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

Discovery of an expansive bacteriophage family that includes the most abundant viruses from the human gut

Natalya Yutin[1], Kira S. Makarova[1], Ayal B. Gussow[1], Mart Krupovic[2], Anca Segall[1,3], Robert A. Edwards[3] and Eugene V. Koonin[1]*

Metagenomic sequence analysis is rapidly becoming the primary source of virus discovery[1–3]. A substantial majority of the currently available virus genomes come from metagenomics, and some of these represent extremely abundant viruses, even if never grown in the laboratory. A particularly striking case of a virus discovered via metagenomics is crAssphage, which is by far the most abundant human-associated virus known, comprising up to 90% of sequences in the gut virome[4]. Over 80% of the predicted proteins encoded in the approximately 100 kilobase crAssphage genome showed no significant similarity to available protein sequences, precluding classification of this virus and hampering further study. Here we combine a comprehensive search of genomic and metagenomic databases with sensitive methods for protein sequence analysis to identify an expansive, diverse group of bacteriophages related to crAssphage and predict the functions of the majority of phage proteins, in particular those that comprise the structural, replication and expression modules. Most, if not all, of the crAss-like phages appear to be associated with diverse bacteria from the phylum Bacteroidetes, which includes some of the most abundant bacteria in the human gut microbiome and that are also common in various other habitats. These findings provide for experimental characterization of the most abundant but poorly understood members of the human-associated virome.

Viruses are the most abundant biological entities on Earth. In most environments, from ocean water to the content of animal guts, the number of detected virus particles exceeds that of cells by one to two orders of magnitude[2]. Among these viruses, more than 90% are tailed bacteriophages[1]. More than 99% of the prokaryotic diversity in the biosphere is represented by bacteria and archaea that fail to grow in laboratory cultures and, accordingly, the great majorityof the viruses are thought to infect these uncultivated microbes[1]. Moreover, analysis of the human gut virome shows that most of the sequences, in contrast to the bacterial and archaeal sequences, have no matches in the current sequence databases, suggesting a vast virome consisting primarily of 'dark matter'[5–7].

The crAssphage is the utmost manifestation of this trend. The complete crAssphage (after Cross Assembly) genome was assembled by joining contigs obtained from several human faecal viral metagenomes as a circular double-stranded (ds) DNA molecule of ~97 kilobases (kb)[4]. The circular genome map apparently results from terminal redundancy and/or circular permutation. The crAssphage is extremely abundant, accounting for up to 90% of the reads in the virus-like particle-enriched fraction of the gut metagenome and about 22% of the reads in the total metagenome. Numerous reads matching the crAssphage genome have been identified in numerous gut metagenomes collected in diverse geographic locations, indicating that crAssphage is not only the most abundant virus in the human gut microbiome but also a (nearly) ubiquitous one[4,8,9]. Read co-occurrence analysis points to bacteria of the phylum Bacteroidetes as the host(s) of crAssphage[4,10]. This assignment is compatible with the presence in the crAssphage genome of a protein containing carbohydrate-binding domains (BACON domains) that is highly similar to a homologous protein from Bacteroides and with partial matches between two crAssphage sequences and CRISPR spacers from two species of Bacteroides[4]. Members of the Bacteroidetes dominate the gut microbiome, but most of these bacteria so far have not been grown in culture[11,12]. Thus, it is hardly surprising that the most abundant—but never isolated—phage from this environment appears to be a parasite of Bacteroidetes. Analysis of the protein sequences encoded in the crAssphage genome failed to identify specific relationships with other bacteriophages[4]. Several proteins implicated in phage genome replication have been identified, including a family of B DNA polymerase (DNAP), a primase and a flavin-dependent thymidylate synthase, but neither the major capsid protein nor other structural and morphogenetic proteins were detected. In an attempt to clarify the provenance of this most abundant but enigmatic human-associated virus, we reanalysed the crAssphage genome using the most sensitive available methods for protein sequence analysis and taking advantage of database growth since the time of crAssphage discovery. The result is the identification of a previously unknown, expansive bacteriophage family that appears to be associated with diverse members of Bacteroidetes and for which we now recognize the structural, replication and expression gene modules.

The sequences of the crAssphage proteins were compared (using PSI-BLAST) to the non-redundant protein sequence database (nr) and the Whole Genome Shotgun (WGS) databases (NCBI, NIH, Bethesda) containing microbial genomic and metagenomic sequences. Sequences with significant similarity to crAssphage proteins were detected in four genomes of previously identified bacteriophages and numerous contigs assigned to bacterial genomes (possibly, prophages) and metagenomic contigs. These sequences

[1]National Center for Biotechnology Information, National Library of Medicine, Bethesda, MD, USA. [2]Institut Pasteur, Unité Biologie Moléculaire du Gène chez les Extrêmophiles, Paris, France. [3]Viral Information Institute, Department of Biology, San Diego State University, San Diego, CA, USA.
*e-mail: koonin@ncbi.nlm.nih.gov

were highly diverse and most were not closely related (despite the statistical significance of the detected similarity) to the crAssphage proteins (Supplementary Table 1). Thus, crAssphage relatives identified here might comprise a previously unidentified, large, diverse family of bacteriophages (henceforth crAss-like family) or potentially even two or more families. Altogether, we identified several hundred putative representatives of the crAss-like phage family (Supplementary Fig. 1). Of these, 37 diverse representatives, for which (nearly) complete genomes were available, were selected for in-depth analysis (Supplementary Table 1). We then constructed multiple alignments of the crAssphage proteins and their homologues and used these alignments as queries for profile–profile searches against a comprehensive collection of protein families using the HHpred software, one of the most sensitive current methods for protein sequence analysis (see Methods for details). We identified a block of five genes that appear to comprise the structural module of the emerging family of bacteriophages. These genes encode a predicted major capsid protein (MCP) of the HK97 fold, portal protein, large terminase subunit and two uncharacterized proteins that are conserved throughout the crAss-like family and, given the consistent adjacency to the MCP, could be components of the virion or the morphogenetic machinery (Table 1, Fig. 1, Supplementary Table 2 and Supplementary Note 1).

Despite the low sequence conservation, even within the family, and remote similarity to proteins from other phages, the gene order in the capsid module of the crAss-like family is nearly invariant (Fig. 1), suggesting congruent evolution of these genes. A concatenated alignment of all five genes was used to construct a phylogenetic tree of the crAss-like family, which includes a strongly supported clade of crAssphage relatives and several other distinct groups of (predicted) bacteriophages (Fig. 1). Three of these groups included previously identified phages, namely Azobacteroides phage ProJPt-Bp[13], *Flavobacterium psychrophilum* phage Fpv3 and *Cellulophaga* phage phi14:2, which are widespread, although apparently not highly abundant phages in the oceans[14]. The same group that included the *Cellulophaga* phage also contained the genome of the IAS (immunodeficiency-associated stool) virus, which is highly abundant in gut viromes of HIV-infected individuals[15]. Most of the other members of the crAss-like family are unassigned metagenomic sequences, but several come from bacterial genome assemblies and might represent prophages. All experimentally characterized crAss-like phages are associated with Bacteroidetes. Among the other sequences included in the family, several are assigned to other bacteria, in particular, *Chlamydia trachomatis*, as well as several members of the recently identified candidate phyla radiation (CPR)[16]. However, in the phylogenetic trees of the predicted MCP, these sequences are embedded within groups consisting of sequences associated with Bacteroidetes (Fig. 1 and Supplementary Fig. 1), and none of these contigs contained genes that could be linked to the host bacteria. Thus, the available data appear compatible with an exclusive association between the crAss-like phages and Bacteroidetes.

In the linear genome maps shown in Fig. 2 (see Supplementary Table 1 and Supplementary Note 1 for the crAssphage and IAS phage gene annotations), the capsid structural module occupies about 10 kb near one end of the crAss-like phage genomes. Downstream of this module are the genes encoding predicted tail proteins and two proteins homologous to the bacterial integration host factor (IHF) that is essential for chromatin packaging in bacteria and some phages[17] (Fig. 2, Table 1 and Supplementary Table 2). The two most conserved tail proteins encoded by the crAss-like group are homologous to tail components gene product (gp)4 (tubular tail protein) and gp10 (tail stabilization protein) of bacteriophage P22[18]. In the same putative operon, crAssphage also encodes a homologue of the tail needle protein gp26 of phage P22. These three proteins are sufficient for the formation of a short tail similar to that of bacteriophage P22[19].The gp10 homologues of crAss-like phages are large (>1,400 aa), apparently multidomain proteins in which the gp10-homologous region accounts only for ~150 aa. The additional domains of this protein could be involved in host recognition similarly to the tail spike protein of P22-like phages[20]. Some crAss-like phages encode additional, auxiliary proteins in the tail module, for example, an IAS protein homologous to the tail-associated lysozyme gp13 of short-tailed phage phi29[21]. Thus, the crAss-like phages can be predicted to possess short, stubby tails, a hallmark of the family Podoviridae. One of the isolated crAss-like phages, *Cellulophaga* phage phi14:2, is indeed a typical podophage[14]. Unlike phages with long tails (families Myoviridae and Syphoviridae), P22-like phages do not encode maturation proteases[22], consistent with the apparent absence of such a protease among gene products of the crAss-like phages. However, in the midst of the genes for predicted tail components, some of the crAss-like phages, including the crAssphage group, encode a predicted Zn-dependent protease (Fig. 2) that might be involved in processing of the tail and/or capsid proteins. In crAssphage group genomes, the protease gene is embedded within a block of genes encoding putative additional tail components, which are highly similar to homologues from uncharacterized prophages integrated in Bacteroides genomes that are otherwise unrelated to crAss-like phages (Fig. 2 and Supplementary Table 2). Thus, evolution of the crAssphage group apparently involved relatively recent recombination with an unrelated (pro)phage from the same host(s).

The lytic replication module genes occupy about 30 kb on the opposite end of the genome from the capsid module and are transcribed towards the middle of genome (Fig. 2). This module encodes a versatile suite of proteins implicated in DNA replication and repair and shows a patchy gene distribution, without a single universally conserved gene and a much greater variability within the crAss-like family than the structural modules (Table 1 and Fig. 3). The most conserved replicative gene is a predicted DnaG family primase. Many crAss-like family members also encode a putative superfamily 2 (SNF2 family) helicase implicated in DNA replication (this helicase is inactivated in the crAssphage subfamily, as indicated by mutiple amino acid replacements in the catalytic sites), an ATP-dependent DNA ligase, a uracyl-DNA glycosylase (UDG), a flavin-dependent thymidylate synthase (ThyX) and one or two diverged single-stranded DNA-binding proteins (SSB) (Table 1 and Fig. 3). The crAss-like family viruses encode one of the two distinct DNA polymerases (DNAPs): the crAssphage clade has a family B DNAP, whereas the other family members have either a family B or a family A DNAP, or no DNAP at all (Table 1 and Fig. 3). Evolutionary reconstruction suggests that the family A DNAP is ancestral in crAss-like phages and was lost or replaced with the family B DNAP on several occasions (Supplementary Fig. 2). The only near-universal replicative protein, DNA primase, appears to be monophyletic in the crAss-like family and forms a strongly supported clade with the primases of Bacteroidetes (Supplementary Fig. 2). This is the only conserved crAss-like family gene that shows a deep connection to Bacteroidetes, indicating that a founder crAss-like phage acquired this gene from a Bacteroidetes host and implying that the virus–host link is evolutionarily ancient. An unusual evolutionary connection was detected for the phage ligase: in the phylogenetic tree, the crAss-like family ligases clustered with those of eukaryotic giant dsDNA viruses, suggesting that these viruses acquired the ligase from crAss-like phages (Supplementary Fig. 2). Portions of the replicative gene block of the IAS virus and its closest relatives show high similarity to homologues from putative prophages of Bacteroidetes, suggesting another recombination event (Fig. 2 and Supplementary Table 2).

The structural and replicative gene blocks of the crAss-like phages are separated by an array of uncharacterized genes that are transcribed in the same direction as the structural genes and are universally conserved across the crAss-like family (Fig. 2 and Table 1). Several of these

Table 1 | Conserved genes in the crAss-like phage family

Gene	crAss phage gene number	IAS virus gene number	cr Ass phage*	CDZK 010154 69*	CEAR 0102 9167*	LSPZ 0100 0006*	LAZR 0100 0126	CESO 0103 0555	BCSF 0100 0013	LSPY 0100 0004	*Azo bact_* ph_ AP01 7903
Core genes of crAss-like family[a]											
Terminase	79	15	Y	Y	Y	Y	Y	Y	Y	Y	Y
Portal	78	16	Y	Y	Y	Y	Y	Y	Y	Y	Y
putative structural protein	77	32	Y	Y	Y	Y	Y	Y	Y	Y	Y
MCP	76	33	Y	Y	Y	Y	Y	Y	Y	Y	Y
putative structural protein	75	34	Y	Y	Y	Y	Y	Y	Y	Y	Y
putative structural protein	74	35,36	Y	Y	Y	Y	Y	Y	Y		Y
putative structural protein	73	–	Y	Y	Y	Y		Y	Y	Y	
IHF subunit	54	40	Y	Y	Y	Y	Y	Y	Y	Y	Y
Tail tubular protein (P22 gp4-like)	52	34	Y	Y	Y	Y	Y	Y	Y	Y	Y
Tail stabilization protein (P22 gp10-like)	51	44	Y	Y	Y	Y	Y	Y	Y	Y	Y
Putative RNAP subunit	47	47	Y	Y	Y	Y	Y	Y	Y	Y	Y
β–β′ RNAP	46N	49	Y	Y	Y	Y	Y	Y	Y	Y	Y
Putative RNAP subunit	46C	48	v	Y	Y	Y	?	?	Y	?	Y
DnaG family primase	22	80_79	Y	Y	Y	Y	Y	Y	Y	Y	
Conserved genes in the crAssphage group											
Uncharacterized protein	86	13	Y	Y	Y	Y	Y	Y	Y		
Uncharacterized protein	49	45	Y	Y	Y	Y		Y	Y		
Uncharacterized protein	34	–	Y	Y	Y	Y					
Siphovirus Gp157 homolog	32	–	Y	Y	Y	Y	Y				
Uncharacterized protein	23	–	Y	Y	Y	Y					
ssb	21	–	Y	Y	Y	Y					
RecT	19	–	Y	Y	Y	Y					
SNF2 helicase[b]	18	87	Y	Y	Y	Y	Y	Y	Y		
PolB	17	–	Y	Y	Y	Y					
Uncharacterized protein	20	–	Y	Y	Y	Y					
UDG	16	75	Y	Y	Y	Y	Y		Y		
SF1 helicase	15	–	Y	Y	Y	Y					

Continued

Table 1 | Conserved genes in the crAss-like phage family (continued)

Gene	*Chlamydia_* CVNZ01000019ext	contig 0001	FUFK01003914 1	Woesebacteria_ MGFQ01000035	*Chitinophaga_* FOJF01000001	*Flavobact_ph_* NC_031904	MDTC01014143	CEUT01009082/	*Cellulophaga_ph_NC_* 021806	IAS_virus_ KJ003983
Core genes of crAss-like family[a]										
Terminase	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Portal	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
putative structural protein	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
MCP	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
putative structural protein	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
putative structural protein	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
putative structural protein		Y	Y	Y	Y	Y	Y	Y	Y	
IHF subunit		Y	Y	Y	Y	Y	Y	Y	Y	Y
Tail tubular protein (P22 gp4-like)		Y	Y		Y	Y	Y	Y	Y	Y
Tail stabilization protein (P22 gp10-like)		Y	Y	Y	Y	Y	Y	Y	Y	Y
Putative RNAP subunit	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
β–β′ RNAP	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Putative RNAP subunit	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
DnaG family primase	Y				Y	Y	Y	Y	Y	Y
Conserved genes in the crAssphage group										
Uncharacterized protein		Y	Y			Y	Y	Y	Y	Y
Uncharacterized protein					Y	Y			Y	Y
Uncharacterized protein		Y	Y					Y		
Siphovirus Gp157 homolog										
Uncharacterized protein										
ssb										
RecT										
SNF2 helicase[b]						Y	Y	Y	Y	Y
PolB	Y				Y	Y			Y	
Uncharacterized protein										
UDG					Y	Y		Y	Y	Y
SF1 helicase										

[a]The core genes include those that are represented in all 5 major branches of the family according to the MCP tree (Fig. 1 and Supplementary figure S1). Members of the crAssphage group are highlighted by asterisks in the table header. [b]The predicted SNF2 family helicase is inactivated in the crAssphage group but active in the other phages. [c]Other crAss-like family members encode distant members of the PD-(D/E)XK nuclease family.

genes encode giant proteins, up to 6,000 aa in size. An HHpred search initiated with the multiple alignment of the homologues of one of these large proteins (crAssphage gene 46 product) identified a small region of similarity with the β′-subunit of the bacterial RNA polymerase (RNAP), which contained the signature catalytic loop with the metal-binding DxDxD motif[23] (Supplementary Fig. 3). Detailed sequence analysis resulted in the identification of two additional conserved motifs typical of the RNAP β-subunit[23], suggesting that the two subunits are fused in this protein (Fig. 4 and Supplementary Fig. 3). Although the similarity between these crAss-like family protein sequences and the large RNAP subunits was limited, the strict conservation of several predicted key motifs that comprise the RNAP catalytic site across the crAss-like family, the fusion of the putative homologues of two RNAP subunits in a single large protein compatible with the combined lengths of the β and β′ RNAP subunits, and the compatibility of the predicted secondary structure elements with the RNAP core structure (Supplementary Fig. 3) strongly suggest that the crAss-like family phages encode an active RNAP.

The putative β–β′ RNAP fusion protein contains another large region of similarity with other phages, which in some of them resides in a separate protein (for example, gene_53 product of Azobacteroides phage ProJPt-Bp1; Fig. 4). Another protein conserved in most of the crAss-like phages (crAssphage gene 47) is typically encoded next to or is fused to the β–β′ RNAP (Fig. 4). Most probably, all three proteins are functionally linked and form a multisubunit RNAP. Although no homologues of the gene 47 product

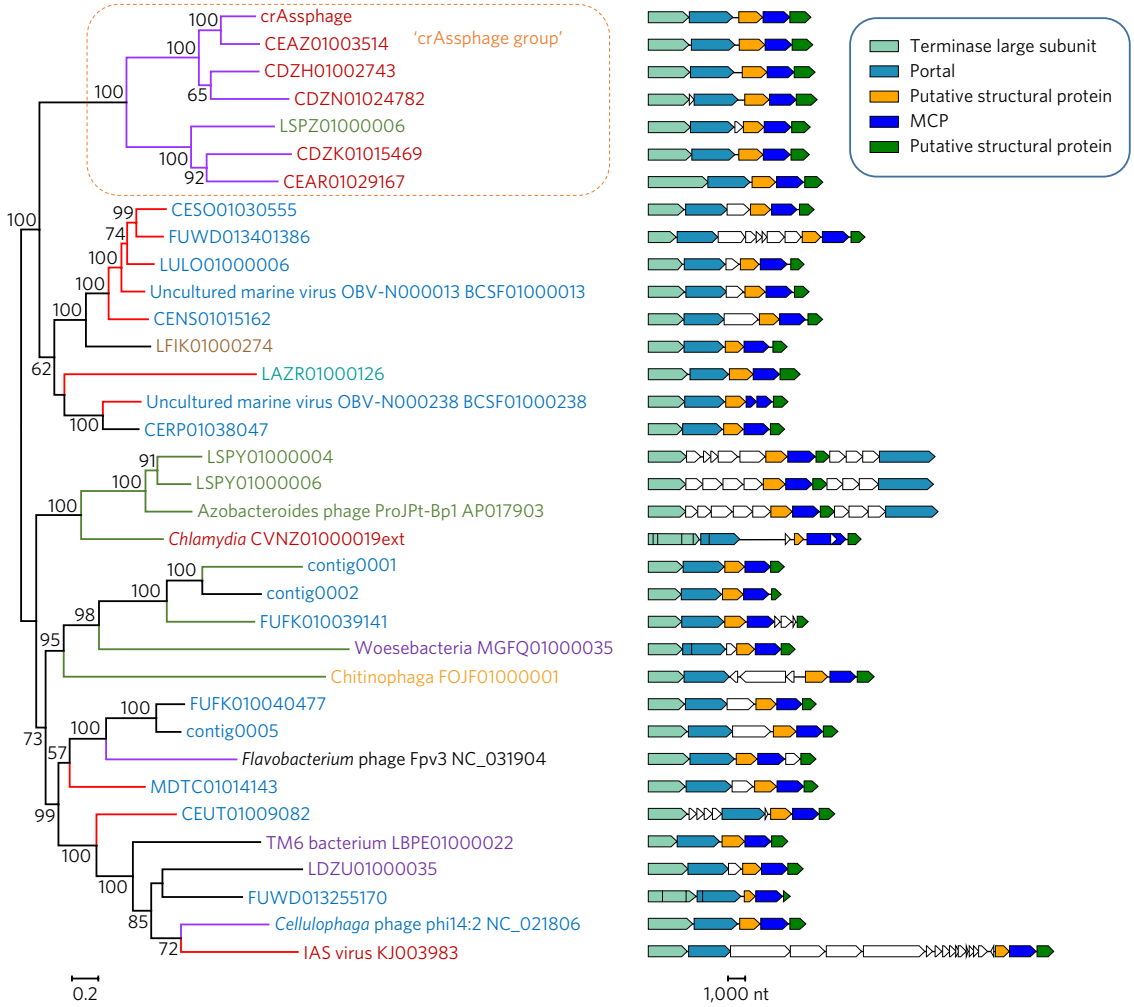

Fig. 1 | Architecture and evolution of the capsid gene module of the crAss-like phage family. The phylogenetic tree was constructed from concatenated multiple alignments of the five proteins of the capsid module. The genomic maps of the capsid gene block are shown for each branch. The five genes of the capsid module are colour-coded, and uncharacterized adjacent genes are shown by empty block arrows. The colours of labels and branches indicate the host or metagenome source: red, human gut or faecal metagenome; green, termite gut metagenome; purple, terrestrial/groundwater; brown, soda lake (hypersaline brine); turquoise, marine sediment; orange, populus root microbiome; black, *Flavobacterium psychrophilum* 950106-1/1 (fish pathogen). Tree branch colours indicate the DNA polymerase family represented in the respective genome (contig): purple, family B DNAP; red, family A DNAP; green, no DNAP; black, unknown (incomplete genomes). Support values were obtained using 100 bootstrap replications; values greater than 50% are shown. The scale bar on the left represents the number of amino acid (aa) substitutions per site, and the scale bar on the right represents the lengths of the genomic segments (number of nucleotides, nt) shown in the figure. No outgroup was included due to the low (or absent) similarity between the crAss-like family protein to homologues from other phages.

were detected, the size and association with the RNAP subunits suggest that this could be a highly diverged α subunit. In most crAss-like phages that encode fused subunits of the predicted RNAP, a putative zincin-like protease domain is encoded in the vicinity, for example, within the gene 45–46 fusion product of the crAssphage, whereas most of the phages that encode the RNAP subunits in separate genes lack the predicted protease (Fig. 4 and Supplementary Fig. 3). Thus, the fused RNAP subunits might be cleaved by the Zn-dependent protease to produce the mature proteins. Fusions of zincin family proteases are typical of different multidomain phage proteins[24], which is compatible with the RNAP cleavage hypothesis. Nonetheless, it cannot be ruled out that the fusion protein is the active form of the phage RNAP. Only a few groups of phages encode their own multisubunit RNAPs, including *Lactococcus* phage 1706, *Rhodococcus* phage ReqiPepy6, *Bacillus* phage SPbeta, *Thermus* phages P74 and P23, and giant phages of *Pseudomonas*[25–28]. In most of these phages, the β and β′ subunits are fused[25,29], whereas in *Pseudomonas* group phages, each subunit is split into two proteins[28]. The phage RNAPs belong to diverged families that can be considered signatures of each respective phage group[26,28]. The crAss-like family RNAPs are even more extremely divergent than those of other phages. To our knowledge, processing of RNAP polyproteins by a dedicated protease so far has not been identified in viruses or cellular organisms. Phage RNAPs transcribe either early (replicative)[28] or late (structural)[27] genes; in the former case, the RNAP presumably is packaged into the virion. We attempted to identify promoters of early and late genes of crAss-like phages by searching the sequences upstream of the genes for potential conserved nucleotide motifs, but no such motifs were detected.

The discovery of crAssphage, the most abundant virus in the human gut virome, appeared particularly striking because the genome was *terra incognita*, with few homologues detected in other viruses or bacteria, and the virion proteins not identified[4]. The present analysis changes this by showing that crAssphage belongs to an expansive phage family that is only distantly related to other known phages and has unusual predicted features, in particular,

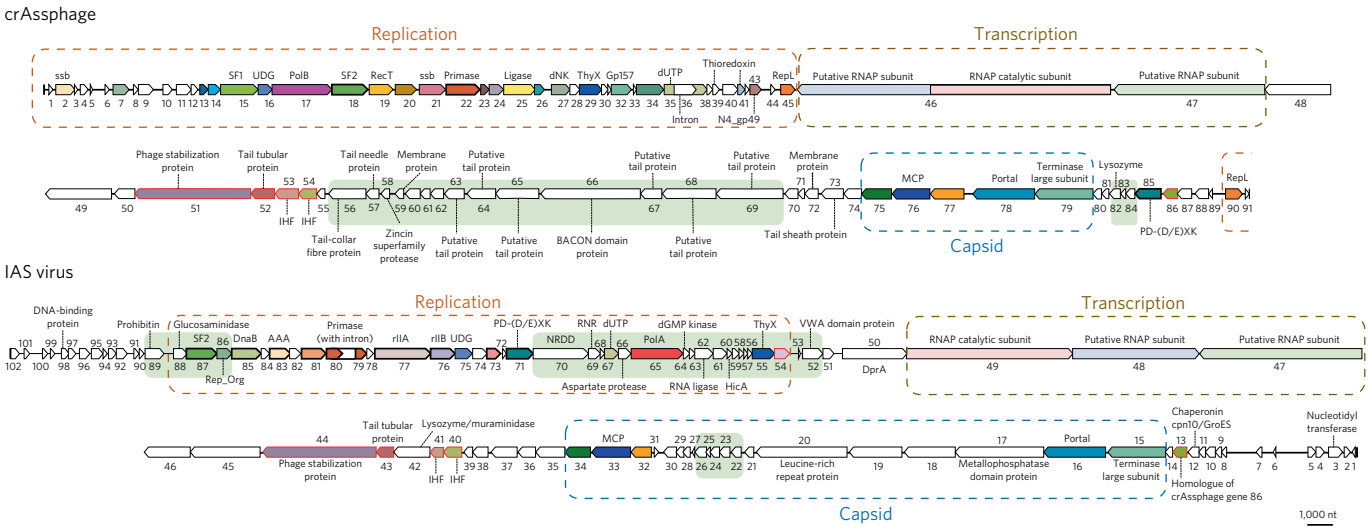

Fig. 2 | Whole-genome maps of crAssphage and IAS virus, the two members of the crAss-like family that are abundant in the human gut virome. Conserved crAss-like family genes are colour-coded. Dashed boxes highlight capsid, tail, replication and transcription gene blocks. Genome regions encoding proteins with strong similarity to Bacteroidetes are shaded in pale green. Gene numbers are according to the crAssphage and IAS virus MetaGeneMark translations. Abbreviations: ssb, single-stranded DNA-binding protein; SF1, SF1 helicase; UDG, uracyl-DNA glycosylase; PolB, DNA polymerase family B; SF2, SNF2-family helicase; RecT, phage RecT recombinase; primase, DnaG family primase; ligase, ATP-dependent DNA ligase; dNK, deoxynucleoside monophosphate kinase; ThyX, flavin-dependent thymidylate synthase; Gp157, Siphovirus Gp157; dUTP, dUTPase; N4_gp49, phage protein of N4_gp49/Sf6_gp66 family; RepL, plasmid replication initiation protein RepL; IHF, integration host factor IHF subunit; PD-(D/E)XK, PD-(D/E)XK family nuclease; Rep_Org, putative replisome organizer protein; DnaB, DnaB replicative DNA helicase; AAA, AAA domain ATPase; rIIA, rIIA-like protector protein; rIIB, rIIB-like protector protein; NRDD, anaerobic ribonucleoside-triphosphate reductase; RNR, anaerobic ribonucleoside-triphosphate reductase activating protein; PolA, DNA polymerase family A; DprA, DNA processing protein DprA. For further details on the annotation, see Supplementary Table 2.

a previously unknown putative mechanism of RNAP maturation via polyprotein processing. The MCP of these phages, a distinct form of the HK97 class of icosahedral capsid proteins[30], is now confidently predicted and amenable for experiments aimed at direct identification and characterization of the phage. Altogether, homologues with characterized functions have been detected for 53% (48 out of 91) of MetaGeneMark-predicted crAssphage proteins (compared to 26% in the original analysis and 14% in the current RefSeq annotation; Supplementary Table 2).

The crAss-like family includes at least one additional phage, IAS, that can reach high abundance in the human gut[15]. Generally, the crAss-like family appears to be abundant and widespread in diverse habitats, both animal-associated and environmental. Various bacteria of the phylum Bacteroidetes appear to be the primary hosts of crAss-like phages, as indicated by the presence of several genes apparently derived from these bacteria, including the DNA primase that is ancestral in the family and the BACON domain protein implicated in phage adhesion to mucus that could increase the frequency of the encounters with the host bacteria[31]. This virus–host association is supported by CRISPR spacer analysis. In addition to the previously reported imperfect matches to Bacteroides genomes[4], we detected perfect matches of crAssphage sequences to two spacers from *Porhyromonas* sp. (Supplementary Note 2). It seems likely that crAssphage has a broad host range among the Bacteroidetes, which could contribute to the (near) ubiquity of this phage in humans. For the IAS virus, spacers with partial matches were detected in CRISPR arrays of *Prevotella* (Supplementary Note 2), again indicating a Bacteroidetes host. Although some of the crAss-like family sequences identified here are assigned to genomes of other bacteria, these assignments could be erroneous (see above), suggesting that the crAss-like virus family is Bacteroidetes-specific. Our results indicate that some of the crAss-like phages are temperate and lysogenize their hosts by integrating into their genomes with the aid of phage-encoded tyrosine integrases. The temperate lifestyle increases the opportunities for recombination with other phages and could account for the presence of regions with high similarity to otherwise unrelated Bacteroidetes prophages.

Our analysis of the predicted tail proteins indicates that the crAss-like phages possess short, podovirus-like tails. Thus, under the classical morphology-guided classification scheme, these phages would be classified into the family Podoviridae. However, given that phage taxonomy is moving towards sequence-based approaches[32], crAss-like phages are likely to become a family within the order Caudovirales. The general lesson from this study is that, with the current proliferation of genomic and especially metagenomic sequence databases and advances in database search approaches, any discovered abundant virus or microbe is likely to become a prototype of a previously undetected, often highly diverse group of organisms. Such advanced analyses can guide the experimental study of viruses and microbes that are currently known only through genomic sequences but could be major players in the microbiota.

Methods

The search for crAssphage structural proteins was performed as follows. The sequences of the crAssphage proteins were first used as queries in a PSI-BLAST search[33] of the NCBI non-redundant (nr) database. Proteins that produced no statistically significant hits to nr proteins with predicted functions were considered candidates for crAssphage structural proteins. For each of these proteins, homologues from both nr and and environmental (env_nr) protein sequence databases were collected (using PSI-BLAST with default parameters until convergence or for 5–6 iterations). The homologues detected in this search were aligned with the query crAssphage protein and the alignments were used as queries for HHpred[34] searches. Three of these queries produced statistically significant hits to phage structural proteins (terminase large subunit, portal and HK97 family MCP).

The sequences of the predicted crAssphage MCP and its homologues ('initial MCP set') were used as queries for translating blast (TBLASTN) searches against the wgs and nr nucleotide databases. Nucleotide sequences of the hits were retrieved and translated in six frames and the sequences of the putative

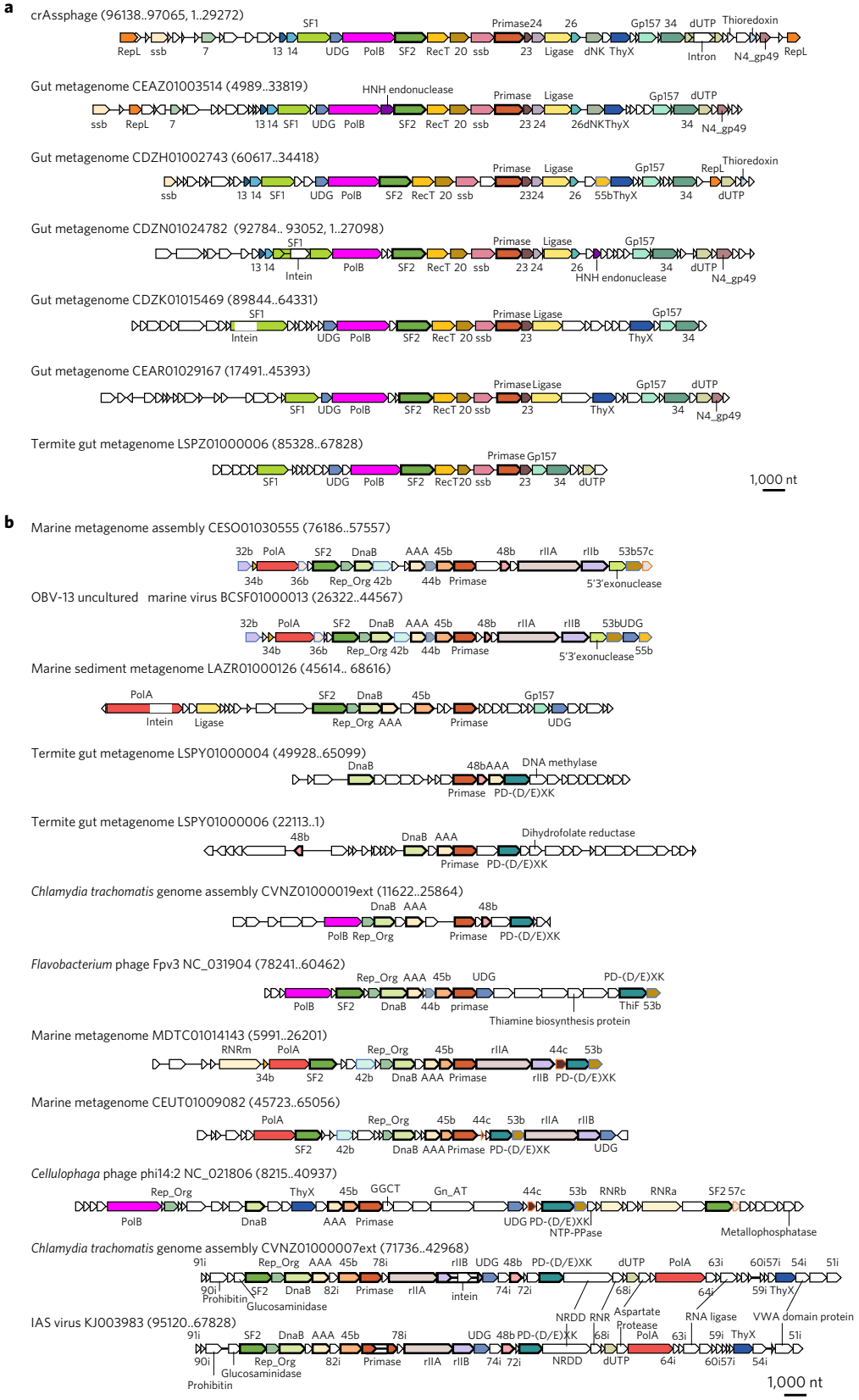

Fig. 3 | Replicative gene module of the crAss-like phage family. a, The crAssphage group. **b**, The rest of the crAss-like family. Homologous genes are marked by the same colours and labels. Genes with no predicted function are numbered according to crAssphage translation, OBV-13 virus translation (suffix 'b'), *Cellulophaga* phage phi14:2 translation (suffix 'c'), or IAS virus translation (suffix 'i'). Abbreviations: RNRm, class II ribonucleotide reductase; RNRa, ribonucleoside reductase alpha chain; RNRb, ribonucleoside reductase beta chain; GGCT, gamma-glutamyl cyclotransferase; Gn_AT, glucosamine-fructose-6-phosphate aminotransferase; NTP-PPase, nucleoside triphosphate pyrophosphohydrolase. The rest of the abbreviations are as in Fig. 2.

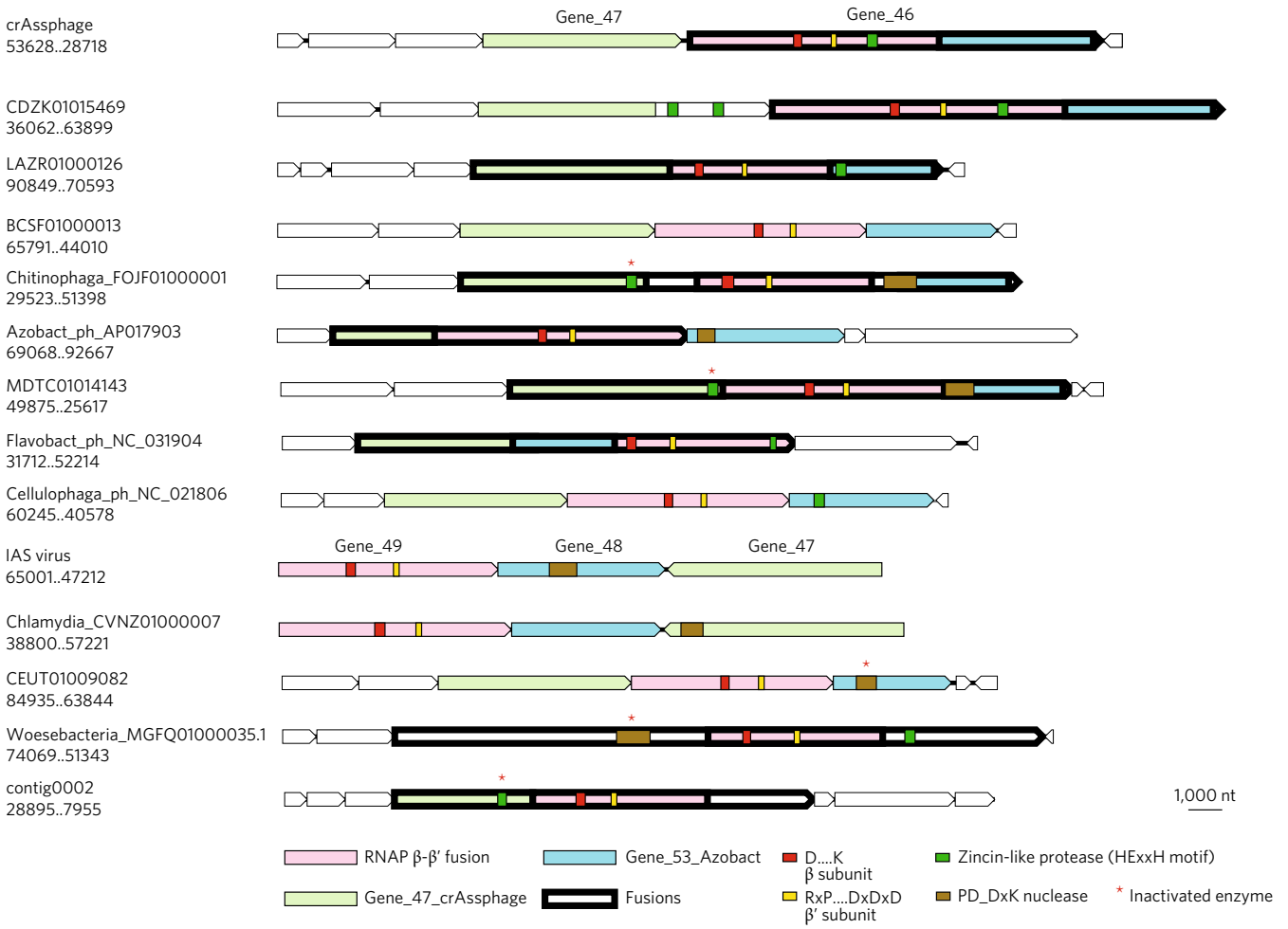

Fig. 4 | Genome expression gene module of the crAss-like phage family. The predicted RNAP subunits as well as the RNAP and protease motifs are colour-coded as shown at the bottom of the figure. The PD-DxK nucleases are most probably encoded in Group I introns.

MCP homologues were validated by comparison to the initial MCP set. The extracted protein sequences of the putative MCP homologues (either complete proteins or fragments longer than 150 aa) were clustered with blastclust at 90% identity, aligned with MUSCLE[35], and used for the phylogenetic reconstruction shown in Supplementary Fig. 1.

In addition, other conserved crAssphage proteins were searched against GenBank databases by protein blast (BLASTP) and TBLASTN[33]. These searches led to the identification of several complete and partial crAss-like viral genomes in the nr database as well as several hundred crAss-like contigs in the wgs databases (Supplementary Fig. 1; see ftp://ftp.ncbi.nih.gov/pub/yutinn/crassphage_2017/ for additional data).

For in-depth analysis, 37 representative genomes (contigs) were selected among the crAss-like family members identified in the nr and wgs databases. In some cases (namely CVNZ01000019ext, CVNZ0100007ext, contig0001, contig0002 and contig0005), representative contigs were obtained by additional assembly using the Geneious software. The original contigs used for these assemblies are listed in Supplementary Table 1.

All representative genomes were translated using MetaGeneMark[36]. The set of open reading frames produced by this translation for crAssphage was virtually identical to the original set that was produced using the Glimmer software[37]; all genes detected by MetaGeneMark, for which a function was predicted, were also represented in the original Glimmer translation[4] (Supplementary Table 2). Homologous protein sequences were aligned using MUSCLE[35]. To identify putative homologues outside the crAss-like family, the alignments of all conserved proteins were used as queries for PSI-BLAST[33] and HHpred[34] searches.

The analysis reported here differed from the original analysis of the crAssphage genome in the following respects: (1) expanded databases of genomic and metagenomics sequences were searched, (2) the updated versions of the PSI-BLAST and HHpred software were used, and (3) multiple alignments of crAssphage proteins and their homologues, rather than individual crAssphage protein sequences, were used to initiate the searches. Furthermore, no preset cutoffs were used in database searches, and all search results and alignments were examined individually for conservation of diagnostic sequence motifs. Together, these amendments to the sequence analysis protocol yielded substantially enhanced search results.

For phylogenetic reconstruction of crAss-like family MCP, PolA, PolB, primase and ligase (Supplementary Figs 1 and 2), gapped columns (more than 30% of gaps) and columns with low information content were removed from the alignments[38], and filtered alignments were used for tree reconstructions using the FastTree program[39]. For the phylogenetic tree shown in Fig. 1, the alignments of five conserved proteins of the capsid module were concatenated and used for phylogenetic analysis with the PhyML program (http://www.atgc-montpellier.fr/phyml-sms/)[40]. The best model identified by PhyML was LG+G+I+F (LG substitution model, gamma-distributed site rates with gamma shape parameter estimated from the alignment; fraction of invariable sites estimated from the alignment; and empirical equilibrium frequencies).

Search for nucleotide sequence motifs was performed using the MEME[41] and Gibbs Centroid Sampler[42] programs.

Life Sciences Reporting Summary. Further information on experimental design is available in the Life Sciences Reporting Summary.

Data availability. All the data used for the analysis reported in this work are publicly available through GenBank. Genbank accession numbers for the representative set of contigs that have been analysed in detail are provided in Supplementary Dataset 1 and the complete set of accession numbers for all crAss-like contigs is available at ftp://ftp.ncbi.nih.gov/pub/yutinn/crassphage_2017/MCP_containing_contig_sources.xlsx. Annotation of the crAssphage and IAS virus genes is provided in Supplementary Dataset 1. Further supplementary information is available in Supplementary Note 1 and at ftp://ftp.ncbi.nih.gov/pub/yutinn/crassphage_2017/.

Received: 30 June 2017; Accepted: 4 October 2017;
Published online: 13 November 2017

References

1. Rohwer, F. Global phage diversity. *Cell* **113**, 141 (2003).
2. Suttle, C. A. Marine viruses—major players in the global ecosystem. *Nat. Rev. Microbiol.* **5**, 801–812 (2007).
3. Simmonds, P. et al. Consensus statement: virus taxonomy in the age of metagenomics. *Nat. Rev. Microbiol.* **15**, 161–168 (2017).
4. Dutilh, B. E. et al. A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. *Nat. Commun.* **5**, 4498 (2014).
5. Dutilh, B. E. Metagenomic ventures into outer sequence space. *Bacteriophage* **4**,e979664 (2014).
6. Ogilvie, L. A. & Jones, B. V. The human gut virome: a multifaceted majority. *Front. Microbiol.* **6**, 918 (2015).
7. Hurwitz, B. L., U'Ren, J. M. & Youens-Clark, K. Computational prospecting the great viral unknown. *FEMS Microbiol. Lett.* **363**, fnw077 (2016).
8. Yarygin, K. et al. Abundance profiling of specific gene groups using precomputed gut metagenomes yields novel biological hypotheses. *PLoS ONE* **12**,e0176154 (2017).
9. Manrique, P. et al. Healthy human gut phageome. *Proc. Natl Acad. Sci. USA* **113**,10400–10405 (2016).
10. Ahlgren, N. A., Ren, J., Lu, Y. Y., Fuhrman, J. A. & Sun, F. Alignment-free d_2^* oligonucleotide frequency dissimilarity measure improves prediction of hosts from metagenomically-derived viral sequences. *Nucleic Acids Res.* **45**, 39–53 (2017).
11. Wexler, A. G. & Goodman, A. L. An insider's perspective: bacteroides as a window into the microbiome. *Nat. Microbiol.* **2**, 17026 (2017).
12. Whitaker, W. R., Shepherd, E. S. & Sonnenburg, J. L. Tunable expression tools enable single-cell strain distinction in the gut microbiome. *Cell* **169**, 538–546 (2017).
13. Pramono, A. K. et al. Discovery and complete genome sequence of a bacteriophage from an obligate intracellular symbiont of a cellulolytic protist in the termite gut. *Microbes Environ.* **32**, 112–117 (2017).
14. Holmfeldt, K. et al. Twelve previously unknown phage genera are ubiquitous in global oceans. *Proc. Natl Acad. Sci. USA* **110**, 12798–12803 (2013).
15. Oude Munnink, B. B. et al. Unexplained diarrhoea in HIV-1 infected individuals. *BMC Infect. Dis.* **14**, 22 (2014).
16. Brown, C. T. et al. Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature* **523**, 208–211 (2015).
17. Burroughs, A. M., Kaur, G., Zhang, D. & Aravind, L. Novel clades of the HU/IHF superfamily point to unexpected roles in the eukaryotic centrosome, chromosome partitioning, and biologic conflicts. *Cell Cycle* **16**, 1093–1103 (2017).
18. Lander, G. C. et al. The P22 tail machine at subnanometer resolution reveals the architecture of an infection conduit. *Structure* **17**, 789–799 (2009).
19. Casjens, S. R. & Molineux, I. J. Short noncontractile tail machines: adsorption and DNA delivery by podoviruses. *Adv. Exp. Med. Biol.* **726**, 143–179 (2012).
20. Bhardwaj, A., Molineux, I. J., Casjens, S. R. & Cingolani, G. Atomic structure of bacteriophage Sf6 tail needle knob. *J. Biol. Chem.* **286**, 30867–30877 (2011).
21. Xiang, Y. et al. Crystal and cryoEM structural studies of a cell wall degrading enzyme in the bacteriophage phi29 tail. *Proc. Natl Acad. Sci. USA* **105**, 9552–9557 (2008).
22. Casjens, S. R. & Thuman-Commike, P. A. Evolution of mosaically related tailed bacteriophage genomes seen through the lens of phage P22 virion assembly. *Virology* **411**, 393–415 (2011).
23. Lane, W. J. & Darst, S. A. Molecular evolution of multisubunit RNA polymerases: sequence analysis. *J. Mol. Biol.* **395**, 671–685 (2010).
24. Iyer, L. M., Burroughs, A. M., Anand, S., de Souza, R. F. & Aravind, L. Polyvalent proteins, a pervasive theme in the intergenomic biological conflicts of bacteriophages and conjugative elements. *J. Bacteriol.* **199**, e00245–17 (2017).
25. Berdygulova, Z. et al. Temporal regulation of gene expression of the *Thermus thermophilus* bacteriophage P23-45. *J. Mol. Biol.* **405**, 125–142 (2011).
26. Iyer, L. M. & Aravind, L. Insights from the architecture of the bacterial transcription apparatus. *J. Struct. Biol.* **179**, 299–319 (2012).
27. Yakunina, M. et al. A non-canonical multisubunit RNA polymerase encoded by a giant bacteriophage. *Nucleic Acids Res.* **43**, 10411–10420 (2015).
28. Lavysh, D. et al. The genome of AR9, a giant transducing *Bacillus* phage encoding two multisubunit RNA polymerases. *Virology* **495**, 185–196 (2016).
29. Ruprich-Robert, G. & Thuriaux, P. Non-canonical DNA transcription enzymes and the conservation of two-barrel RNA polymerases. *Nucleic Acids Res.* **38**, 4559–4569 (2010).
30. Krupovic, M. & Koonin, E. V. Multiple origins of viral capsid proteins from cellular ancestors. *Proc. Natl Acad. Sci. USA* **114**, E2401–E2410 (2017).
31. Barr, J. J. et al. Subdiffusive motion of bacteriophage in mucosal surfaces increases the frequency of bacterial encounters. *Proc. Natl Acad. Sci. USA* **112**,13675–13680 (2015).
32. Krupovic, M. et al. Taxonomy of prokaryotic viruses: update from the ICTV bacterial and archaeal viruses subcommittee. *Arch. Virol.* **161**, 1095–1099 (2016).
33. Altschul, S. F. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
34. Soding, J. Protein homology detection by HMM-HMM comparison. *Bioinformatics* **21**, 951–960 (2005).
35. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
36. Besemer, J. & Borodovsky, M. GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. *Nucleic Acids Res.* **33**, W451–454 (2005).
37. Kelley, D. R., Liu, B., Delcher, A. L., Pop, M. & Salzberg, S. L. Gene prediction with Glimmer for metagenomic sequences augmented by classification and clustering. *Nucleic Acids Res.* **40**, e9 (2012).
38. Yutin, N., Makarova, K. S., Mekhedov, S. L., Wolf, Y. I. & Koonin, E. V. The deep archaeal roots of eukaryotes. *Mol. Biol. Evol.* **25**, 1619–1630 (2008).
39. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5**, e9490 (2010).
40. Guindon, S. et al. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59**,307–321 (2010).
41. Bailey, T. L., Williams, N., Misleh, C. & Li, W. W. MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res.* **34**, W369–373 (2006).
42. Thompson, W. A., Newberg, L. A., Conlan, S., McCue, L. A. & Lawrence, C. E. The Gibbs Centroid Sampler. *Nucleic Acids Res.* **35**, W232–237 (2007).

Acknowledgements

The authors thank Y.I. Wolf and S. Shmakov for technical help and Koonin group members for discussion. N.Y., K.S.M., A.B.G. and E.V.K. are supported by intramural funds of the US Department of Health and Human Services (to the National Library of Medicine).

Author contributions

E.V.K. conceived of the study. N.Y., K.S.M. and M.K. performed research. N.Y., K.S.M., A.B.G., M.K., A.S., R.A.E. and E.V.K. analysed the data. E.V.K. wrote the manuscript, which was read, edited and approved by all authors.

Competing interests

The authors declare no competing financial interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41564-017-0053-y.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to E.V.K.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

nature research

Corresponding author(s): Eugene V Koonin

☐ Initial submission ☐ Revised version ☒ Final submission

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Experimental design

1. Sample size

Describe how sample size was determined.

This is not an experimental study but rather, a computational/bioinformatic one, therefore the concept of sample size is not applicable.

2. Data exclusions

Describe any data exclusions.

No data were excluded from the analysis.

3. Replication

Describe whether the experimental findings were reliably reproduced.

This is not an experimental study but rather, a computational/bioinformatic one, therefore, the concept of replication is not applicable.

4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

This is not an experimental study but rather, a computational/bioinformatic one, therefore, the concept of allocation of samples is not applicable. In all sequence comparisons and phylogenetic analyses, the appropriate randomization was performed.

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

This is not an experimental study but rather, a computational/bioinformatic one, therefore the concept of blinding is not applicable.

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a	Confirmed	
☒	☐	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
☒	☐	A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	A statement indicating how many times each experiment was replicated
☒	☐	The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
☒	☐	A description of any assumptions or corrections, such as an adjustment for multiple comparisons
☒	☐	The test results (e.g. *P* values) given as exact values whenever possible and with confidence intervals noted
☒	☐	A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
☒	☐	Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.	BLAST package (BLASTP, PSI-BLAST, TBLASTN), HHPred, GeneMark, Geneious, FAstTree, PhyML, MEME, Gibbs Centroid Sampler. All software is in the public domain.

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.	All materials used in this study are in the public domain; all datasets can be obtained according to the full data availability statement.

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).	N/A

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.	N/A
b. Describe the method of cell line authentication used.	N/A
c. Report whether the cell lines were tested for mycoplasma contamination.	N/A
d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.	N/A

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.	N/A

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.	N/A