences also extend to crAss-like phages, we performed differential prevalence analyses by comparing samples from IBD patients with (1) CD versus UC, and (2) an exclusively colonic versus an ileal-inclusive disease location. In both tests, we did not observe any significant differences between the test groups at any of the taxonomic levels (FDR > 0.05; Table S6).

Finally, to investigate whether metabolic syndrome is associated with the crAss-like phage prevalence, we stratified samples from the 300OB cohort based on the presence or absence of metabolic syndrome according to the National Cholesterol Education Program ATP III criteria (Ter Horst et al., 2020). We found no significant associations between the two groups at any of the taxonomic levels (FDR > 0.05; Table S6).

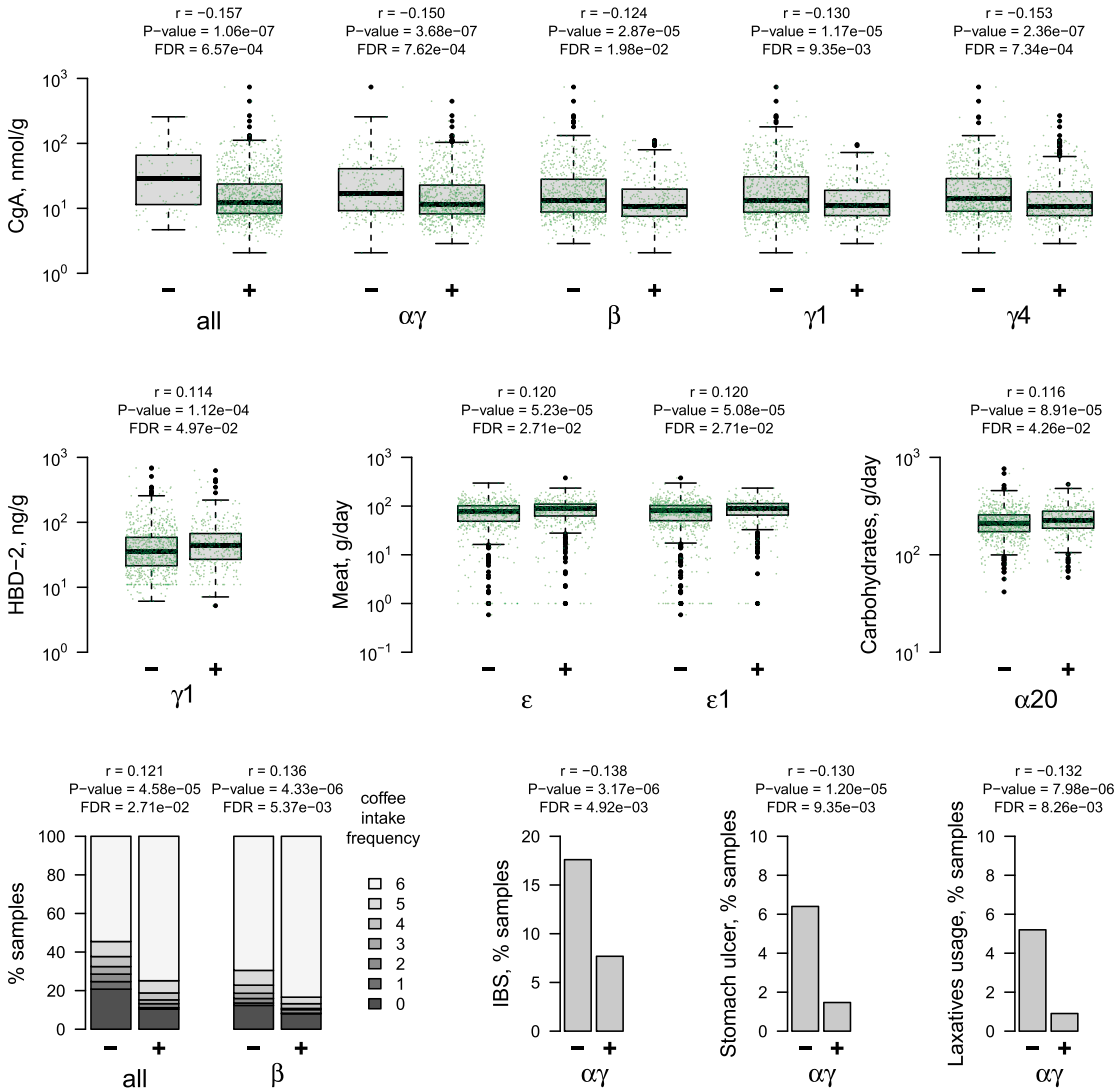


Figure 6. Statistically significant associations between LLD cohort phenotypes and presence of crAss-like phage taxonomic clusters

Fourteen associations, significant (FDR < 0.05) according to the Spearman correlation analysis, are shown. For each association, phenotype values are shown separately for the samples where a taxonomic cluster was (plus sign [+] on the x axis) and was not detected (minus sign [−]). Individual phenotype values are indicated by green points on top of the boxplots summarizing them. For individuals who do not consume meat, meat consumption was recorded as 1 g/day to obtain finite values after log transformation.

DISCUSSION

In this study, we relied on existing knowledge about the genomes and proteomes of crAss-like phages (Dutilh et al., 2014; Guerin et al., 2018; Yutin et al., 2018, 2021) to systematically analyze the diversity and abundance of crAss-like phages in gut metagenomic samples collected from participants of four cohorts set up in the Netherlands: LLD, LLD follow-up, 300OB, and IBD. Using the three conserved capsid and genome-packaging proteins of crAss-like phages (TerL, portal, and MCP) as markers, we were able to detect 1,556 crAss-like contigs. 76% of these contigs clustered into species-level vOTUs populated by previously discovered crAss-like phages, while the remaining 24% formed 125 vOTUs absent from the reference databases used. At the genus level, 96% of the newly discovered contigs clustered with the known crAss-like phages, while the remaining 4% formed 32 genus-level clusters absent from the reference databases used. Notably, all the crAss-like contigs we detected fell into the five known groups associated with the human gut (Guerin et al., 2018; Yutin et al., 2021). This may indicate that most groups of gut-associated crAss-like phages present in the Western European population have already been discovered, although, importantly, our detection method was based on information about the known groups and may be less sensitive for detection of other distantly related groups.

Interestingly, we observed that crAss-like phage genomes can be almost identical throughout a large portion of their length but possess strikingly different genome regions responsible for transcription. This phenomenon is likely to be widespread among crAss-like phages, because its manifestation—differential sequencing read coverage of the transcription gene module—was observed in multiple representatives of the alpha/gamma, beta, delta, and zeta groups of crAss-like phages. Lower coverage of this region compared with the rest of the genome can be explained by the presence of a crAss-like phage that is closely related to the reference genome used for read mapping but possesses a highly divergent transcription gene module. Higher coverage of this region compared with the rest of the genome can be explained by the presence of a distant crAss-like phage that possesses a transcription gene module similar to the one in the reference genome. The likely mechanism driving the emergence of diverse genome regions responsible for transcription among closely related crAss-like phages is recombination. Acquisition of a divergent functional region through recombination may be beneficial for a virus, for instance, allowing it to adapt to changing conditions. For example, non-segmented (+)ssRNA genomes of the *Enterovirus C* virus species include a type-specific capsid coding region and a non-structural protein coding region subject to frequent inter-typic homologous recombination, which may lead to the emergence of novel virus lineages (Smura et al., 2014). A similar mechanism of genetic diversification can be envisioned for the crAss-like phages.

Analysis of samples from 388 individuals taken at two time points (Chen et al., 2021; Zhernakova et al., 2016) allowed us to explore the temporal stability of crAss-like phages. Despite the 4-year period separating the two time points, we observed a degree of stability both at the level of individual genus-level clusters (the numbers of individuals positive for frequently detected genus-level clusters at both time points were considerably higher than what would be expected if the crAss-like phage composition at the two time points was completely independent) and at the level of the overall composition of the crAss-like fraction of the virome (on average, dissimilarity between samples collected from the same individual at the two time points was lower than the dissimilarity between pairs of samples collected from different individuals at the same time point). These observations are in line with previous reports of temporal stability of crAss-like phages (Edwards et al., 2019; Shkoporov et al., 2019; Siranosian et al., 2020), which may be explained by a relatively peaceful co-existence of phage and host (Guerin et al., 2021; Shkoporov et al., 2018, 2021).

Consistent with the results of previous studies (Dutilh et al., 2014; Fitzgerald et al., 2021; Guerin et al., 2021; Hryckowian et al., 2020; Marbouty et al., 2021; Shkoporov et al., 2018; Yutin et al., 2021), our analysis of matches between CRISPR spacers from microbial genomes and crAss-like phage genomes, as well as our analysis of co-abundance of microbial and crAss-like phage taxa, pointed primarily to bacteria of the phylum Bacteroidetes as the hosts of the crAss-like phages. Predicting the hosts of phages is challenging because CRISPR spacer sequences are short and may match phage genomes just by chance, while the CRISPR defense system itself was only detected in a fraction of microbial genomes (Edwards et al., 2016; Shmakov et al., 2017). Co-abundance analysis can detect a linear or monotonous relationship but may not be able to capture a more complex relationship, e.g., a time-lagged phage-host relationship (Edwards et al., 2016), or distinguish between a phage-host relationship and other ecological interactions. Combining multiple techniques of host prediction can increase confidence in the results. In this study, both CRISPR-based and co-abundance analyses strongly pointed to *Prevotella copri* as a host of the crAss-like phages belonging to the genus-level cluster delta27. The fact that the correlation between the relative abundance of the delta27 phages and their potential host is positive suggests that the dynamics in this phage-host pair may be similar to the co-existence in equilibrium reported for several crAss-like phages isolated in cultures of their hosts (Guerin et al., 2021; Hryckowian et al., 2020; Shkoporov et al., 2018, 2021).

The interplay between phages and bacteria may be influenced by structural variation in bacterial genomes. For instance, several studies have linked phase variation in the bacterial genome loci responsible for cell surface coating biosynthesis to acquisition of resistance to phage infection (Porter et al., 2020; Shkoporov et al., 2021). In our study, we identified 22 replicable significant associations between bacterial SVs and crAss-like phage taxonomic clusters. All the bacteria involved in these associations were from the phyla Actinobacteria and Firmicutes, while the majority of the taxonomic clusters of crAss-like phages involved were linked to potential hosts from the order Bacteroidales by CRISPR spacer matches. Consequently, the detected associations are unlikely to stem from phage-host relationships and may instead reflect yet unknown interactions in the broader ecological community of the gut.

Our analysis of the correlations between human phenotype levels and crAss-like phage taxonomic clusters in the population-based LLD cohort revealed several associations with intrinsic (levels of CgA and beta-defensin-2 secretory proteins in stool), dietary (coffee, meat, and carbohydrate intake), and health (IBS and stomach ulcer diagnosis, usage of laxatives) phenotypes. An association between a phage and a human phenotype may reflect the association between the microbial host of the phage and the phenotype in question. Although the exact hosts of the majority of the crAss-like phages remain unknown, both previous reports (Dutilh et al., 2014; Yutin et al., 2021) and our data point to the bacteria of the phylum Bacteroidetes, and in particular the order Bacteroidales, as the major hosts of the crAss-like phages associated with the human gut. Accordingly, four of the phenotypes mentioned above were reported to be associated with bacteria belonging to the order Bacteroidales, and in three of these cases (levels of CgA and beta-defensin-2 and IBS), the directionality of the association matched that of the crAss-like phage associations with the phenotype in question, whereas in one case (laxative usage), the directionality was opposite (Zhernakova et al., 2016). The notion that phage-phenotype associations may reflect the host-phenotype associations is further supported by our observation that, after the adjustment of the logistic regression model of the phage-phenotype relationship for the abundance of the four potential host genera (*Bacteroides*, *Prevotella*, *Porphyromonas*, and *Parabacteroides*), the identified phage-phenotype associations became only nominally significant.

The prevalence of the crAss-like phages was strongly associated with an IBD diagnosis: the assemblage, all five groups,

and 19 genus-level clusters of crAss-like phages were less prevalent in the IBD patient cohort compared with the population-based LLD cohort. In contrast, there was no statistically significant difference in prevalence of crAss-like phages between the LLD and 300OB cohorts at any taxonomic level. The depletion of crAss-like phages in the IBD cohort may be explained by the depletion of their hosts. A decrease in bacterial biomass and bacterial richness has consistently been described in patients with IBD, including the depletion of several members of the order Bacteroidales (Schirmer et al., 2019; Vieira-Silva et al., 2019). Although, interestingly, the majority of the associations between prevalence of crAss-like phages and IBD diagnosis remained significant after the adjustment of the logistic regression model for the abundance of the four potential host genera.

In conclusion, we have expanded the known diversity of crAss-like phages by a third at the species level, thus contributing to the ongoing discovery and cataloging of the human gut virome. Our study shows that identifying, classifying, and analyzing novel viral genomes can both offer new clues about the biology of viruses and further our understanding of the interactions between viruses and their surrounding ecosystem and the role of viruses in human health and disease.

Limitations of the study

Limitations of the study include the possibility of read misalignment and misassembly because of the employment of short-read sequencing data, as well as estimation of the relative, rather than absolute, abundance of microorganisms and viruses. Importantly, the study relied on total metagenome sequencing data, and thus does not have the benefit of focusing specifically on the active virome fraction of the gut ecosystem that can be provided by sequencing nucleic acid extracted from virus-like particles (Garmaeva et al., 2019). Finally, selection of the vOTU detection threshold, which is defined as >10% reference genome length coverage by reads in our study, has an influence on the results, determining the selectivity and specificity of crAss-like phages detection.

STAR★METHODS

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

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - LifeLines-DEEP cohort
  - LifeLines-DEEP follow-up cohort
  - 300OB cohort
  - IBD cohort
- METHOD DETAILS
  - Sequencing reads quality control
  - Contigs assembly
  - crAss-like phage detection
  - Nucleotide sequence characterization
  - Species- and genus-level genome clustering
  - Open reading frame prediction
  - Proteome functional annotation
  - Genetic code determination
  - Phylogeny reconstruction
  - Assignment to crAss-like phage groups
  - Mapping reads to reference genomes
  - Genomes nucleotide identity analysis
  - Ecological measurements
  - Taxonomic profiling of microbial communities
  - Host prediction based on CRISPR spacers
  - Detection of bacterial structural variants
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Associations with microbial abundance
  - Associations with bacterial SVs
  - crAss-like phage relative abundance ~ SV + Age + Sex + BMI + Reads number
  - Associations with human phenotypes
  - Associations with IBD and obesity

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.celrep.2021.110204.

ACKNOWLEDGMENTS

We thank Natalya Yutin and Eugene V. Koonin for helpful discussion and advice. We would like to thank Lianmin Chen and Sergio Andreu-Sánchez for technical support; Kate Mc Intyre for editing the manuscript; the Center for Information Technology of the University of Groningen for their support and for providing access to the Peregrine high-performance computing cluster; and the volunteers of the LifeLines-DEEP, 300OB, and IBD cohorts for their participation. S.G. and R.A.A.A.R. hold scholarships from the Graduate School of Medical Sciences, University of Groningen and the Junior Scientific Masterclass, University of Groningen, respectively. D.W. is supported by China Scholarship Council (CSC201904910478). A.Z. is supported by the European Research Council (ERC) Starting Grant 715772, the Netherlands Organisation for Scientific Research (NWO) VIDI grant 016.178.056, and NWO Gravitation grant ExposomeNL 024.004.017. J.F. is supported by NWO Gravitation grant Netherlands Organ-on-Chip Initiative 024.003.001, ERC Consolidator grant 101001678, and NWO VICI grant VI.C.202.022. N.P.R., M.G.N., J.F., and A.Z. are supported by the Netherlands Heart Foundation CVON grant 2018-27. R.K.W. is supported by the Seerave Foundation and the Dutch Digestive Foundation (16-14). C.W. is supported by NWO Gravitation grant 024.003.001 and NWO Spinoza Prize SPI 92-266.

AUTHOR CONTRIBUTIONS

A.G. and A.Z. designed the study. N.P.R., M.G.N., C.W., R.K.W., J.F., and A.Z. provided the data. A.G., S.G., R.A.A.A.R., D.W., A.V.V., and A.K. performed data analysis. A.G., S.G., R.A.A.A.R., D.W., A.V.V., A.K., and A.Z. wrote the manuscript. All authors reviewed and edited the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: June 7, 2021
Revised: September 3, 2021
Accepted: December 13, 2021
Published: January 11, 2022

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.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. *215*, 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2.

Balduzzi, S., Rücker, G., and Schwarzer, G. (2019). How to perform a meta-analysis with R: a practical tutorial. Evid. Based Ment. Health. *22*, 153–160. https://doi.org/10.1136/ebmental-2019-300117.

Bartolomucci, A., Possenti, R., Mahata, S.K., Fischer-Colbrie, R., Loh, Y.P., and Salton, S.R. (2011). The extended granin family: structure, function, and biomedical implications. Endocr. Rev. *32*, 755–797. https://doi.org/10.1210/er.2010-0027.

Beghini, F., McIver, L.J., Blanco-Miguez, A., Dubois, L., Asnicar, F., Maharjan, S., Mailyan, A., Manghi, P., Scholz, M., Thomas, A.M., et al. (2021). Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. Elife *10*. https://doi.org/10.7554/eLife.65088.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B (Methodolog.) *57*, 289–300.

Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114–2120. https://doi.org/10.1093/bioinformatics/btu170.

Brister, J.R., Ako-Adjei, D., Bao, Y., and Blinkova, O. (2015). NCBI viral genomes resource. Nucleic Acids Res. *43*, D571–D577. https://doi.org/10.1093/nar/gku1207.

Camarillo-Guerrero, L.F., Almeida, A., Rangel-Pineros, G., Finn, R.D., and Lawley, T.D. (2021). Massive expansion of human gut bacteriophage diversity. Cell *184*, 1098–1109 e9. https://doi.org/10.1016/j.cell.2021.01.029.

Castillo, D., and Middelboe, M. (2016). Genomic diversity of bacteriophages infecting the fish pathogen Flavobacterium psychrophilum. FEMS Microbiol. Lett. *363*. https://doi.org/10.1093/femsle/fnw272.

Charif, D., and Lobry, J.R. (2007). SeqinR 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. In Structural Approaches to Sequence Evolution: Molecules, Networks, Populations, U. Bastolla, M. Porto, H.E. Roman, and M. Vendruscolo, eds. (Springer Berlin Heidelberg), pp. 207–232. https://doi.org/10.1007/978-3-540-35306-5_10.

Chen, L., Wang, D., Garmaeva, S., Kurilshikov, A., Vich Vila, A., Gacesa, R., Sinha, T., Lifelines Cohort, S., Segal, E., Weersma, R.K., et al. (2021). The long-term genetic stability and individual specificity of the human gut microbiome. Cell. https://doi.org/10.1016/j.cell.2021.03.024.

Clooney, A.G., Sutton, T.D.S., Shkoporov, A.N., Holohan, R.K., Daly, K.M., O'Regan, O., Ryan, F.J., Draper, L.A., Plevy, S.E., Ross, R.P., and Hill, C. (2019). Whole-virome analysis sheds light on viral dark matter in inflammatory bowel disease. Cell Host Microbe. *26*, 764–778.e5. https://doi.org/10.1016/j.chom.2019.10.009.

Drobysheva, A.V., Panafidina, S.A., Kolesnik, M.V., Klimuk, E.I., Minakhin, L., Yakunina, M.V., Borukhov, S., Nilsson, E., Holmfeldt, K., Yutin, N., et al. (2021). Structure and function of virion RNA polymerase of a crAss-like phage. Nature *589*, 306–309. https://doi.org/10.1038/s41586-020-2921-5.

Dutilh, B.E., Cassman, N., McNair, K., Sanchez, S.E., Silva, G.G., 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*, 4498. https://doi.org/10.1038/ncomms5498.

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. https://doi.org/10.1093/femsre/fuv048.

Edwards, R.A., Vega, A.A., Norman, H.M., Ohaeri, M., Levi, K., Dinsdale, E.A., Cinek, O., Aziz, R.K., McNair, K., Barr, J.J., et al. (2019). Global phylogeography and ancient evolution of the widespread human gut virus crAssphage. Nat. Microbiol. *4*, 1727–1736. https://doi.org/10.1038/s41564-019-0494-6.

Ewels, P., Magnusson, M., Lundin, S., and Kaller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics *32*, 3047–3048. https://doi.org/10.1093/bioinformatics/btw354.

Fitzgerald, C.B., Shkoporov, A.N., Upadrasta, A., Khokhlova, E.V., Ross, R.P., and Hill, C. (2021). Probing the "dark matter" of the human gut phageome: culture assisted metagenomics enables rapid discovery and host-linking for novel bacteriophages. Front. Cel. Infect. Microbiol. *11*. https://doi.org/10.3389/fcimb.2021.616918.

Franzosa, E.A., Sirota-Madi, A., Avila-Pacheco, J., Fornelos, N., Haiser, H.J., Reinker, S., Vatanen, T., Hall, A.B., Mallick, H., McIver, L.J., et al. (2019). Gut microbiome structure and metabolic activity in inflammatory bowel disease. Nat. Microbiol. *4*, 293–305. https://doi.org/10.1038/s41564-018-0306-4.

Garmaeva, S., Sinha, T., Kurilshikov, A., Fu, J., Wijmenga, C., and Zhernakova, A. (2019). Studying the gut virome in the metagenomic era: challenges and perspectives. BMC Biol. *17*, 84. https://doi.org/10.1186/s12915-019-0704-y.

Grant, B.J., Rodrigues, A.P., ElSawy, K.M., McCammon, J.A., and Caves, L.S. (2006). Bio3d: an R package for the comparative analysis of protein structures. Bioinformatics *22*, 2695–2696. https://doi.org/10.1093/bioinformatics/btl461.

Grigoriev, A. (1998). Analyzing genomes with cumulative skew diagrams. Nucleic Acids Res. *26*, 2286–2290. https://doi.org/10.1093/nar/26.10.2286.

Guerin, E., Shkoporov, A., Stockdale, S.R., Clooney, A.G., Ryan, F.J., Sutton, T.D.S., Draper, L.A., Gonzalez-Tortuero, E., Ross, R.P., and Hill, C. (2018). Biology and taxonomy of crAss-like bacteriophages, the most abundant virus in the human gut. Cell Host Microbe. *24*, 653–664 e6. https://doi.org/10.1016/j.chom.2018.10.002.

Guerin, E., Shkoporov, A.N., Stockdale, S.R., Comas, J.C., Khokhlova, E.V., Clooney, A.G., Daly, K.M., Draper, L.A., Stephens, N., Scholz, D., et al. (2021). Isolation and characterisation of ΦcrAss002, a crAss-like phage from the human gut that infects Bacteroides xylanisolvens. Microbiome *9*, 89. https://doi.org/10.1186/s40168-021-01036-7.

Hoang, D.T., Chernomor, O., von Haeseler, A., Minh, B.Q., and Vinh, L.S. (2018). UFBoot2: improving the ultrafast bootstrap approximation. Mol. Biol. Evol. *35*, 518–522. https://doi.org/10.1093/molbev/msx281.

Holmfeldt, K., Solonenko, N., Shah, M., Corrier, K., Riemann, L., Verberkmoes, N.C., and Sullivan, M.B. (2013). Twelve previously unknown phage genera are ubiquitous in global oceans. Proc. Natl. Acad. Sci. U S A. *110*, 12798–12803. https://doi.org/10.1073/pnas.1305956110.

Hryckowian, A.J., Merrill, B.D., Porter, N.T., Van Treuren, W., Nelson, E.J., Garlena, R.A., Russell, D.A., Martens, E.C., and Sonnenburg, J.L. (2020). Bacteroides thetaiotaomicron-infecting bacteriophage isolates inform sequence-based host range predictions. Cell Host Microbe. *28*, 371–379 e5. https://doi.org/10.1016/j.chom.2020.06.011.

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. https://doi.org/10.1186/1471-2105-11-119.

Imhann, F., Van der Velde, K.J., Barbieri, R., Alberts, R., Voskuil, M.D., Vich Vila, A., Collij, V., Spekhorst, L.M., Van der Sloot, K.W.J., Peters, V., et al. (2019). The 1000IBD project: multi-omics data of 1000 inflammatory bowel disease patients; data release 1. BMC Gastroenterol. *19*, 5. https://doi.org/10.1186/s12876-018-0917-5.

Katoh, K., and Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. *30*, 772–780. https://doi.org/10.1093/molbev/mst010.

Koonin, E.V., Dolja, V.V., Krupovic, M., Varsani, A., Wolf, Y.I., Yutin, N., Zerbini, F.M., and Kuhn, J.H. (2020). Global organization and proposed megataxonomy of the virus world. Microbiol. Mol. Biol. Rev. *84*. https://doi.org/10.1128/MMBR.00061-19.

Koonin, E.V., and Yutin, N. (2020). The crAss-like phage group: how metagenomics reshaped the human virome. Trends Microbiol. *28*, 349–359. https://doi.org/10.1016/j.tim.2020.01.010.

Kurilshikov, A., van den Munckhof, I.C.L., Chen, L., Bonder, M.J., Schraa, K., Rutten, J.H.W., Riksen, N.P., de Graaf, J., Oosting, M., Sanna, S., et al. (2019).

Gut microbial associations to plasma metabolites linked to cardiovascular phenotypes and risk. Circ. Res. *124*, 1808–1820. https://doi.org/10.1161/CIRCRESAHA.118.314642.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359. https://doi.org/10.1038/nmeth.1923.

Lawrence, M., Huber, W., Pages, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M.T., and Carey, V.J. (2013). Software for computing and annotating genomic ranges. PLoS Comput. Biol. *9*, e1003118. https://doi.org/10.1371/journal.pcbi.1003118.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Genome Project Data Processing, S. (2009). The sequence alignment/map format and SAMtools. Bioinformatics *25*, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352.

Lobry, J.R. (1996). Asymmetric substitution patterns in the two DNA strands of bacteria. Mol. Biol. Evol. *13*, 660–665. https://doi.org/10.1093/oxfordjournals.molbev.a025626.

Lole, K.S., Bollinger, R.C., Paranjape, R.S., Gadkari, D., Kulkarni, S.S., Novak, N.G., Ingersoll, R., Sheppard, H.W., and Ray, S.C. (1999). Full-length human immunodeficiency virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination. J. Virol. *73*, 152–160. https://doi.org/10.1128/JVI.73.1.152-160.1999.

Marbouty, M., Thierry, A., Millot, G.A., and Koszul, R. (2021). MetaHiC phage-bacteria infection network reveals active cycling phages of the healthy human gut. Elife *10*. https://doi.org/10.7554/eLife.60608.

Marcais, G., Delcher, A.L., Phillippy, A.M., Coston, R., Salzberg, S.L., and Zimin, A. (2018). MUMmer4: a fast and versatile genome alignment system. Plos Comput. Biol. *14*, e1005944. https://doi.org/10.1371/journal.pcbi.1005944.

Mende, D.R., Letunic, I., Huerta-Cepas, J., Li, S.S., Forslund, K., Sunagawa, S., and Bork, P. (2017). proGenomes: a resource for consistent functional and taxonomic annotations of prokaryotic genomes. Nucleic Acids Res. *45*, D529–D534. https://doi.org/10.1093/nar/gkw989.

Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., von Haeseler, A., and Lanfear, R. (2020). IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Mol. Biol. Evol. *37*, 1530–1534. https://doi.org/10.1093/molbev/msaa015.

Nayfach, S., Camargo, A.P., Schulz, F., Eloe-Fadrosh, E., Roux, S., and Kyrpides, N.C. (2020). CheckV assesses the quality and completeness of metagenome-assembled viral genomes. Nat. Biotechnol. https://doi.org/10.1038/s41587-020-00774-7.

Nurk, S., Meleshko, D., Korobeynikov, A., and Pevzner, P.A. (2017). metaSPAdes: a new versatile metagenomic assembler. Genome Res. *27*, 824–834. https://doi.org/10.1101/gr.213959.116.

Oude Munnink, B.B., Canuti, M., Deijs, M., de Vries, M., Jebbink, M.F., Rebers, S., Molenkamp, R., van Hemert, F.J., Chung, K., Cotten, M., et al. (2014). Unexplained diarrhoea in HIV-1 infected individuals. BMC Infect. Dis. *14*, 22. https://doi.org/10.1186/1471-2334-14-22.

Paradis, E., and Schliep, K. (2019). Ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics *35*, 526–528. https://doi.org/10.1093/bioinformatics/bty633.

Porter, N.T., Hryckowian, A.J., Merrill, B.D., Fuentes, J.J., Gardner, J.O., Glowacki, R.W.P., Singh, S., Crawford, R.D., Snitkin, E.S., Sonnenburg, J.L., and Martens, E.C. (2020). Phase-variable capsular polysaccharides and lipoproteins modify bacteriophage susceptibility in Bacteroides thetaiotaomicron. Nat. Microbiol. *5*, 1170–1181. https://doi.org/10.1038/s41564-020-0746-5.

Pourcel, C., Touchon, M., Villeriot, N., Vernadet, J.P., Couvin, D., Toffano-Nioche, C., and Vergnaud, G. (2020). CRISPRCasdb a successor of CRISPRdb containing CRISPR arrays and cas genes from complete genome sequences, and tools to download and query lists of repeats and spacers. Nucleic Acids Res. *48*, D535–D544. https://doi.org/10.1093/nar/gkz915.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842. https://doi.org/10.1093/bioinformatics/btq033.

R Core Team (2020). R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing).

Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS: the European molecular biology open software suite. Trends Genet. *16*, 276–277. https://doi.org/10.1016/s0168-9525(00)02024-2.

Roux, S., Adriaenssens, E.M., Dutilh, B.E., Koonin, E.V., Kropinski, A.M., Krupovic, M., Kuhn, J.H., Lavigne, R., Brister, J.R., Varsani, A., et al. (2019). Minimum information about an uncultivated virus genome (MIUViG). Nat. Biotechnol. *37*, 29–37. https://doi.org/10.1038/nbt.4306.

Roux, S., Emerson, J.B., Eloe-Fadrosh, E.A., and Sullivan, M.B. (2017). Benchmarking viromics: an in silico evaluation of metagenome-enabled estimates of viral community composition and diversity. Peer J. *5*, e3817. https://doi.org/10.7717/peerj.3817.

Roux, S., Enault, F., Hurwitz, B.L., and Sullivan, M.B. (2015). VirSorter: mining viral signal from microbial genomic data. Peer J. *3*, e985. https://doi.org/10.7717/peerj.985.

Roux, S., Paez-Espino, D., Chen, I.A., Palaniappan, K., Ratner, A., Chu, K., Reddy, T.B.K., Nayfach, S., Schulz, F., Call, L., et al. (2021). IMG/VR v3: an integrated ecological and evolutionary framework for interrogating genomes of uncultivated viruses. Nucleic Acids Res. *49*, D764–D775. https://doi.org/10.1093/nar/gkaa946.

Sayers, E.W., Cavanaugh, M., Clark, K., Pruitt, K.D., Schoch, C.L., Sherry, S.T., and Karsch-Mizrachi, I. (2021). Genbank. Nucleic Acids Res. *49*, D92–D96. https://doi.org/10.1093/nar/gkaa1023.

Schirmer, M., Garner, A., Vlamakis, H., and Xavier, R.J. (2019). Microbial genes and pathways in inflammatory bowel disease. Nat. Rev. Microbiol. *17*, 497–511. https://doi.org/10.1038/s41579-019-0213-6.

Schliep, K.P. (2011). phangorn: phylogenetic analysis in R. Bioinformatics *27*, 592–593. https://doi.org/10.1093/bioinformatics/btq706.

Schneider, V.A., Graves-Lindsay, T., Howe, K., Bouk, N., Chen, H.C., Kitts, P.A., Murphy, T.D., Pruitt, K.D., Thibaud-Nissen, F., Albracht, D., et al. (2017). Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. Genome Res. *27*, 849–864. https://doi.org/10.1101/gr.213611.116.

Schoch, C.L., Ciufo, S., Domrachev, M., Hotton, C.L., Kannan, S., Khovanskaya, R., Leipe, D., McVeigh, R., O'Neill, K., Robbertse, B., et al. (2020). NCBI Taxonomy: a comprehensive update on curation, resources and tools. Database (Oxford) *2020*. https://doi.org/10.1093/database/baaa062.

Shkoporov, A.N., Clooney, A.G., Sutton, T.D.S., Ryan, F.J., Daly, K.M., Nolan, J.A., McDonnell, S.A., Khokhlova, E.V., Draper, L.A., Forde, A., et al. (2019). The human gut virome is highly diverse, stable, and individual specific. Cell Host Microbe. *26*, 527–541.e5. https://doi.org/10.1016/j.chom.2019.09.009.

Shkoporov, A.N., Khokhlova, E.V., Fitzgerald, C.B., Stockdale, S.R., Draper, L.A., Ross, R.P., and Hill, C. (2018). PhiCrAss001 represents the most abundant bacteriophage family in the human gut and infects Bacteroides intestinalis. Nat. Commun. *9*, 4781. https://doi.org/10.1038/s41467-018-07225-7.

Shkoporov, A.N., Khokhlova, E.V., Stephens, N., Hueston, C., Seymour, S., Hryckowian, A.J., Scholz, D., Ross, R.P., and Hill, C. (2021). Long-term persistence of crAss-like phage crAss001 is associated with phase variation in Bacteroides intestinalis. BMC Biol. *19*, 163. https://doi.org/10.1186/s12915-021-01084-3.

Shmakov, S.A., Sitnik, V., Makarova, K.S., Wolf, Y.I., Severinov, K.V., and Koonin, E.V. (2017). The CRISPR spacer space is dominated by sequences from species-specific mobilomes. mBio. *8*. https://doi.org/10.1128/mBio.01397-17.

Siranosian, B.A., Tamburini, F.B., Sherlock, G., and Bhatt, A.S. (2020). Acquisition, transmission and strain diversity of human gut-colonizing crAss-like phages. Nat. Commun. *11*, 280. https://doi.org/10.1038/s41467-019-14103-3.

Smura, T., Blomqvist, S., Vuorinen, T., Ivanova, O., Samoilovich, E., Al-Hello, H., Savolainen-Kopra, C., Hovi, T., and Roivainen, M. (2014). Recombination in the evolution of enterovirus C species sub-group that contains types

CVA-21, CVA-24, EV-C95, EV-C96 and EV-C99. PLoS One *9*, e94579. https://doi.org/10.1371/journal.pone.0094579.

Stajich, J.E., Block, D., Boulez, K., Brenner, S.E., Chervitz, S.A., Dagdigian, C., Fuellen, G., Gilbert, J.G., Korf, I., Lapp, H., et al. (2002). The Bioperl toolkit: perl modules for the life sciences. Genome Res. *12*, 1611–1618. https://doi.org/10.1101/gr.361602.

Ter Horst, R., van den Munckhof, I.C.L., Schraa, K., Aguirre-Gamboa, R., Jaeger, M., Smeekens, S.P., Brand, T., Lemmers, H., Dijkstra, H., Galesloot, T.E., et al. (2020). Sex-specific regulation of inflammation and metabolic syndrome in obesity. Arterioscler Thromb. Vasc. Biol. *40*, 1787–1800. https://doi.org/10.1161/ATVBAHA.120.314508.

Tigchelaar, E.F., Zhernakova, A., Dekens, J.A., Hermes, G., Baranska, A., Mujagic, Z., Swertz, M.A., Munoz, A.M., Deelen, P., Cenit, M.C., et al. (2015). Cohort profile: LifeLines DEEP, a prospective, general population cohort study in the northern Netherlands: study design and baseline characteristics. BMJ Open *5*, e006772. https://doi.org/10.1136/bmjopen-2014-006772.

Tisza, M.J., and Buck, C.B. (2021). A catalog of tens of thousands of viruses from human metagenomes reveals hidden associations with chronic diseases. Proc. Natl. Acad. Sci. U S A. *118*. https://doi.org/10.1073/pnas.2023202118.

Vich Vila, A., Imhann, F., Collij, V., Jankipersadsing, S.A., Gurry, T., Mujagic, Z., Kurilshikov, A., Bonder, M.J., Jiang, X., Tigchelaar, E.F., et al. (2018). Gut microbiota composition and functional changes in inflammatory bowel disease and irritable bowel syndrome. Sci. Transl. Med. *10*. https://doi.org/10.1126/scitranslmed.aap8914.

Vieira-Silva, S., Sabino, J., Valles-Colomer, M., Falony, G., Kathagen, G., Caenepeel, C., Cleynen, I., van der Merwe, S., Vermeire, S., and Raes, J. (2019). Quantitative microbiome profiling disentangles inflammation- and bile duct obstruction-associated microbiota alterations across PSC/IBD diagnoses. Nat. Microbiol. *4*, 1826–1831. https://doi.org/10.1038/s41564-019-0483-9.

Wang, D., Doestzada, M., Chen, L., Andreu-Sánchez, S., van den Munckhof, I.C.L., Augustijn, H.E., Koehorst, M., Ruiz-Moreno, A.J., Bloks, V.W., Riksen, N.P., et al. (2021). Characterization of gut microbial structural variations as determinants of human bile acid metabolism. Cell Host Microbe. https://doi.org/10.1016/j.chom.2021.11.003.

Waterhouse, A.M., Procter, J.B., Martin, D.M., Clamp, M., and Barton, G.J. (2009). Jalview Version 2–a multiple sequence alignment editor and analysis workbench. Bioinformatics *25*, 1189–1191. https://doi.org/10.1093/bioinformatics/btp033.

Wingett, S.W., and Andrews, S. (2018). FastQ Screen: a tool for multi-genome mapping and quality control. F1000Res. *7*, 1338. https://doi.org/10.12688/f1000research.15931.2.

Yutin, N., Benler, S., Shmakov, S.A., Wolf, Y.I., Tolstoy, I., Rayko, M., Antipov, D., Pevzner, P.A., and Koonin, E.V. (2021). Analysis of metagenome-assembled viral genomes from the human gut reveals diverse putative CrAss-like phages with unique genomic features. Nat. Commun. *12*, 1044. https://doi.org/10.1038/s41467-021-21350-w.

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. https://doi.org/10.1038/s41564-017-0053-y.

Zeevi, D., Korem, T., Godneva, A., Bar, N., Kurilshikov, A., Lotan-Pompan, M., Weinberger, A., Fu, J., Wijmenga, C., Zhernakova, A., and Segal, E. (2019). Structural variation in the gut microbiome associates with host health. Nature *568*, 43–48. https://doi.org/10.1038/s41586-019-1065-y.

Zhernakova, A., Kurilshikov, A., Bonder, M.J., Tigchelaar, E.F., Schirmer, M., Vatanen, T., Mujagic, Z., Vila, A.V., Falony, G., Vieira-Silva, S., et al. (2016). Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. Science *352*, 565–569. https://doi.org/10.1126/science.aad3369.

CellPress
OPEN ACCESS

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
LLD data	Tigchelaar et al. (2015); Zhernakova et al. (2016)	https://ega-archive.org/datasets/EGAD00001001991, https://www.lifelines.nl/researcher/
LLD follow-up data	Chen et al. (2021)	https://ega-archive.org/datasets/EGAD00001006959, https://www.lifelines.nl/researcher/
300OB data	Kurilshikov et al. (2019)	https://ega-archive.org/studies/EGAS00001003508
IBD data	Imhann et al. (2019); Vich Vila et al. (2018)	https://ega-archive.org/studies/EGAS00001002702
Human reference genome GRCh38.p13	Schneider et al. (2017)	https://www.ncbi.nlm.nih.gov/grc/human
Viral RefSeq 202	Brister et al. (2015)	https://www.ncbi.nlm.nih.gov/genome/viruses/
A database of 249 crAss-like phage genomes	Guerin et al. (2018)	N/A
A database of 673 crAss-like phage genomes; MSAs of crAss-like phage domains	Yutin et al. (2021)	N/A
A database of 1556 crAss-like genomes and genome fragments	This paper	https://doi.org/10.6084/m9.figshare.16708696
A database of CRISPR spacers	Shmakov et al. (2017)	N/A
CRISPR-Cas++ spacers database (21.01.2021)	Pourcel et al. (2020)	https://crisprcas.i2bc.paris-saclay.fr/
proGenomes database	Mende et al. (2017)	http://progenomes1.embl.de/
Software and algorithms		
KneadData 0.7.4	https://huttenhower.sph.harvard.edu/kneaddata/	https://huttenhower.sph.harvard.edu/kneaddata/
Trimmomatic 0.33	Bolger et al. (2014)	https://github.com/usadellab/Trimmomatic
FastQC 0.11.7	Wingett and Andrews (2018)	https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
MultiQC 1.9	Ewels et al. (2016)	https://multiqc.info/
MetaSPAdes 3.14.1	Nurk et al. (2017)	https://cab.spbu.ru/software/spades/
BBMap 37.93	https://sourceforge.net/projects/bbmap/	https://sourceforge.net/projects/bbmap/
EMBOSS 6.6.0	Rice et al. (2000)	http://emboss.sourceforge.net/
Prodigal 2.6.3	Hyatt et al. (2010)	https://github.com/hyattpd/Prodigal
HMMER 3.3	http://hmmer.org/	http://hmmer.org/
rhmmer 0.1.0 R package	https://CRAN.R-project.org/package=rhmmer	https://CRAN.R-project.org/package=rhmmer
Cluster_genomes_5.1.pl script	Roux et al. (2017)	https://github.com/simroux/ClusterGenomes
MUMmer 4.0.0beta2	Marcais et al. (2018)	https://github.com/mummer4/mummer
CheckV 0.7.0	Nayfach et al. (2020)	https://bitbucket.org/berkeleylab/CheckV
IRanges 2.22.2 R package	Lawrence et al. (2013)	https://bioconductor.org/packages/IRanges/
MAFFT 7.453	Katoh and Standley (2013)	https://mafft.cbrc.jp/alignment/software/

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
seqinr 3.6-1 R package	Charif and Lobry (2007)	https://CRAN.R-project.org/package=seqinr
Jalview 2.11.1.3	Waterhouse et al. (2009)	https://www.jalview.org/
IQ-TREE 2.0.3	Minh et al. (2020)	http://www.iqtree.org/
phangorn 2.5.5 R package	Schliep (2011)	https://CRAN.R-project.org/package=phangorn
ape 5.4-1 R package	Paradis and Schliep (2019)	https://CRAN.R-project.org/package=ape
Bowtie2 2.4.2	Langmead and Salzberg (2012)	http://bowtie-bio.sourceforge.net/bowtie2/
BEDTools 2.29.2	Quinlan and Hall (2010)	https://bedtools.readthedocs.io/
SAMtools 1.10	Li et al. (2009)	https://www.htslib.org/
Seqtk 1.3	https://github.com/lh3/seqtk	https://github.com/lh3/seqtk
Bio3D 2.4-1 R package	Grant et al. (2006)	http://thegrantlab.org/bio3d/
BLASTN 2.10.1+	Altschul et al. (1990)	https://blast.ncbi.nlm.nih.gov/Blast.cgi
vegan 2.5-6 R package	https://CRAN.R-project.org/package=vegan	https://CRAN.R-project.org/package=vegan
vioplot 0.3.6 R package	https://github.com/TomKellyGenetics/vioplot	https://github.com/TomKellyGenetics/vioplot
MetaPhlAn 3.0.7	Beghini et al. (2021)	https://huttenhower.sph.harvard.edu/metaphlan/
BioPerl 1.6.924	Stajich et al. (2002)	https://bioperl.org/
SGV-Finder	Zeevi et al. (2019)	https://github.com/segalab/SGVFinder
meta 4.18-0 R package	Balduzzi et al. (2019)	https://CRAN.R-project.org/package=meta
foreach 1.5.1 R package	https://CRAN.R-project.org/package=foreach	https://CRAN.R-project.org/package=foreach
Custom scripts	This paper	https://doi.org/10.6084/m9.figshare.17138852

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Prof. A. Zhernakova (a.zhernakova@umcg.nl; sashazhernakova@gmail.com).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- Previously published data used in this study are listed in the key resources table. A FASTA file with the 1,556 crAss-like genomes and genome fragments identified in this study has been deposited at the Figshare repository and is publicly available as of the date of publication. The DOI is listed in the key resources table.
- All original code has been deposited at the Figshare repository and is publicly available as of the date of publication. The DOI is listed in the key resources table.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Data used in the project included gut metagenome sequencing data and human phenotype data from four Dutch human cohorts. For all cohorts, DNA isolation and metagenomic sequencing was done in the same laboratory and used the same protocols and pipelines, and was described in earlier publications (Chen et al., 2021; Imhann et al., 2019; Kurilshikov et al., 2019; Vich Vila et al., 2018; Zhernakova et al., 2016).

LifeLines-DEEP cohort

The LifeLines-DEEP (LLD) cohort includes individuals from the general population (Tigchelaar et al., 2015; Zhernakova et al., 2016). Data for 1135 LLD cohort participants, whose raw gut metagenome sequencing data included more than 15 million reads, were used in the study. The age of the participants ranged from 18 to 81 years. There were 41.8% male and 58.2% female participants. The study has been approved by the institutional ethics review board of the University Medical Center Groningen (ref. M12.113965).

LifeLines-DEEP follow-up cohort

LLD follow-up cohort includes data about 338 LLD cohort participants collected 4 years after the initial LLD cohort data collection (Chen et al., 2021). The age of the participants ranged from 21 to 84 years. There were 44.4% male and 55.6% female participants. The study has been approved by the institutional ethics review board of the University Medical Center Groningen (ref. M12.113965).

300OB cohort

The 300OB cohort includes individuals with BMI >27 kg/m^2 (Kurilshikov et al., 2019; Ter Horst et al., 2020). Data for 298 300OB cohort participants were used in this study. The age of the participants ranged from 54 to 81 years. There were 55.7% male and 44.3% female participants. All participants received detailed printed and oral information and subsequently gave written informed consent. The study was approved by the Ethical Committee of the Radboud University (nr. 46846.091.13). Experiments were conducted according to the principles expressed in the Declaration of Helsinki.

IBD cohort

The IBD cohort includes individuals with inflammatory bowel disease (Imhann et al., 2019; Vich Vila et al., 2018). A total of 520 gut metagenome samples from the IBD cohort containing $\geq$10 million sequencing reads were screened for the presence of crAss-like phages. To conduct association analyses, we excluded data obtained from 62 individuals with stoma and ileoanal pouches, one duplicated sample and two samples without metadata from the statistical analysis, bringing the number of individuals to 455. The age of the 455 individuals ranged from 18 to 82 years. There were 40.7% male and 59.3% female individuals, respectively. All participants signed an informed consent form prior to sample collection. Institutional ethics review board (IRB) approval is available for the 1000IBD (IRB-number 2008.338) cohort.

METHOD DETAILS

Sequencing reads quality control

Sequencing reads from each sample were independently processed using the KneadData 0.7.4 pipeline (https://huttenhower.sph.harvard.edu/kneaddata/). Specifically, reads mapping to the human reference genome GRCh38.p13 (Schneider et al., 2017) were filtered out, and adapter and quality trimming of the reads was performed using Trimmomatic 0.33 (Bolger et al., 2014) with the parameters ''ILLUMINACLIP:NexteraPE-PE.fa:2:30:10:1:true LEADING:20 TRAILING:20 SLIDINGWINDOW:4:20 MINLEN:50''. Read quality was assessed using FastQC 0.11.7 (Wingett and Andrews, 2018) and MultiQC 1.9 (Ewels et al., 2016).

Contigs assembly

Contigs were assembled for each sample independently using metaSPAdes 3.14.1 (Nurk et al., 2017). In some cases (14, 1, 6 and 24 samples from LLD, LLD follow-up, 300OB and IBD cohorts, respectively), metaSPAdes did not progress to finish its read error correction step. For these samples, read error correction was conducted using BBMap 37.93 program tadpole.sh with ''mode=correct ecc=t prefilter=2'' parameters (https://sourceforge.net/projects/bbmap/), followed by metaSPAdes assembly with the ''–only-assembler'' parameter.

crAss-like phage detection

The datasets screened for the presence of crAss-like phages included sets of contigs $\geq$10 kb from individual samples belonging to the LLD, LLD follow-up, 300OB and IBD cohorts, as well as genomes from four control datasets: Viral RefSeq 202 (Brister et al., 2015), a database of 249 crAss-like phages described in (Guerin et al., 2018), a database of 673 crAss-like phages described in (Yutin et al., 2021) and a database of 146 nucleotide sequences $\geq$10 kb recognized as ''crAss-like'' by the NCBI Taxonomy resource on 24.12.2020 (Schoch et al., 2020). Genomes and contigs were translated in six frames by the EMBOSS 6.6.0 program transeq with a ''-clean'' parameter (Rice et al., 2000), and their proteomes were predicted by Prodigal 2.6.3 in the ''meta'' mode (Hyatt et al., 2010). The HMMER 3.3 program hmmsearch with an E-value threshold of 0.001 (http://hmmer.org/) was used to compare proteomes, and, separately, translations in six frames to the following three highly conserved structural and genome-packaging proteins of crAss-like phages: terminase large subunit (three profiles from (Yutin et al., 2021)), portal protein (three profiles from (Yutin et al., 2021)) and major capsid protein (nine profiles from (Yutin et al., 2021; Yutin et al., 2018)). The output was processed with the help of R package *rhmmer* 0.1.0 (https://CRAN.R-project.org/package=rhmmer). A genome or a contig was considered as crAss-like if either its predicted proteome or its translations in six frames were hit by profiles representing each of the three proteins.

Nucleotide sequence characterization

To identify potentially complete genomes, direct overlaps of ≥10 nt at the ends of the genome sequences were detected using the approach from (Roux et al., 2015). The nucleotide content was estimated as the percent of each of the four nucleotides in a window sliding along the genome. GC and AT skew were calculated as (G − C)/(G + C) and (A − T)/(A + T) in a window sliding along the genome, respectively (Lobry, 1996). A 1001 nt window with a 200 nt step was used in all calculations; the most 3′-terminal window was extended to include up to 200 3′-terminal nucleotides. Cumulative GC (AT) skew corresponding to a window was calculated as a sum of GC (AT) skew values in this and all preceding windows (Grigoriev, 1998).

Species- and genus-level genome clustering

Species-level virus operational taxonomic units (vOTUs) were defined under the 95% ANI over the 85% AF threshold (Roux et al., 2019) using the Cluster_genomes_5.1.pl script (https://github.com/simroux/ClusterGenomes) relying on MUMmer 4.0.0beta2 (Marcais et al., 2018; Roux et al., 2017). If available, a tentatively complete genome belonging to a vOTU was selected to represent it in all subsequent analyses. Otherwise, the longest genome fragment belonging to the vOTU was used. Three potentially chimeric contigs (artifacts of sequencing and assembly or genuine biological entities such as proviruses integrated into bacterial genomes or recombinant viruses) representing vOTUs were recognized based on abrupt changes in nucleotide content and coverage by reads along the length of the contig and CheckV 0.7.0 contamination assessment (Nayfach et al., 2020). The non-crAss-like regions of these contigs were excluded from consideration (Table S2), followed by re-calculation of the species-level clusters. Representatives of the vOTUs were further clustered into genus-level clusters under the 50% AF threshold (Adriaenssens and Brister, 2017) by the Cluster_genomes_5.1.pl script.

Open reading frame prediction

Open reading frames (ORFs) of every crAss-like phage representing a vOTU were independently predicted under the standard bacterial genetic code (translation table 11, c11) and under the two alternative genetic codes employed by crAss-like phages: c11 genetic code with stop codon TAG reassignment for glutamine and c11 genetic code with stop codon TGA reassignment for tryptophan (Yutin et al., 2021). The prediction under the standard bacterial genetic code was made by Prodigal 2.6.3 (Hyatt et al., 2010). Predictions under the alternative genetic codes were made by a specifically designed modification of Prodigal published in (Yutin et al., 2021). The predictions were made for the individual genomes in the "meta" mode, unless the genome sequence length was ≥100 kb and the "single" mode was used.

Proteome functional annotation

Predicted proteins of crAss-like phages were functionally annotated by the hits to profiles representing the highly conserved proteins of crAss-like phages (Yutin et al., 2021). The HMMER 3.3 program hmmsearch with an E-value threshold of 0.001 and the "–domtblout" output parameter was used (http://hmmer.org/). Hit envelope coordinates were used to define domain position. When envelopes of hits identifying the same domain overlapped, their coordinates were combined with the help of R package *IRanges* 2.22.2 (Lawrence et al., 2013).

Genetic code determination

For each crAss-like phage representing a vOTU (Table S2), we considered three genome maps, each with ORFs predicted using a different genetic code as described above. The proteomes corresponding to each genome map were functionally annotated as described above. The genome maps (Data S1) were inspected to determine if a crAss-like phage may employ an alternative genetic code. If conserved genes of a crAss-like phage were split into multiple small ORFs under a standard bacterial genetic code, an alternative genetic code allowing for expression of the conserved genes with less interruption was assigned to the genome in question.

Phylogeny reconstruction

A multiple sequence alignment (MSA) of the TerL protein sequences was built based on the MSA from (Yutin et al., 2021). A subset of the TerL MSA from (Yutin et al., 2021) corresponding to the crAss-like phages representing the vOTUs in this study was supplemented with TerL sequences of crAss-like phages that represent vOTUs and are specific to this study using MAFFT 7.453 with the "–add" parameter (Katoh and Standley, 2013). Protein fragments (potentially split by introns (Yutin et al., 2021)) were joined, when necessary, with the help of R package *seqinr* 3.6-1 (Charif and Lobry, 2007), and the alignment was inspected with the help of Jalview 2.11.1.3 (Waterhouse et al., 2009). The MSA was used to reconstruct a phylogenetic tree using IQ-TREE 2.0.3 with 1000 replicates of ultrafast bootstrap (Hoang et al., 2018; Minh et al., 2020). The tree was midpoint-rooted using R package *phangorn* 2.5.5 and visualized using R package *ape* 5.4-1 (Paradis and Schliep, 2019; Schliep, 2011). An identical procedure was used to reconstruct a phylogenetic tree based on the portal protein sequences.

Assignment to crAss-like phage groups

Assignments of crAss-like phages to the five groups made in (Yutin et al., 2021) were extended to the vOTUs of this study. First, the most recent common ancestor (MRCA) of crAss-like phages that received an assignment to a certain group in (Yutin et al., 2021) was identified on the TerL protein phylogenetic tree. Next, we verified that all descendants of the MRCA were either assigned to the group

in question in (Yutin et al., 2021) or were not part of that study. Finally, all descendants of the MRCA were assigned to the group in question. The procedure was conducted using R packages *phangorn* 2.5.5 and *ape* 5.4-1 (Paradis and Schliep, 2019; Schliep, 2011).

Mapping reads to reference genomes

Sequencing reads of each individual sample, filtered and quality-trimmed as described above, were competitively mapped to a database of 378 crAss-like phage genomes representing vOTUs using Bowtie2 2.4.2 with a "–very-sensitive" parameter (Langmead and Salzberg, 2012). Breadth of genome coverage by reads was calculated using the BEDTools 2.29.2 command coverage (Quinlan and Hall, 2010). Depth of genome coverage by reads was calculated using the SAMtools 1.10 command depth (Li et al., 2009). Abundance of a vOTU in a sample was considered to be zero if the breadth of the representative genome coverage by reads did not exceed 10%, otherwise it was estimated as $(N \cdot 10^6)/(L \cdot S)$, where N is the number of reads mapped to a genome, L is the length of a genome and S is the number of sample reads after filtering and quality trimming. The abundance of genus- and higher-level taxonomic clusters in a sample was calculated as a sum of abundances of vOTUs belonging to the cluster. A vOTU was considered to be detected in a sample if >10% of the length of its representative sequence was covered by reads. A genus-level or higher-level taxonomic cluster was considered to be detected in a sample if at least one vOTU belonging to it was detected.

Genomes nucleotide identity analysis

All genomes belonging to a genus-level cluster that were tentatively complete or nearly complete and represented a vOTU were collected with the help of Seqtk 1.3 (https://github.com/lh3/seqtk), reverse-transcribed and permuted where necessary to make their termini match those of a selected genome, and aligned using MAFFT 7.453 (Katoh and Standley, 2013). To estimate nucleotide identity between the selected genome and any other genome, a pairwise alignment was derived from the MSA using R package *Bio3D* 2.4-1 (Grant et al., 2006), its columns containing gaps in the selected genome row were removed, and the identity was recorded using a 1001 nt window sliding along the selected genome with a 200 nt step. The most 3′-terminal window was extended to include up to 200 3′-terminal nucleotides. A fragment of the delta27 genome MSA corresponding to the 50656-72821 nt of the OLNE01000081 genome (RNAP and VP02740 ORFs predicted under the TAG stop codon reassignment for glutamine) was extracted and used to reconstruct a phylogenetic tree using IQ-TREE 2.0.3 with the "-B 1000" parameter (Hoang et al., 2018; Minh et al., 2020). The tree was subsequently midpoint-rooted using *phangorn* 2.5.5 (Schliep, 2011). The RNAP and VP02740 ORFs of OLNE01000081 were also used to query the set of 2616 crAss-like phage genomes and GenBank (14.02.2021, parameters "-db 'nt' -remote") using BLASTN 2.10.1+ (Altschul et al., 1990).

Ecological measurements

The Bray-Curtis dissimilarities between samples were calculated based on the genus-level data using the function vegdist() from the R package *vegan* 2.5-6. Genus-level clusters that were not detected in any sample and samples without crAss-like phages were excluded from consideration. Distributions of the Bray-Curtis dissimilarities were visualized using R package *vioplot* 0.3.6. Wilcoxon signed-rank tests comparing intra-individual and inter-individual dissimilarities were performed using the R function wilcox.test() with the "alternative = 'less', paired = FALSE" parameters (R Core Team, 2020). Significance of the Bray-Curtis dissimilarities difference was also confirmed (empirical P-value < 0.0001) in a permutational test with 10,000 iterations, designed to account for the highly unequal number of intra-individual and inter-interindividual dissimilarities (Chen et al., 2021).

Taxonomic profiling of microbial communities

The relative abundance of microbial taxa was estimated using MetaPhlAn 3.0.7 (Beghini et al., 2021).

Host prediction based on CRISPR spacers

The database of CRISPR spacers published in (Shmakov et al., 2017) and the CRISPR-Cas++ spacers database (21.01.2021) (Pourcel et al., 2020) were independently compared to the set of 2616 crAss-like phage genomes using BLASTN 2.10.1+ (Altschul et al., 1990) with the parameters "-task blastn-short -dust no -evalue 1 -max_target_seqs 1000000". A crAss-like genome was linked to a host if there was a spacer–protospacer match characterized by $\geq$95% identity over the length of the spacer, or multiple spacer–protospacer matches characterized by $\geq$80% identity over the length of each spacer (Roux et al., 2021). When multiple spacers of a host organism matched exactly the same region of a crAss-like phage genome, or if multiple regions of a crAss-like phage genome matched the same spacer of a host organism, a single spacer–protospacer match characterized by the highest bit-score was considered. Host taxonomy was retrieved from Genbank with the help of BioPerl 1.6.924 (Stajich et al., 2002).

Detection of bacterial structural variants

Prior to the detection of bacterial structural variants (SVs), human genome reads and low-quality reads were filtered from raw sequencing data using KneadData 0.7.4, Bowtie 2.3.4.3 and Trimmomatic 0.39 in a two-step procedure: (1) removal of human genome reads by aligning metagenomic reads to the human reference genome (GRCh37/hg19) and (2) removal of low-quality reads and adaptor sequences using Trimmomatic (parameters: SLIDINGWINDOW:4:20 MINLEN:50). Bacterial SVs were detected using a previously published tool, SGV-Finder (Zeevi et al., 2019), which is composed of two main steps. The first step aligns the cleaned metagenomic reads to reference bacterial genomes and solves ambiguously aligned reads using the iterative coverage-based

read assignment algorithm. The second step splits the reference genomes into genomic bins and examines the coverage of these bins to detect highly variable genomic segments for species with enough coverage and generate variable SV and deletion SV profiles. We used the reference database provided by SGV-Finder, which is based on the proGenomes database (http://progenomes1.embl.de/) (Mende et al., 2017). In total, we detected 7,999 deletion SVs and 3,559 variable SVs from 55 bacteria using default parameters. All bacterial species with SV calling were present in at least 5% of total samples.

QUANTIFICATION AND STATISTICAL ANALYSIS

Associations with microbial abundance

Correlation between the relative abundance of microbial taxa (from kingdoms to species) and the abundance of crAss-like phages (assemblage, groups, genus-level clusters) was assessed using the R function cor.test() with the "method = 'spearman'" parameter (R Core Team, 2020) for each cohort independently. The IBD cohort was represented by the selected 455 samples (see above). Only the taxa present in >10 samples in a given cohort were considered. Meta-analysis of the results obtained for the LLD, 300OB and IBD cohorts was conducted using the R package *meta* 4.18-0 (Balduzzi et al., 2019), function metacor() employing Fisher's Z-transformation of correlations and the Sidik-Jonkman between-study variance estimation method (parameters "sm = 'ZCOR', method.tau = 'SJ'"). Multiple testing correction was conducted by the R function p.adjust() using the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995).

Associations with bacterial SVs

Prior to the analysis, all continuous variables were standardized to follow a standard normal distribution using empirical normal quantile transformation. Associations between SVs and relative abundances of crAss-like phages at genus and higher taxonomic levels (30 taxonomic clusters present in >5% LLD samples) were assessed in the LLD, 300OB and IBD cohorts separately using linear models with the following formula:

crAss-like phage relative abundance ~ SV + Age + Sex + BMI + Reads number

The association results from different cohorts were then integrated via meta-analysis (random-effect model) and heterogeneity analysis. To control the FDR, Benjamini-Hochberg P-value corrections were performed using the p.adjust() function in R. The association analysis and P-value correction were conducted for vSVs and dSVs separately. Replicable significant associations were defined using the following criteria: $FDR_{meta} < 0.05$, $P\text{-value}_{heterogeneity} > 0.05$, $P\text{-value}_{LLD} < 0.05$, $P\text{-value}_{300OB} < 0.05$ and $P\text{-value}_{IBD} < 0.05$.

Associations with human phenotypes

The analysis was applied to 1135 LLD cohort samples, 207 phenotypes (missing values imputed) and 30 taxonomic clusters of crAss-like phages present in >5% LLD samples. The associations were analyzed using five methods: (1) Spearman correlation with abundance of crAss-like taxonomic clusters, (2) Spearman correlation with presence-absence of crAss-like taxonomic clusters, (3) unadjusted logistic regression, (4) logistic regression adjusted for the age and sex of the cohort participants and (5) logistic regression adjusted for the age and sex of the cohort participants and for the log-transformed abundance of the bacterial genera *Bacteroides*, *Prevotella*, *Porphyromonas* and *Parabacteroides*. Spearman correlation was assessed using R function cor.test() with the "method = 'spearman'" parameter (R Core Team, 2020). Logistic regression was fitted using the R function glm() with the "family = 'binomial'" parameter. Multiple testing correction was conducted using the R function p.adjust() using the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995).

Associations with IBD and obesity

The analysis involved 455 IBD cohort samples from patients without a stoma (colostomy or ileostomy) or ileoanal pouch, 298 300OB cohort samples and 1135 LLD cohort samples; 30 taxonomic clusters of crAss-like phages present in >5% LLD samples were considered. Logistic regression was performed to compare the prevalence of the crAss-like phage taxonomic clusters between the following test groups: (1) LLD vs. IBD cohort, (2) LLD vs. 300OB cohort, (3) within IBD cohort: CD vs. UC, (4) within IBD cohort: exclusively colonic vs. ileal-inclusive disease location, (5) within 300OB cohort: absence vs. presence of metabolic syndrome. CrAss-like phage abundances were coded as 0 for absence (zero values) and 1 for presence (non-zero values). The logistic regression tests were performed using the R function glm() with the parameters "family = binomial(link = 'logit')". Age and sex were included as covariates in the regression model for all comparisons. For comparisons involving IBD cohort samples, we also included carbohydrates and meat consumption, frequency of coffee consumption, usage of laxatives, and CgA and HBD-2 levels as covariates; these variables were not available for the 300OB cohort. In addition, we conducted a version of the LLD vs. IBD cohort and the LLD vs. 300OB cohort comparisons with the logistic regression further adjusted for the log-transformed abundance of the bacterial genera *Bacteroides*, *Prevotella*, *Porphyromonas* and *Parabacteroides*. Benjamini-Hochberg correction for multiple testing was applied using the R function p.adjust() with the parameter "method = 'fdr'". A significance threshold of FDR < 0.05 was used.