A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes

Bas E. Dutilh1,2,3,4, Noriko Cassman3,†, Katelyn McNair2, Savannah E. Sanchez3, Genivaldo G.Z. Silva5, Lance Boling3, Jeremy J. Barr3, Daan R. Speth6, Victor Seguritan3, Ramy K. Aziz2,7, Ben Felts8, Elizabeth A. Dinsdale3,5, John L. Mokili3 & Robert A. Edwards2,4,5,9

Metagenomics, or sequencing of the genetic material from a complete microbial community, is a promising tool to discover novel microbes and viruses. Viral metagenomes typically contain many unknown sequences. Here we describe the discovery of a previously unidentified bacteriophage present in the majority of published human faecal metagenomes, which we refer to as crAssphage. Its ~97 kbp genome is six times more abundant in publicly available metagenomes than all other known phages together; it comprises up to 90% and 22% of all reads in virus-like particle (VLP)-derived metagenomes and total community metagenomes, respectively; and it totals 1.68% of all human faecal metagenomic sequencing reads in the public databases. The majority of crAssphage-encoded proteins match no known sequences in the database, which is why it was not detected before. Using a new co-occurrence profiling approach, we predict a Bacteroides host for this phage, consistent with Bacteroides-related protein homologues and a unique carbohydrate-binding domain encoded in the phage genome.

1 Centre for Molecular and Biomolecular Informatics, Radboud Institute for Molecular Life Sciences, Radboud university medical centre, Geert Grootplein 28, 6525 GA Nijmegen, The Netherlands. 2 Department of Computer Science, San Diego State University, 5500 Campanile Drive, San Diego, California 92182, USA. 3 Department of Biology, San Diego State University, 5500 Campanile Drive, San Diego, California 92182, USA. 4 Department of Marine Biology, Institute of Biology, Federal University of Rio de Janeiro, Av. Carlos Chagas Fo. 373, Prédio Anexo ao Bloco A do Centro de Ciências da Saúde, Ilha do Fundão, CEP 21941-902 Rio de Janeiro, Brazil. 5 Computational Science Research Center, San Diego State University, 5500 Campanile Drive, San Diego, California 92182, USA. 6 Department of Microbiology, Institute for Water and Wetland Research, Radboud University, Heyendaalseweg 135, 6525 AJ Nijmegen, The Netherlands. 7 Department of Microbiology and Immunology, Faculty of Pharmacy, Cairo University, Kasr El-Aini Street, Cairo 11562, Egypt. 8 Department of Mathematics, San Diego State University, 5500 Campanile Drive, San Diego, California 92182, USA. 9 Division of Mathematics and Computer Science, Argonne National Laboratory, 9700 S Cass Ave B109, Argonne, Illinois 60439, USA. † Present address: Netherlands Institute of Ecology, Wageningen, The Netherlands. Correspondence and requests for materials should be addressed to B.E.D. (email: bedutilh@gmail.com).
n the past decade, metagenomic sequencing efforts have started to reveal the microbes inhabiting our planet1–3 and our body4–17, including bacteria and viruses. The first snapshots of the bacterial microbiome in the human gut that were taken by using metagenomics revealed a high interindividual diversity and many unknown genes14. More recently, large scale sequencing efforts revealed that in fact, many people share a similar intestinal flora, regardless of whether these similarities are viewed as discrete enterotypes15 or as gradients18. Endeavours including the Human Microbiome Project (HMP)19 and MetaHIT20 laid a baseline for the healthy human microbiome, enabling research laboratories around the world to anchor their work. Among the numerous breakthroughs in this field are the discovery of important functions for microbes in healthy individuals14, and the associations between specific members of our intestinal flora and unexpected diseases ranging from obesity19 to cancer20.

The viral community in our gut, also known as the human gut virome is dominated by bacteriophages (phages). Viral metagenomic studies have revealed that, in sharp contrast to the bacteria on which these phages depend for their replication, the viral sequences are mostly unknown, that is, they have no homologues in the database1–13. Moreover, gut viromes are thought to be highly individual specific as a result of the rapid sequence evolution of phages and the dynamic microbial ecosystem in the gut4–6. This leads to a virome with a vast, uncharted sequence space that is often referred to as biological ‘dark matter’ and provides an unprecedented opportunity for the discovery of novel viruses21.

Virus discovery efforts increasingly employ metagenomics as a relatively unbiased tool to explore and chart the virosphere. Several studies have used metagenomic shotgun reads to assemble complete viral genomes, including phages22 and viruses that infect humans23. In these studies, DNA may be derived from isolated virus-like particles (VLPs), but interestingly, DNA isolated from total community samples may also contain sequences of viral origin, estimates ranging up to 17% based on homology of the reads to known reference sequences6,17,24,25. Bypassing the need for sequence homology to identify specific phages, a recent study employed a search image based on a tetranucleotide usage profile to identify potential Bacterioides-like phages in published metagenomes24. While Bacteroides is one of the major bacterial taxa inhabiting our gut, only two phage genomes that infect this taxon were previously described22,26. In this previous study, a total of 85 contigs were identified that were potentially derived from Bacteroides-infesting phages24. The sequences were in the length order of known Bacteroides phage genomes, but the genome sequences were not closed.

There are two important hurdles to overcome in metagenomic virus discovery efforts. First, the sequences derived from one viral genome need to be identified among the mixed metagenomic reads (‘binning’) and assembled. Second, the role of the assembled genome in the gut ecosystem needs to be unravelled. In the case of a phage, this starts with the identification of its bacterial or archaeal host. Here, we address both these issues by exploiting the idea that interacting sequences co-occur across samples. Co-occurrence analysis is a strong tool to identify functionally related entities. For example, correlation of gene presence/absence across genomes, also known as phylogenetic profiling has been exploited to predict functional relationships between genes27. Similarly, co-occurrence across metagenomic samples has been used to predict ecological interactions between bacterial species28,29. Recently, similarity in read depth profiles across metagenomes has been introduced to bin contigs derived from the same genome, allowing the assembly of draft genomes from rare bacteria in the microbial community. This approach relies on the availability of multiple similar metagenomes, such as time series samples30 or metagenomes obtained after extracting DNA with different protocols31.

Here, we re-analyse previously published viral metagenomes, isolated from human faeces of 12 different individuals. These subjects comprised four pairs of healthy female monozygotic twins and their mothers from four unrelated families8. By using cross-assembly32, we create depth profiles of cross-contigs across individuals to discover sequences that co-vary in abundance across these metagenomes. After separate assembly of one viral metagenome, we obtain an ~97 kbp circular genome sequence of a novel bacteriophage (crAssphage) that is highly abundant in publicly available metagenomes. Next, we exploit the same concept of co-occurrence profiling to predict the phage–host relationship, since we expect that phages can only thrive when their host species is present. We perform phage–host prediction by using co-occurrence in parallel with several independent phage–host prediction approaches, including homology searches and identification of CRISPR spacers for the in silico prediction of phage–host relationships, all of which consistently suggest a Bacteroides host for crAssphage.

Results

Metagenome assembly and binning of ubiquitous contigs.

Metagenomes, and especially viral metagenomes, are often characterized by a majority of unknown sequences that have no homologues in the database21. Assembly of the short, second generation sequencing reads into longer contigs may facilitate their annotation and interpretation. Depending on the sequence diversity of the underlying microbial community, their genome sizes, and sequencing statistics including depth and read length, metagenome assembly can yield short or longer contigs, comprising genome sequences with different degrees of fragmentation. Binning of contigs derived from the same genome followed by re-assembly of the reads associated to these contigs can greatly improve assembly statistics31. Several binning strategies are commonly used. First, contigs derived from the same genome may be identified if they are homologous to a known reference genome, although this approach seems less suitable for completely novel genomes. Second, information from mate-paired sequencing reads may be exploited for de novo binning, if available. Third, oligonucleotide usage (k-mer profile binning) is a de novo binning approach suitable for longer contigs >1,000 nt, but it is not robust for short contigs33. Moreover, genomes with similar k-mer usage cannot be distinguished. Fourth, a recently introduced depth profile binning approach allowed the assembly of near-complete draft genomes from metagenome data30,31. Combining different metagenomic data sets in a cross-assembly is a simple way to create an occurrence profile for contigs, where the number of reads from a metagenome that are assembled into a contig represents the occurrence of that contig in the metagenome. The cross-assembly programme crAss32 generates such an occurrence profile for each contig. By normalizing for metagenome size, these occurrence profiles can be transformed into depth profiles for all contigs.

Here, we generated a de novo cross-assembly of 1,584,658 metagenomic reads derived from faecal viral metagenomes from twelve different individuals8. The 7,584 resulting cross-contigs had an N50 value of 2,638 nucleotides, and one short contig, contig07548 contained reads from all 12 individuals, indicating that it was possibly derived from a ubiquitous viral entity. Next, we used depth profile binning as well as homology binning to identify other contigs that were likely derived from the same ubiquitous viral genome as contig07548. First, we calculated Spearman’s correlation scores between the depth profile of contig07548 across the 12 samples, and the profiles of all other
contigs (Supplementary Data 1). This resulted in many contigs with highly similar abundance patterns across the viral metagenomes (Supplementary Data 1 and Supplementary Fig. 1), indicating that they had a similar occurrence in the original samples. Second, we performed a blastn search (e-value <10^-10) against Genbank, resulting in significant hits for 1,591 cross-contigs. The Genbank sequence that was most frequently identified as the top hit for a contig (214 times out of 7,584 contigs) was a clone in an unrelated human gut metagenome (Genbank PopSet identifier 259114965, see Methods). Although this sequence represents an unannotated clone and not a complete reference genome, these results strongly suggest that many of the assembled cross-contigs are actually derived from a single genome that was present in all the 12 personal viral metagenome data sets, hidden among the unknown reads.

**crAssphage genome assembly.** To investigate if the cross-contigs identified above were indeed derived from the same genome, we carefully re-assembled all the reads from one of the personal viral metagenomes. Because we expected the viral quasispecies to contain sequence heterogeneities, we used assembly settings that allowed such diversity to be collapsed (specifically, a short word length and a large bubble size, see Methods). We used the viral metagenome of Twin 1 from Family 2 (F2T1) for this assembly, because most reads in the ubiquitous cross-contigs were derived from this metagenome (Supplementary Data 1). This allowed us to construct a circular chromosome of 97,065 base pairs (bp) at an average depth of 230-fold in the F2T1 viral metagenomic sample (Fig. 1). We named this phage ‘crAssphage’ after the cross-assembly programme originally used to discover it. Although we used *de novo* assembly, the crAssphage genome aligned a total of 31 sequences from the unrelated human gut metagenome in Genbank mentioned above over 99.3% of its length (blastn e-value <10^-43, four small gaps of 7–406 bp remained). Note that this was an unrelated metagenome sequenced by a different laboratory, yet the average aligned sequence identity was 97.4%. This illustrates the extensive evolutionary conservation of the crAssphage genome sequence between the intestines of unrelated human individuals. Due to the

---

**Figure 1 | Schematic representation of the circular crAssphage genome.** The genome contains 80 ORFs that were predicted with Glimmer trained on Caudovirales. The total coverage of each nucleotide in the F2T1 metagenome, and in all public metagenomes in MG-RAST is indicated (466 human faecal and 2,440 other metagenomes, as determined by blastn mapping: ≥75 bp aligned with ≥95% identity, see Methods). Green bars indicate the 36 regions that were validated by long-range PCR (see Table 2 and Supplementary Table 1). Selected regions of several PCR amplicons (indicated as light green regions in the green bars) were sequenced by Sanger dideoxynucleotide sequencing to validate that the amplicons were indeed derived from the crAssphage genome (Supplementary Table 1). See Supplementary Fig. 6 for the fully annotated figure.
permissive assembly settings, the genome sequence is a consensus of the viral quasispecies population that occurs in the F2T1 personal viral metagenome, as is common in metagenome assembles46. To validate that this sequence represented an existing chromosome in the original viral sample, we used the metagenomic sequences to design long-range PCR primers along the entire length of the genome and were able to amplify products of the expected size from the F2T1 sample (Fig. 1 and Supplementary Fig. 2, original viral preparation kindly provided by the authors37) as well as from an unrelated faecal viral preparation (TSDC8.2, see Methods). A majority of the amplicons were partially sequenced by using Sanger sequencing to validate that they were indeed derived from the assembled crAssphage genome. As expected, higher sequence identity to the assembled crAssphage genome was observed between the Sanger sequencing traces derived from sample F2T1, than those from TSDC8.2 (Supplementary Table 1).

Encoded proteins. Open reading frame (ORF) prediction identified 80 protein coding genes. They are largely co-oriented, being organized in two large blocks of ORFs encoded on the same strand (Fig. 1). This genomic organization is typical of phage genomes, which frequently have a much larger fraction of their proteins encoded in sequential stretches on the same strand than bacterial genomes37. Functions could be annotated to a minority of the ORFs by a combination of homology searches in Genbank35 and the Phage Annotation Tools and Methods database (PhAnToMe, http://www.phantome.org/), and domain searches using HHPre38 (Supplementary Data 2). Viral structural proteins were predicted using iVireons39. Among the annotated ORFs, a tentative clustering of functionally related proteins could be observed (Supplementary Data 2). The HHPre search identified two hits to a Firmicute plasmid replication protein (RepL; PFam identifier PF05732 (ref. 40)) in orf00050 and orf00102. Interestingly, these two ORFs occur close to the location where the coding directionality switches from the forward to the reverse strand. These conserved domains may facilitate the independent replication of the phage genome. Moreover, we identified several Bacteroidetes-associated carbohydrate-binding (BACON; PFam identifier PF13004 (ref. 41)) domains present within orf00074 (Supplementary Fig. 5). BACON is a recently identified domain mediating adherence to glycoproteins41. It has thus far only been described as positive controls also cluster among Bacteroides species (Fig. 2). We also searched for sequence similarity of the 991 phage sequences previously identified by CRISPR targeting in human faecal metagenomes48 (blastn, e-value ≤10), but no matches were found between these sequences and the crAssphage genome.

Phage-host prediction. Clustered regularly interspersed short palindromic repeats provide a form of acquired immunity to phages and plasmids in bacteria, consisting of multiple short direct repeats, and spacers derived from the encountered foreign DNA such as phage genomes45. Thus, CRISPR spacers have been used to recognize the phages that previously infected a certain bacterial genome3. To determine the bacterial host of crAssphage, we performed CRISPR searches46 in 3,177 complete bacterial genomes. To provide a working immunity, CRISPR spacers should contain perfect matches to the viral genome, although recent studies have suggested that imperfect CRISPR matches could indicate recent interaction, facilitating rapid primed CRISPR adaptation47. In our analysis, none of the 93,276 identified CRISPR spacers had a perfect match to crAssphage. The most similar spacers were found in Prevotella intermedia 17 and in Bacteroides sp. 20_3, two intestinal species from the phylum Bacteroidetes (Fig. 2). We also searched for sequence similarity of the 991 phage sequences previously identified by CRISPR targeting in human faecal metagenomes48 (blastn, e-value ≤10), but no matches were found between these sequences and the crAssphage genome.

Next, we attempted to link crAssphage to a bacterial host by using co-occurrence profiling. Because phages can only thrive in an environment if their cellular host is present, we expected the occurrence of crAssphage and its host to show a correlated pattern, similar to the correlations between the depth profiles of contigs derived from the same genome. Thus, we compared the depth profile of the crAssphage genome and of 404 intestinal bacterial strains forming potential hosts, across an expanded data set of 151 faecal community shotgun metagenomes from the HMP16 (see Supplementary Data 3). These metagenomes contained both viral and bacterial sequences. We calculated Spearman’s correlation values between all depth profiles and created a co-occurrence cladogram to identify which bacterial depth profiles clustered with that of crAssphage. As shown in Fig. 3, crAssphage clusters deep within a group of Bacteroidetes genomes, suggesting one of the Bacteroides species as the most likely host. Similarly, two Bacteroides phages,22,26 that we included as positive controls also cluster among Bacteroides, although they recruited only a small fraction of metagenomic reads compared with crAssphage (104,076,280; 13,770; and 12,852 reads for crAssphage, B40-8 and B124-14, respectively).

A recent study that identified Bacteroides phages by using a tetranucleotide search image identified 85 sequences of > 10 kbp in size in published human faecal metagenomes44. None of these 85 sequences matched crAssphage in a homology search (blastn, e-value ≤10), while the data sets that were scanned for that previous study44, in fact contained 11 sequences >10 kbp and 72 sequences in total that were homologous to crAssphage, with a combined length of 408 kbp (results not shown). This emphasizes the uniqueness of the crAssphage genome both in terms of the sequence and in terms of its nucleotide-usage profile.

infected many, easily infectable bacteria13. Moreover, we suspect that the range of host taxa observed among the phage protein homologues reflects the sparseness of annotated sequence information available from gut phages. As a result, we mainly detected conserved or widespread phage proteins, including proteins involved in nucleic acid manipulation, and phage-structural proteins. Notably, the structural proteins are mostly associated with phage classes rather than with host classes. Finally, crAssphage rooted deeply in the phage protocline tree44 without close relatives (Supplementary Fig. 4). Together, this shows that crAssphage represents a highly divergent genome sequence, with few clues for determining its role in the intestinal ecosystem.
Table 1 | CrAssphage ORFs with homology to known proteins or domains.

| ORF | Function of top hits | Species | Host phylum |
|-----|----------------------|---------|-------------|
| orf00014 | Hypothetical protein | Phages | Ambiguous |
| orf00017 | Uracil-DNA glycosylase | *Acetivibrio cellulosolyticus* | Firmicutes |
| orf00016 | DNA helicase | *Francisella philomiragia* | Proteobacteria |
| orf00018 | DNA polymerase | *Labrenzia* | Proteobacteria |
| orf00025 | DNA primase/helicase | *Veillonella* sp. | Firmicutes |
| orf00029 | DNA ligase | *Erwinia* phage | Proteobacteria |
| orf00031 | Deoxynucleoside monophosphate kinase | *Enterobacteria* phage | Proteobacteria |
| orf00032 | Baseplate hub | *Aeromonas* phage | Proteobacteria |
| orf00033 | Thymidylate synthase complementing protein ThyX | *Prevotella* sp. | Bacteroidetes |
| orf00035 | Hypothetical protein | *Bacteroides* sp. | Bacteroidetes |
| orf00037 | Phage/plasmid-related protein | *Muclaginibacter* | Bacteroidetes |
| orf00038 | Deoxyuridine 5’-triphosphate nucleotidohydrolase | *Acinetobacter* phage | Proteobacteria |
| orf00039 | Endonuclease | *Paeibiculis* sp. | Firmicutes |
| orf00040 | Deoxyuridine 5’-triphosphate nucleotidohydrolase | *Salmonella* phage | Proteobacteria |
| orf00042 | Glutaredoxin/thioredoxin | Phages | Ambiguous |
| orf00047 | Hypothetical protein | *Clostridium* botellae | Firmicutes |
| orf00050 | Plasmid replication protein domain | *Firmicutes* | Firmicutes |
| orf00052 | Phage-structural protein | *Cellulophaga* phage | Bacteroidetes |
| orf00053 | Phage-structural protein | *Synechococcus* phage | Cyanobacteria |
| orf00056 | Hypothetical protein | *Escherichia* phage | Proteobacteria |
| orf00065 | Hypothetical protein | *Alistipes* putredinis | Bacteroidetes |
| orf00066 | Phage-related protein | *Escherichia* phage | Proteobacteria |
| orf00070 | Predicted protein | *Bacteroides* sp. | Bacteroidetes |
| orf00071 | Predicted protein | *Bacteroides* sp. | Bacteroidetes |
| orf00072 | Hypothetical protein | *Acinetobacter schindleri* | Proteobacteria |
| orf00073 | Phage-related protein | *Bacteroides* stercoris | Bacteroidetes |
| orf00074 | Phage-structural protein, contains BACON domains | *Bacteroides* sp. | Bacteroidetes |
| orf00075 | Phage-structural protein | *Mycobacterium* phage | Actinobacteria |
| orf00077 | Recombination endonuclease sunbunit | *Bacteroides* vulgatus | Bacteroidetes |
| orf00076 | Phage-related protein | *Desulfotobacterium* hafniense | Firmicutes |
| orf00086 | Phage-structural protein | *Veillonella* sp. | Firmicutes |
| orf00088 | Phage-structural protein | *Pseduomonas* phage | Proteobacteria |
| orf00091 | Phage-structural protein | *Cellulophaga* phage | Bacteroidetes |
| orf00092 | Hypothetical protein | *Veillonella* sp. | Firmicutes |
| orf00093 | DNA helicase | *Staphylococcus* phage | Firmicutes |
| orf00094 | Endolysin | *Marinilabilia* salmonicolor | Bacteroidetes |
| orf00095 | Endolysin | *Phage* | Proteobacteria |
| orf00096 | Phage-related protein | *Marinimicrobia* sp. | Marinimicrobia |
| orf00102 | Plasmid replication protein domain | *Firmicutes* | Firmicutes |

ORF, open reading frame. Function and taxonomy information of the hits are displayed. For details see Supplementary Data 2.

Figure 2 | CRISPR spacers similar to regions of the crAssphage genome.

CRISPR spacers were identified in 2,773 complete bacterial genomes from Genbank, and in 404 genomes of intestinal isolates from HMP and MetaHIT. The CRISPR spacers that were most similar to the crAssphage genome were found in *Prevotella intermedia* 17 (Genbank genomes) and in *Bacteroides* sp. 20_3 (HMP and MetaHIT genomes). Conserved A, C, G, and T nucleotides are displayed in red, green, yellow and blue, respectively.

Finally, plaque assays performed with phages from independent faecal isolates gave no additional insights (see Methods). None of the plaques that were observed on potential *Bacteroides* host strains tested positive using crAssphage-specific PCR primers.

Ubiquity of crAssphage in public metagenomes. Next, we determined the ubiquity of the crAssphage sequence in all 2,944 publicly available DNA shotgun metagenomes in the MG-RAST database49, and compared it with previously known phages. These public metagenomes were sampled from a range of microbial environments, and were not limited to the human gut. To do this, we created a database consisting of 1,193 phage genomes including crAssphage and all complete phage genomes in the PhAnToMe database, and used this to align the public metagenomic reads (blastn, ≥75 bp aligned, ≥95% identity, ambiguously aligned reads were discarded). We detected reads that uniquely aligned to crAssphage in 940 of these metagenomes, including both total community metagenomes and viral metagenomes, among which were 342 of the 466 faecal metagenomes (73%). Previous reports have estimated that total community samples may also contain sequences of viral origin, estimates ranging up to 17% (refs 6,8,17,24,25). Here, we observed that the crAssphage genome accommodated up to 90% of the sequencing reads in the faecal viral (that is, VLP-derived) metagenomes from the twin study that we used as a starting point6; up to 24% of the reads in an unrelated faecal viral metagenome from Korea13; and up to 22% of the reads in total faecal community metagenomes from USA16 (Supplementary Data 4). Across all the metagenomes, 235.8 million reads aligned to crAssphage.
uniquely to crAssphage. Almost all these hits (235.5 million or 99.9%) were derived from human gut metagenomes, comprising 1.68% of all the reads in the metagenomes that were annotated as being derived from human faeces. Thus, crAssphage is significantly more abundant in human faeces than in other environments (one-tailed Kolmogorov–Smirnov test, $P < 2.2 e^{-16}$). Based on this analysis, crAssphage is by far the most abundant, and one of the most ubiquitous bacteriophages in the publicly available metagenomes (Fig. 4 and Supplementary Data 5).

The read depth or recruitment profile of these reads along the crAssphage genome was stable in most of the public metagenomes (Fig. 5), although several conspicuous gaps remained, that were apparent in many of the public metagenomes (indicated with arrowheads at the top of Fig. 5). Similar recruitment gaps have recently been observed in marine bacteriophages50, and have been suggested to result from a Constant-Diversity or Red Queen dynamics. Dubbed ‘metaviromic islands’, the rare or divergent regions, may contain genes that are under positive evolutionary selection, including host recognition and other proteins50. Two such ORFs on the crAssphage genome, orf00039 and orf00050, encode proteins with homology to an endonuclease and a plasmid replication protein, respectively.

**Figure 3 | Phage-host prediction based on co-occurrence across metagenomes.** Unrooted co-occurrence cladogram of correlated depth profiles across 151 HMP faecal metagenomes10 of the crAssphage, two known Bacteroides fragilis-infecting phages22,26, and 404 potential hosts. Colours indicate bacterial phyla. The phages are indicated with blue dashed lines. See Supplementary Fig. 7 for the fully annotated figure.

**Figure 4 | Abundance ubiquity plot of phage genomes in public metagenomes.** Reads from 2,944 publicly available shotgun metagenomes49 were aligned to a database of 1,193 phage genomes (see Methods). The average depth of aligned reads per nucleotide of the phage genome (abundance) is plotted against the number of metagenomes it is found in (ubiquity). See Supplementary Data 5 for details.
Discussion

Many studies in the relatively young field of metagenomics aim to characterize environments in terms of the taxa and functions that are encoded on their communal genomes, or metagenomes. To do this, environmental shotgun sequencing reads are aligned to a reference database of annotated sequences using fast algorithms and intelligent shortcuts, but this approach may leave a large fraction of the reads unannotated. Viruses are especially notorious in this respect. Being both under-sampled and rapidly evolving, it has been suggested that the unknown fraction of metagenomes is enriched for viral sequences. These sequences may either be disregarded altogether, or reported in a pie chart as a slice labelled 'unknown', but either way they represent a sizeable elephant in the room. Everyone agrees the unknowns may be important, yet they are commonly ignored.

Here, we report the assembly of an ~97 kbp circular sequence from a previously published faecal viral metagenome. Our results strongly suggest that this sequence represents a phage genome that is highly divergent from the known phages currently present in the databases. First, the genome is highly abundant in some viral metagenomes that have been size- and density-filtered for VLPs, such as the F2T1 viral metagenome used to assemble it. Second, the ORFs encoded on the genome show similarity to bacteriophage and bacterial proteins. Although more than two thirds of the ORFs have no inferable function, we observe several functions that are typical of bacteriophages. Specifically, functions for DNA manipulation, replication and phage-structural proteins are encoded, while there is a distinct lack of any conserved bacterial or archaeal metabolic genes. Moreover, we observe a pattern of modularity among these functions. Third, the coding structure of the ORFs is typical of a phage genome. While the ORFs encoded on the circular genome change strand directionality at two positions, the ORFs encoded on bacterial genomes change strand directionality much...
more frequently. Fourth, although we did not find any transfer RNA genes (either prokaryotic or eukaryotic) in the genome sequence, the plethora of putative prokaryotic promoter patterns pillars the premise that this is a prokaryotic phage. Finally, we detected this genome in many metagenomes sequenced by different laboratories around the world, derived from different samples and sequenced using different sequencing technologies (Supplementary Data 4). This limits the possibility that the sequence is derived from contamination of reagents.

In this study, we exploited co-occurrence profiles across metagenomes in two ways to predict associations between sequence elements. First, we created cross-contigs by assembling the metagenomic reads from 12 personal viral metagenomes, and analysed their depth profiles using crAss. These depth profiles were highly similar for the majority of the ubiquitous cross-contigs, suggesting that they were derived from the same genome (Supplementary Data 1). This analysis also identified F2T1 as the person with the highest read depth for these ubiquitous contigs. Using this information, we then assembled the circular crAssphage genome from the F2T1 data set. By assembling only one metagenome, we minimized the heterogeneity in the assembled genome sequence. The genome assembly was validated with long-range PCR and Sanger dideoxynucleotide sequencing of several regions in two independent samples. We note that the presented genome sequence should be interpreted as a metapopulation consensus genome of the viral quasispecies that occurs in the sample, where some sequence diversity remains. For example, this can be observed in the sequencing results of the PCR products, where higher sequence identity to the crAssphage genome is observed in the traces derived from the original sample, F2T1 than in the unrelated sample TS8C8.2, although some mutations relative to the assembled sequence still remain, even in F2T1 (Supplementary Table 1).

Second, we used co-occurrence profiling of the newly identified phage genome with 404 potential hosts across 151 faecal shotgun metagenomes from the HMP. These total community metagenomes contained bacterial reads (the potential hosts) as well as viral reads (for example, crAssphage). As a result, the occurrence profiles in our analysis contain the same data sets, precluding any spurious correlations due to differences in the sampling or sequencing protocol. Here, we have used the non-parametric Spearman rank correlation to determine links between the phage and its host. We note that other correlation measures or a combination thereof might also be appropriate. Moreover, the correlation associations may be indirect. Although our approach could not specify the host to the level of species due to low correlation scores of crAssphage with the individual Bacteroidetes genomes, the clustering with this group was consistent with the other two Bacteroides phages (Fig. 3). Moreover, this host prediction is in line with the unique Bacteroides-associated BACR domain, the 12 annotated homologues in the phylum Bacteroidetes (Table 1) and the two most similar CRISPR spacers (Fig. 2). We are currently integrating these and other signals into a bioinformatic phage-host prediction framework to facilitate this important first step towards identifying the role of newly identified phages in the microbial ecosystem.

Although phages and the associated kill-the-winner dynamics may result in a loss of correlation between the occurrence patterns of phages and their hosts, we postulate that this process acts at the level of strains, or even sub-strain types diverging only at the level of, for example, their surface proteins. While one bacterial strain may be killed by expansion of a specific cognate phage, closely related bacteria may still be present in those same metagenomes because they will have similar niche preferences. This is also supported by the apparent constant-diversity (also known as Red Queen) dynamics acting on only a few of the ORFs, that is, those that under-recruit reads from public metagenomes and are observed as metaviromic islands (Fig. 5).

To summarize, here we identified and validated the ~97 kbp genome sequence of a novel bacteriophage, crAssphage. We show that its genome sequence is highly abundant and ubiquitous in publicly available metagenomes, and predominantly occurs in human faeces. This observation argues against the common view that the intestinal virome is unique to each individual, and suggests that some phages might be highly conserved in people around the world. Notably, little congruency was originally observed between the intestinal bacterial microbiota of different individuals, whereas many people in fact share a similar intestinal flora. Abundant phages and other mobile elements have previously been observed in metagenomes from the ocean and the human gut. Highly abundant, ubiquitous and conserved bacteriophages may remain hidden in the unknown fraction of metagenomes.

Methods

Metagenomic sequencing data. Metagenomic sequencing reads from human faecal viral metagenomes of four female twin pairs and their mothers were downloaded from the NCBI Short Read Archive (accession number SRA012183). Because the original authors showed that the intrapersonal diversity was low, we combined metagenomes for each individual, resulting in 12 personal faecal viral metagenomes.

Total shotgun metagenomes from human faeces, including both viral and bacterial sequences, were downloaded from HMP (151 Faecal Illumina WGS Reads and Assemblies HMIWGS/HMASM http://www.hmpdacc.org/HMASM/, the accession numbers are listed in Supplementary Data 3).

Assembly. An initial metagenomic cross-assembly of the twelve combined faecal viral metagenomes was constructed by using gsAssembler (≥65 nt overlap, ≥98 identity) yielding 7,584 cross-contigs as identified with crass version 1.2. By examining this cross-assembly, we discovered that many of these contigs had highly correlated depth profiles (Supplementary Data 1) and were homologous to a few long contig from the metagenome in Genbank with the PopSet identifier 239114965, suggesting that they were derived from the same genome. As the F2T1 viral metagenome contributed the most reads within these contigs, all the reads from this metagenome were re-assembled de novo with CLC Genomics Workbench 6.0.4 (word size: 35, bubble size: 1,000), yielding three contigs with similarly high depth (lengths 64,230; 31,634; and 1,269 bp) that overlapped, allowing them to be merged into a circular genome of 97,065 bp with an average per base depth of 230-fold. This crAssphage genome sequence was deposited in Genbank with the identifier JQ995537.

Primer design. To validate the crAssphage genome sequence, we designed 24 primer pairs covering the entire genome sequence (Table 2). Genome-wide primer design using iPCR was supplemented with a conservation analysis of the sequence in the 12 personal viral metagenomes to ensure they matched within conserved genomic regions. The regions selected for amplification were between 498 bp and 18,342 bp in length (Supplementary Fig. 2).

Long-range PCR. To validate the genome assembly of crAssphage, we obtained the original F2T1.2 sample of faecal viral DNA from the authors (sample identifier: TW4.2). Moreover, an independent faecal viral sample was also included, obtained from a healthy adult monoygotic twin as part of a human gut microbiome surveym (sample identifier: TSDC8.2). The sample was processed for isolation of VLP-associated DNA as described. The PCR was performed using the Thermo Scientific Hi-Fidelity Extensor Long-Range PCR Enzyme kit with 9 μl of the 2 μl mix and 14 μl of water per reaction (Cat. No. AB-0702/2A and AB-0720/B). A touchdown method was used for the PCR cycling conditions to account for differing melting temperatures of the primers. PCR cycling conditions were as follows: (1) denaturation at 95°C for 2 min; (2) denaturation at 95°C for 10 s; (3) annealing at 58°C for 30 s; touchdown at −0.5°C; (4) extension at 68°C for 6.5 or 10 min; (5) extension at 68°C for 2.5 or 3.5 min. Cycle steps 2–4 thirty times; 6) final elongation at 68°C for 10 or 15 min; store samples at 4°C until further use. Extension times were adjusted for the expected product size, for example, for products of primer

Assembly. An initial metagenomic cross-assembly of the twelve combined faecal viral metagenomes was constructed by using gsAssembler (≥65 nt overlap, ≥98 identity) yielding 7,584 cross-contigs as identified with crass version 1.2. By examining this cross-assembly, we discovered that many of these contigs had highly correlated depth profiles (Supplementary Data 1) and were homologous to a few long contig from the metagenome in Genbank with the PopSet identifier 239114965, suggesting that they were derived from the same genome. As the F2T1 viral metagenome contributed the most reads within these contigs, all the reads from this metagenome were re-assembled de novo with CLC Genomics Workbench 6.0.4 (word size: 35, bubble size: 1,000), yielding three contigs with similarly high depth (lengths 64,230; 31,634; and 1,269 bp) that overlapped, allowing them to be merged into a circular genome of 97,065 bp with an average per base depth of 230-fold. This crAssphage genome sequence was deposited in Genbank with the identifier JQ995537.

Primer design. To validate the crAssphage genome sequence, we designed 24 primer pairs covering the entire genome sequence (Table 2). Genome-wide primer design using iPCR was supplemented with a conservation analysis of the sequence in the 12 personal viral metagenomes to ensure they matched within conserved genomic regions. The regions selected for amplification were between 498 bp and 18,342 bp in length (Supplementary Fig. 2).

Long-range PCR. To validate the genome assembly of crAssphage, we obtained the original F2T1.2 sample of faecal viral DNA from the authors (sample identifier: TW4.2). Moreover, an independent faecal viral sample was also included, obtained from a healthy adult monoygotic twin as part of a human gut microbiome surveym (sample identifier: TSDC8.2). The sample was processed for isolation of VLP-associated DNA as described. The PCR was performed using the Thermo Scientific Hi-Fidelity Extensor Long-Range PCR Enzyme kit with 9 μl of the 2 μl mix and 14 μl of water per reaction (Cat. No. AB-0702/2A and AB-0720/B). A touchdown method was used for the PCR cycling conditions to account for differing melting temperatures of the primers. PCR cycling conditions were as follows: (1) denaturation at 95°C for 2 min; (2) denaturation at 95°C for 10 s; (3) annealing at 58°C for 30 s; touchdown at −0.5°C; (4) extension at 68°C for 6.5 or 10 min; (5) extension at 68°C for 2.5 or 3.5 min. Cycle steps 2–4 thirty times; 6) final elongation at 68°C for 10 or 15 min; store samples at 4°C until further use. Extension times were adjusted for the expected product size, for example, for products of primer
pair F3R3 with expected size 5.4 kbp, the extension time was 6.5 min. A 12

similarly, the occurrence pattern was measured of two positive controls, the known

Table 2 | PCR primer pairs designed to validate the crAssphage genome sequence.

| Nr   | Forward primer sequence                      | Start               | Reverse primer sequence                      | End               |
|------|---------------------------------------------|---------------------|---------------------------------------------|------------------|
| 1    | 5′-TGCACGGAGGATTTAAGTGA-3′                 | 17,851              | 5′-GCATTAAGTCCACGAGA-3′                     | 6,793            |
| 2    | 5′-TCACCTTGTGAGCCATTT-3′                  | 20,735              | 5′-ATGAAATGAGCCGTTCAT-3′                    | 9,073            |
| 3    | 5′-ATTTTACTTATCCTCTTCC-3′                | 23,484              | 5′-TCTTAAGTGAGCCTTTCG-3′                    | 11,923           |
| 4    | 5′-AGTATTTCTTTTCTTGCTCAG-3′              | 25,822              | 5′-ATGGAAACGCTGCTTGC-3′                     | 14,930           |
| 5    | 5′-GCCCGGTGAGCGAGCT-3′                   | 27,112              | 5′-ATCGAAGAGCTGCTTGC-3′                     | 20,687           |
| 6    | 5′-GTCAGATTTCTCCTGCTTCAG-3′              | 29,355              | 5′-ATGAAAGAGCTGCTTGC-3′                     | 23,713           |
| 7    | 5′-TGGACGTTTGTGAGGCT-3′                  | 31,507              | 5′-ATGGAAAGAGCTGCTTGC-3′                    | 26,855           |
| 8    | 5′-GCCCGGTGAGCGAGCT-3′                   | 33,656              | 5′-ATGGAAAGAGCTGCTTGC-3′                    | 29,913           |
| 9    | 5′-ATTTTACTTATCCTCTTCC-3′                | 35,807              | 5′-ATGGAAAGAGCTGCTTGC-3′                    | 33,065           |
| 10   | 5′-GCCCGGTGAGCGAGCT-3′                   | 37,958              | 5′-ATGGAAAGAGCTGCTTGC-3′                    | 36,217           |

Numbers in the first column correspond to the bands in Supplementary Table 1. Numbers in the Start and End columns refer to the position on the crAssphage genome.
Phages in metagenomes. To determine the prevalence of the crAss phage genome across different environments, we downloaded sequencing reads from all 2,944 publicly available shotgun metagenomes available in the MG-RAST database. A database of 1,193 phage genomes was created by adding crAssphage to all complete phage genomes from PhAniToMe (http://www.phantome.org). A blastn search of every metagenomic read versus the phage genome database was performed, and hits were considered if they had at least 95% identity over an aligned length of 75 bp. Ambiguously aligned reads were discarded.

References

1. Tyson, G. W. et al. Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature 428, 37–43 (2004).
2. Breitbart, M. et al. Genomic analysis of uncultured marine viral communities. Proc. Natl Acad. Sci. USA 99, 14250–14255 (2002).
3. Cassman, N. et al. Oxygen minimum zones harbour novel viral communities with low diversity. Environ. Microbiol. 14, 3043–3065 (2012).
4. Minot, S. et al. Rapid evolution of the human gut virome. Proc. Natl Acad. Sci. USA 110, 12450–12455 (2013).
5. Minot, S., Grunberg, S., Wu, G. D., Lewis, J. D. & Bushman, F. N. Hypervariable loci in the human gut virome. Proc. Natl Acad. Sci. USA 109, 3962–3966 (2012).
6. Minot, S. et al. The human gut virome: inter-individual variation and dynamic response to diet. Genome Res. 21, 1616–1625 (2011).
7. Reyes, A., Wu, M., McNulty, N. P., Rohwer, F. L. & Gordon, J. I. Gnotobiotic mouse model of phage-bacterial host dynamics in the human gut. Proc. Natl Acad. Sci. USA 110, 20236–20241 (2013).
8. Reyes, A. et al. Viruses in the faecal microbiota of monozygotic twins and their mothers. Nature 466, 334–338 (2010).
9. Zhang, T. et al. RNA viral community in human feces: prevalence of plant pathogenic viruses. PLoS Biol. 4, e3 (2006).
10. Nakamura, S. et al. Direct metagenomic detection of viral pathogens in nasal and faecal specimens using an unbiased high-throughput sequencing approach. PLoS ONE 4, e2129 (2009).
11. Nakamura, S. et al. Metagenomic diagnosis of bacterial infections. Emerg. Infect. Dis. 14, 1784–1786 (2008).
12. Breitbart, M. et al. Metagenomic analyses of an uncultured viral community from human feces. J. Bacteriol. 185, 6220–6223 (2003).
13. Kim, M. S., Park, E. J., Roh, S. W. & Bae, J. W. Diversity and abundance of single-stranded DNA viruses in human feces. Appl. Environ. Microbiol. 77, 8062–8070 (2011).
14. Gill, S. R. et al. Metagenomic analysis of the human distal gut microbiome. Science 312, 1355–1359 (2006).
15. Knight, R. et al. Enterotypes of the human gut microbiome. Nature 473, 174–180 (2011).
16. Peterson, J. et al. The NIH Human Microbiome Project. Genome Res. 19, 2317–2323 (2009).
17. Qin, J. et al. A human gut microbial gene catalogue established by metagenomic sequencing. Nature 464, 59–65 (2010).
18. Kolen, O. et al. A guide to enterotypes across the human body; meta-analysis of microbial community structures in human microbiome datasets. PLoS Comput. Biol. 9, e1002863 (2013).
19. Turnbaugh, P. J. et al. An obesity-associated gut microbiome with increased capacity for energy harvest. Nature 444, 1027–1031 (2006).
20. Tjalsma, H., Bolej, A., Marchesi, J. R. & Dutilh, B. E. A bacterial driver-passenger model for colorectal cancer: beyond the usual suspects. Nat. Rev. Microbiol. 10, 575–582 (2012).
21. Mokili, J. L., Rohwer, F. & Dutilh, B. E. Metagenomics and future perspectives in virus discovery. Curr. Opin. Virol. 2, 63–77 (2012).
22. Ogilvie, L. A. et al. Comparative (metagenomic) analysis and ecological profiling of human gut-specific bacteriophage phiB124-14. PLoS ONE 7, e35053 (2012).
23. Mokili, J. L. et al. Identification of a novel human papillomavirus by metagenomic analysis of samples from patients with febrile respiratory illness. PLoS ONE 8, e85404 (2013).
24. Ogilvie, L. A. et al. Genome signature-based dissection of human gut metagenomes to extract subliminal viral sequences. Nat. Commun. 4, 2420 (2013).
25. Waller, A. S. et al. Classification and quantification of bacteriophage taxa in human gut metagenomes. ISME J. 8, 1551–1552 (2014).
26. Hawkins, S. A., Layton, A. C., Ripp, S., Williams, D. & Sayler, G. S. Genome sequence of the Bacteroides fragilis phage ATCC 51477-B1. Virol. J. 5, 97 (2008).
27. Kenoue, M. et al. Recent advances in revealing and quantifying bacteriophage diversity in the human gut. Front. Cell. Infect. Microbiol. 4, 1–18 (2014).
28. Dutilh, B. E., Huynen, M. A. & Strous, M. The effect of a double-stranded DNA virus on the human gut virome. PLoS ONE 7, e44904 (2012).
29. Dutilh, B. E., Huynen, M. A. & Strous, M. The effect of a double-stranded DNA virus on the human gut virome. PLoS ONE 7, e44904 (2012).
30. Wrighton, K. C. et al. Metagenomics and future perspectives in virus discovery. Curr. Opin. Virol. 3, 33–40 (2013).
31. Breitbart, M. et al. Microbial co-occurrence relationships in the human gut metagenome. PLoS Comput. Biol. 8, e1002606 (2012).
32. Akhter, S., Aziz, R. K. & Edwards, R. A. PhiSpy: a novel algorithm for finding prophages in bacterial genomes that combines similarity- and composition-based strategies. Nucleic Acids Res. 40, e126 (2012).
33. Soding, J. & Biegert, A. & Lupas, A. N. The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res. 33, W244–W248 (2005).
34. Seguritan, V. et al. Artificial neural networks trained to detect viral and phage structural proteins. PLoS Comput Biol. 8, e1002657 (2012).
35. Finn, R. D. et al. The Pfam protein families database. Nucleic Acids Res. 38, D211–D222 (2010).
36. Mello, L. V., Chen, X. & Ridgen, D. J. Mining metagenomic data for novel domains: BACON, a new carbohydrate-binding module. FEMS Lett. 584, 2421–2426 (2010).
37. Barr, J. J. et al. Bacteriophage adhering to mucus provide a non-host-derived immunity. Proc. Natl Acad. Sci. USA 110, 10771–10776 (2013).
38. Flores, C. O., Meyer, J. R., Valverde, S., Farr, L. & Weitz, J. S. Statistical structure of host-phage interactions. Proc. Natl Acad. Sci. USA 108, E288–E297 (2011).
39. Rohwer, F. et al. The phage proteome tree: a genome-based taxonomy for phage. J. Bacteriol. 184, 4529–4535 (2002).
40. Barrangou, R. et al. CRISPR provides acquired resistance against viruses in prokaryotes. Science 315, 1709–1712 (2007).
41. Edgar, R. C. & Myers, E. W. PILER: identification and classification of genomic repeats. Bioinformatics 21(Suppl 1): i152–i158 (2005).
42. Futterer, P. C. et al. Degenerate target sites mediate rapid primed CRISPR adaptation. Nat. Protoc. Lett. 111(1), E1629–E1638 (2014).
43. Stern, A., Micks, E., Tirosh, I., Sagi, O. & Sorek, R. CRISPR targeting reveals a reservoir of common phages associated with the human gut microbiome. Genome Res. 22, 1985–1994 (2012).
44. Meyer, F. et al. The metagenomics RAST server—a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics 9, 386 (2008).
45. Mizuno, C. M., Ghai, R. & Rodriguez-Valera, F. Evidence for metaviromic islands in marine phages. Front. Microbiol. 5, 27 (2014).
46. Li, S. C. et al. UMAPS: Un-MiAppable Reads Solution. BMC Bioinformatics 12(Suppl 1): S9 (2011).
47. Zhao, Y. et al. Abundant SAR11 viruses in the ocean. Nature 494, 357–360 (2013).
48. Margulies, M. et al. Genome sequencing in micromanufactured high-density polycrile reactors. Nature 437, 376–380 (2005).
NATURE COMMUNICATIONS | DOI: 10.1038/ncomms5498

ACKNOWLEDGEMENTS

We thank Alejandro Reyes, Jeffrey Gordon and Forest Rohwer for access to virome materials and fruitful discussions; Michiyu Wellington-Oguri for initial host predictions; Anca Segall for help with iVireons; and Cynthia Sears (NIH-R01CA151393) for screening Bacteroides genomes. B.E.D. was supported by NWO Veni (016.111.075). CAPES/BRASIL and the Dutch Virgo Consortium (FES0908, NGI 050-060-452); D.R.S. by was supported BE-Basic (fp6702); and R.A.E. was supported by grants from the National Science Foundation (DBI-0850356, MCB-1330800, and DEB-1046413). High performance computation was provided by award CNS-1305112 from the Information and Intelligent Systems Division of the National Science Foundation to R.A.E.

AUTHOR CONTRIBUTIONS

B.E.D., B.F., J.L.M., R.A.E. performed and analysed the initial cross-assembly; B.E.D., D.R.S. assembled crAssphage genome; B.E.D., J.L.M. designed PCR primers; N.C., J.L.M., E.A.D. performed long-range PCR experiments; B.E.D., G.G.Z.S., J.L.M. analysed Sanger sequencing results; B.E.D., J.J.B., V.S., R.K.A., R.A.E. annotated crAssphage genome and ORFs; R.A.E., phage proteomic tree; B.E.D., bioinformatic host predictions (co-occurrence and CRISPRs); S.S., L.B., plaque assays; B.E.D., K.M. analysed public metagenomes; B.E.D. wrote paper with input from all authors.

ADDITIONAL INFORMATION

Accession codes: The crAssphage genome sequence was deposited in the Genbank Nucleotide sequence database with accession code JQ995537. Sanger sequenced regions of the PCR products from F2T1 and TS2C8.2 (indicated by the light green regions of the green bands of Fig. 1) were deposited in the Genbank nucleotide sequence database with accession codes KM000086 to KM000121 (see Supplementary Table 1).

Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications

COMPETING FINANCIAL INTERESTS: The authors declare no competing financial interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

HOW TO CITE THIS ARTICLE: Dutilh, B. E. et al. Unknown sequences in faecal metagenomes reveal a widely distributed and highly abundant bacteriophage. Nat. Commun. 5:4498 doi: 10.1038/ncomms5498 (2014).

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-sa/4.0/