Article

Clades of huge phages from across Earth's ecosystems

https://doi.org/10.1038/s41586-020-2007-4

Received: 22 March 2019

Accepted: 2 January 2020

Published online: 12 February 2020

Open access

Check for updates

Basem Al-Shayeb[1,17], Rohan Sachdeva[1,17], Lin-Xing Chen[1], Fred Ward[1], Patrick Munk[2], Audra Devoto[1], Cindy J. Castelle[1], Matthew R. Olm[1], Keith Bouma-Gregson[3], Yuki Amano[4], Christine He[1], Raphaël Méheust[1], Brandon Brooks[1], Alex Thomas[1], Adi Lavy[1], Paula Matheus-Carnevali[1], Christine Sun[5], Daniela S. A. Goltsman[5], Mikayla A. Borton[6], Allison Sharrar[3], Alexander L. Jaffe[1], Tara C. Nelson[7], Rose Kantor[1], Ray Keren[1], Katherine R. Lane[1], Ibrahim F. Farag[1], Shufei Lei[3], Kari Finstad[8], Ronald Amundson[8], Karthik Anantharaman[3], Jinglie Zhou[9], Alexander J. Probst[1], Mary E. Power[10], Susannah G. Tringe[9], Wen-Jun Li[11], Kelly Wrighton[6], Sue Harrison[12], Michael Morowitz[13], David A. Relman[5], Jennifer A. Doudna[1], Anne-Catherine Lehours[14], Lesley Warren[7], Jamie H. D. Cate[1], Joanne M. Santini[15] & Jillian F. Banfield[1,3,8,16]✉

Bacteriophages typically have small genomes[1] and depend on their bacterial hosts for replication[2]. Here we sequenced DNA from diverse ecosystems and found hundreds of phage genomes with lengths of more than 200 kilobases (kb), including a genome of 735 kb, which is—to our knowledge—the largest phage genome to be described to date. Thirty-five genomes were manually curated to completion (circular and no gaps). Expanded genetic repertoires include diverse and previously undescribed CRISPR–Cas systems, transfer RNAs (tRNAs), tRNA synthetases, tRNA-modification enzymes, translation-initiation and elongation factors, and ribosomal proteins. The CRISPR–Cas systems of phages have the capacity to silence host transcription factors and translational genes, potentially as part of a larger interaction network that intercepts translation to redirect biosynthesis to phage-encoded functions. In addition, some phages may repurpose bacterial CRISPR–Cas systems to eliminate competing phages. We phylogenetically define the major clades of huge phages from human and other animal microbiomes, as well as from oceans, lakes, sediments, soils and the built environment. We conclude that the large gene inventories of huge phages reflect a conserved biological strategy, and that the phages are distributed across a broad bacterial host range and across Earth's ecosystems.

Phages—viruses that infect bacteria—are considered distinct from cellular life owing to their inability to carry out most biological processes required for reproduction. They are agents of ecosystem change because they prey on specific bacterial populations, mediate lateral gene transfer, alter host metabolism and redistribute bacterially derived compounds through cell lysis[2–4]. They spread antibiotic resistance[5] and disperse pathogenicity factors that cause disease in humans and animals[6,7]. Most knowledge about phages is based on laboratory-studied examples, the vast majority of which have genomes that are a few tens of kb in length. Widely used isolation-based methods select against large phage particles, and they can be excluded from phage concentrates obtained by passage through 100-nm or 200-nm filters[1]. In 2017, only 93 isolated phages with genomes that were more than 200 kb in length were published[1]. Sequencing of whole-community DNA can uncover phage-derived fragments; however, large genomes can still escape detection owing to fragmentation[8]. A new clade of human- and animal-associated megaphages was recently described on the basis of genomes that were manually curated to completion from metagenomic datasets[9]. This finding prompted us to carry out a more-comprehensive analysis of microbial communities to evaluate the prevalence, diversity and ecosystem distribution of phages with large genomes. Previously, phages with genomes of more than 200 kb have been referred to as 'jumbophages'[1] or, in the case of phages with genomes of more than 500 kb, as megaphages[9]. As the set reconstructed here span both size ranges we refer to them simply as 'huge phages'. A graphical abstract provides an overview of our approach and main

[1]Innovative Genomics Institute, University of California Berkeley, Berkeley, CA, USA. [2]National Food Institute, Technical University of Denmark, Kongens Lyngby, Denmark. [3]Earth and Planetary Science, University of California Berkeley, Berkeley, CA, USA. [4]Nuclear Fuel Cycle Engineering Laboratories, Japan Atomic Energy Agency, Tokai-mura, Japan. [5]Department of Microbiology & Immunology, Stanford University, Stanford, CA, USA. [6]Department of Soil and Crop Sciences, Colorado State University, Fort Collins, CO, USA. [7]Department of Civil and Mineral Engineering, University of Toronto, Toronto, Ontario, Canada. [8]Environmental Science, Policy and Management, University of California Berkeley, Berkeley, CA, USA. [9]DOE Joint Genome Institute, Berkeley, CA, USA. [10]Integrative Biology, University of California Berkeley, Berkeley, CA, USA. [11]School of Life Sciences, Sun Yat-Sen University, Guangzhou, China. [12]Centre for Bioprocess Engineering Research, University of Cape Town, Cape Town, South Africa. [13]Department of Surgery, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA. [14]Laboratoire Microorganismes: Génome et Environnement, Université Clermont Auvergne, CNRS, Clermont-Ferrand, France. [15]Institute of Structural and Molecular Biology, University College London, London, UK. [16]School of Earth Sciences, University of Melbourne, Melbourne, Victoria, Australia. [17]These authors contributed equally: Basem Al-Shayeb, Rohan Sachdeva. ✉e-mail: jbanfield@berkeley.edu

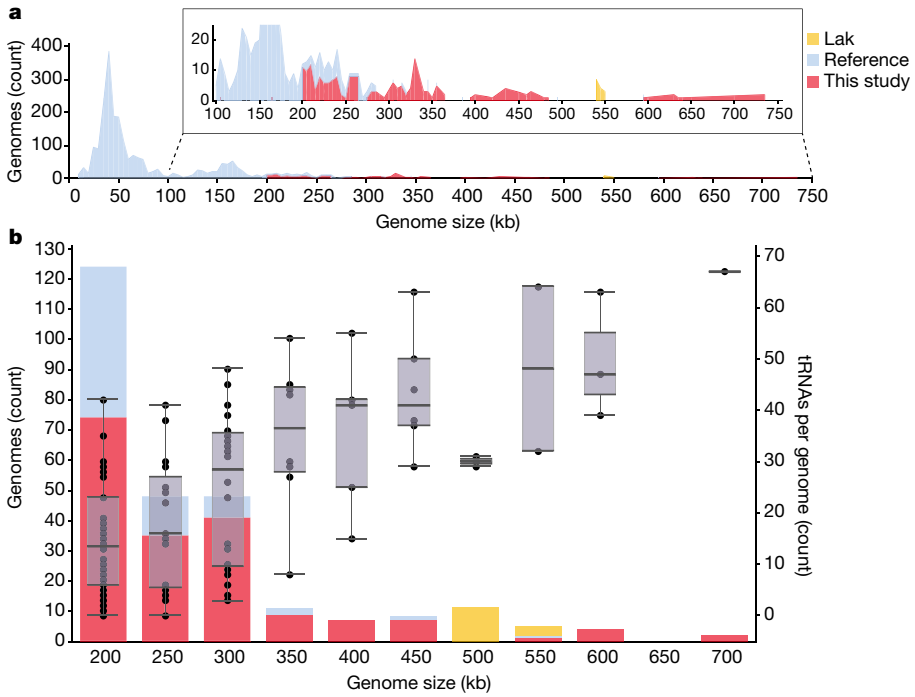


Fig. 1 | Distribution of the genome sizes and tRNAs of phages. a, Size distribution of circularized bacteriophage genomes from this study, Lak megaphage genomes reported recently for a subset of the same samples[9] and reference sources. Reference genomes were collected from all complete RefSeq r92 dsDNA genomes and non-artefactual assemblies with lengths of more than 200 kb from a previous study[14]. **b**, Histogram of the genome size distribution of phages with genomes of more than 200 kb from this study, Lak and reference genomes. Box-and-whisker plot of tRNA counts per genome from this study and Lak phages as a function of genome size (Spearman's $\rho = 0.61$, $P = 4.5 \times 10^{-22}$, $n = 201$ individual phage genomes). The middle line for each box marks the median tRNA count for each size bin, the box marks the interquartile range, and the whiskers represent the maximum and minimum.

findings (Extended Data Fig. 1). This study expands our understanding of phage biodiversity and reveals the wide variety of ecosystems in which phages have genomes with sizes that rival those of small-celled bacteria[10–12]. We postulate that these phages have evolved a distinct 'life' strategy that involves extensive interception and augmentation of host biology while they replicate their huge genomes.

Ecosystem sampling

Metagenomic datasets were acquired from human faecal and oral samples, faecal samples from other animals, freshwater lakes and rivers, marine ecosystems, sediments, hot springs, soils, deep subsurface habitats and the built environment (Extended Data Fig. 2). Genome sequences that were clearly not bacterial, archaeal, archaeal virus, eukaryotic or eukaryotic virus were classified as phage, plasmid-like or mobile genetic elements of uncertain nature on the basis of their gene inventories (Supplementary Information). De novo assembled fragments close to or more than 200 kb in length were tested for circularization and a subset was selected for manual verification and curation to completion (Methods).

Genome sizes and basic features

We reconstructed 351 phage sequences, 6 plasmid-like sequences and 4 sequences of unknown classification (Extended Data Fig. 2). We excluded additional sequences that were inferred to be plasmids (Methods), retaining only those that encoded CRISPR–Cas loci. We included 3 phage sequences of ≤200 kb in length owing to the presence of CRISPR–Cas loci. Consistent with the classification as phages, we identified a wide variety of phage-relevant genes, including those involved in lysis and encoding structural proteins, and documented other expected genomic features of phages (Supplementary Information). Some predicted proteins were large, up to 7,694 amino acids in length; some were tentatively annotated as structural proteins. In total, 175 phage sequences were circularized and 35 were manually curated to completion, in some cases by resolving complex repeat regions, revealing their encoded proteins (Methods and Supplementary Table 1). The remaining genomes are probably incomplete, although some may be complete, but linear. Approximately 30% of genomes show clear GC skew indicative of bidirectional replication and 30% have patterns indicative of unidirectional replication[13] (Extended Data Fig. 3 and Supplementary Information).

Our 4 largest complete, manually curated and circularized phage genomes are 634, 636, 642 and 735 kb in length and are—to our knowledge—the largest phage genomes reported to date. The largest previously reported circularized phage genome was 596 kb in length[14]. The same previous study also reported a circularized genome of 630 kb in length; however, this is an assembly artefact (Supplementary Information). The problem of concatenation artefacts was sufficiently prominent in IMG/VR[15] that we did not include these data in further analyses. We used both complete and circularized genomes from our study and published phage genomes to produce an updated view of the distribution of phage genome sizes (Methods). Without the huge phages reported here, the median genome size for complete phages is around 52 kb (Fig. 1a). Thus, the sequences reported here substantially expand the inventory of phages with unusually large genomes (Fig. 1b).

Some of our reported genomes have a very low coding density (9 genomes have densities of less than 78%) (Supplementary Information), probably owing to the use of a genetic code that is different from the standard code (Methods). This phenomenon has been rarely noted in phages, but has previously been reported for the Lak phage[9] and in a previous study[16]. In the current study, some genomes (mostly those that are associated with humans and/or animals) appear to have reassigned

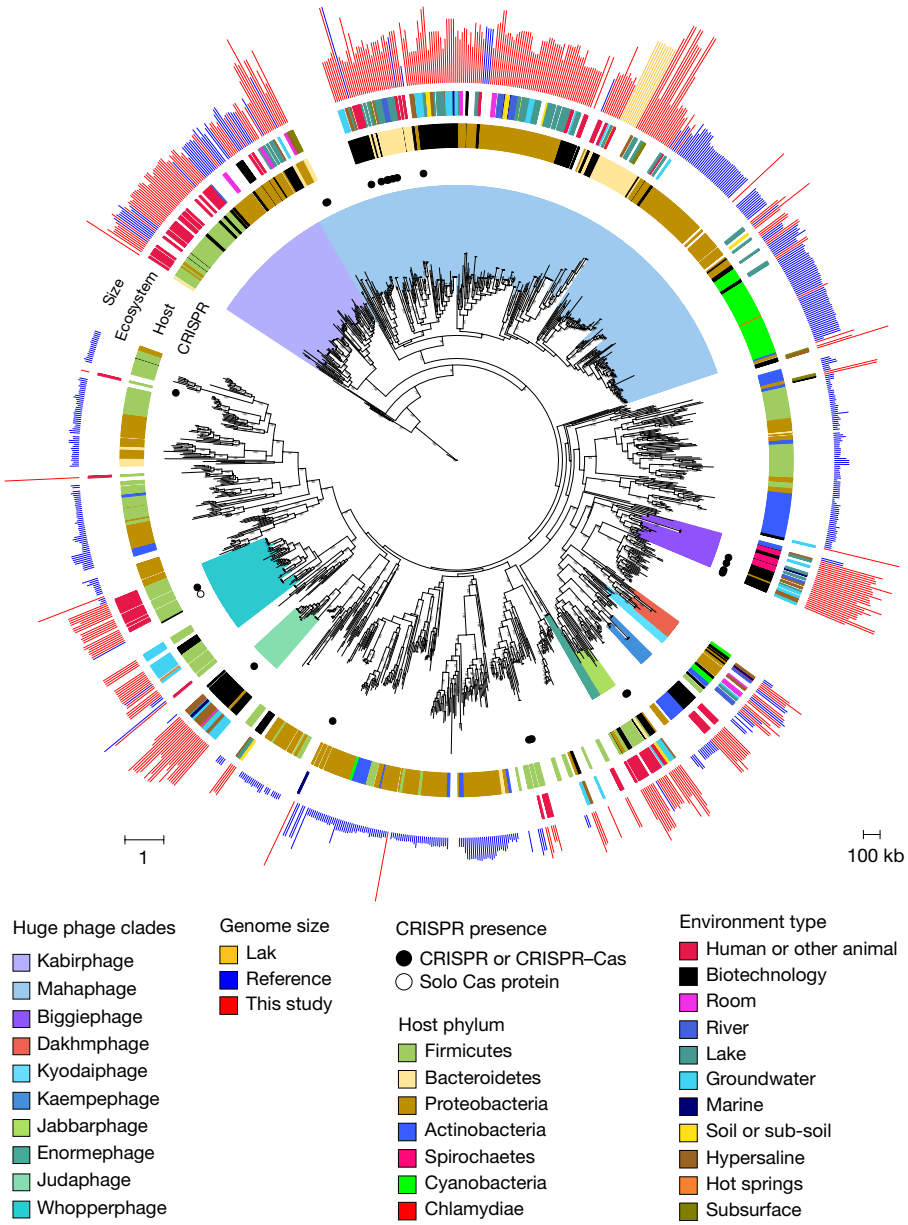


Fig. 2 | Phylogenetic reconstruction of the evolutionary history of huge phages. The phylogeny of phages was reconstructed using large terminase sequences from this study ($n = 397$) and similar matches from all RefSeq r92 proteins ($n = 532$). The tree also includes large terminase sequences from complete RefSeq phage, the Lak megaphage clade[9] ($n = 9$) and non-artefactual phage genomes that are more than 200 kb, from a previous study[14]. Huge phage clades identified in this study were independently corroborated with a phylogenetic reconstruction of major capsid protein (MCP) genes (Extended Data Fig. 5a) and protein clustering (Extended Data Fig. 5b). The tree was rooted using eukaryotic herpesvirus terminases ($n = 7$). The inner to outer rings display the presence of CRISPR–Cas in this study, host phylum, environmental sampling type and genome size. Host phylum and genome size were not included for RefSeq protein database matches for which the sequence may be from an integrated prophage or part of organismal genome projects. Scale bars show the number of substitutions per site (left) and number of base pairs (right).

the UAG (amber) stop codon to encode an amino acid (Extended Data Fig. 4 and Supplementary Information).

In only one case, we identified a sequence of more than 200 kb that was classified as a prophage on the basis of the transition into a flanking bacterial genome sequence. However, around half of the genomes were not circularized, so their potential integration as a prophage cannot be ruled out. The presence of integrases in some genomes is suggestive of a temperate lifestyle under some conditions.

Hosts, diversity and distribution

An intriguing question relates to the evolutionary history of phages with huge genomes; namely, whether they are the result of recent genome expansion within clades of normal-sized phages or whether a large inventory of genes is an established, persistent strategy. To investigate this, we constructed phylogenetic trees for large terminase subunit proteins (Fig. 2) and major capsid proteins (Extended Data Fig. 5a) using sequences from public databases as a context (Methods). Many of the sequences from our phage genomes cluster together with high bootstrap support, thus defining clades. Analysis of the genome size information for database sequences shows that the public sequences that fall into these clades are from phages with genomes of at least 120 kb in length. The largest clade, referred to here as Mahaphage (Maha being Sanskrit for huge), includes all of our largest genomes as well as the 540–552 kb Lak genomes from human and animal microbiomes[9]. We identified nine other clusters of large phages, and refer to them

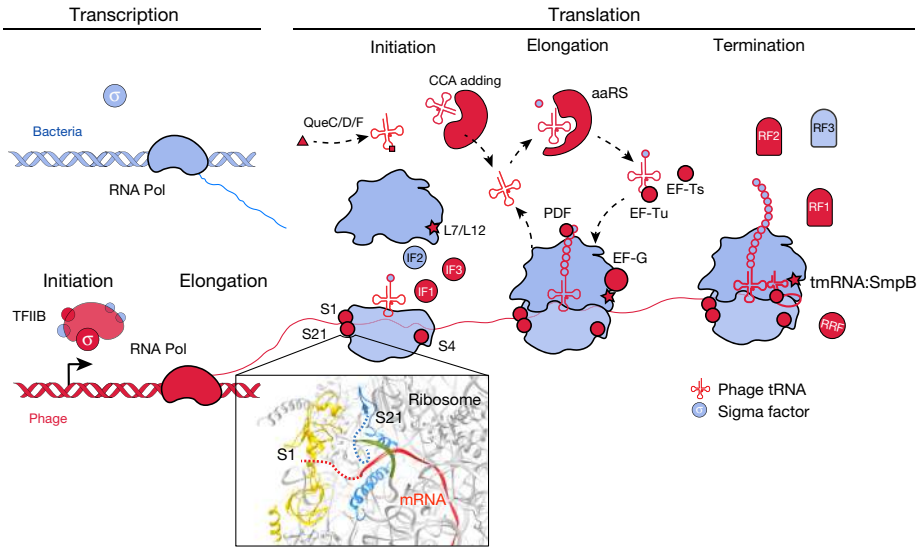


Fig. 3 | A model for phage interception and redirection of host translational systems. Potential mechanisms for how phage-encoded capacities could function to redirect the translational system of the host to produce phage proteins (bacterial components in blue, phage proteins in red). No huge phage encodes all translation-related genes, but many have tRNAs and tRNA synthetases (Supplementary Table 6). Phage proteins with up to six ribosomal protein S1 domains occur in a few genomes. The S1 binds to mRNA to bring it into the site on the ribosome where it is decoded[39]. Phage ribosomal protein S21 might promote translation initiation of phage mRNAs, and many sequences have N-terminal extensions that may be involved in binding RNA (dashed blue line in ribosome insert (RCSB Protein Data Bank (PDB) code: 6BU8[40])), analysed with UCSF Chimera[41]. Many other proteins of the translational apparatus that belong to all steps of the translation cycle are encoded by huge phages. aaRS, aminoacyl-tRNA synthetase; CCA-adding, tRNA nucleotidyltransferase; EF, elongation factor; IF, initiation factor; PDF, peptide deformylase; QueC/D/F, queuosine synthesis and tRNA modification; RF, release factor; RNA Pol, RNA polymerase; RRF, ribosome recycling factor; TFIIB, transcription factor II B.

using the words for 'huge' in the languages of some authors of this paper. We acknowledge that the detailed tree topologies for different genes and datasets vary slightly; however, the clustering is broadly supported by protein family and capsid analyses (Extended Data Fig. 5a, b). The fact that large phages are consistently grouped together into clades establishes that a large genome size is a relatively stable trait. Within each clade, phages were sampled from a wide variety of environment types (Fig. 2), indicating the diversification of these huge phages and their hosts across ecosystems. We also examined the environmental distribution of phages that are so closely related that their genomes can be aligned and we found 20 cases in which the phages occur in at least 2 distinct cohorts or habitat types (Supplementary Table 2).

To determine the extent to which bacterial host phylogeny correlates with phage clades, we identified some phage hosts using CRISPR spacer targeting from bacteria in the same or related samples and phylogenies of normally host-associated phage genes (see below, Supplementary Table 3). We also tested the predictive value of bacterial taxonomic affiliations of the phage gene inventories (Methods) and found that in every case, CRISPR spacer targeting and phylogeny agreed with phylum-level taxonomic profiles. We therefore used taxonomic profiles to predict the bacterial host phylum for many phages (Supplementary Table 4). The results establish the importance of Firmicutes and Proteobacteria as hosts (Extended Data Fig. 2) ($P = 2.5 \times 10^{-5}$, $n = 74$, $W = 606$; one-sided Wilcoxon signed-rank test). The higher prevalence of Firmicutes-infecting huge phages in the human and animal gut compared with other environments reflects the potential host compositions of the microbiomes ($P = 9.3 \times 10^{-7}$, $n = 37$, $U = 238$; one-sided Mann–Whitney U-test). Notably, the 5 genomes that were more than 634 kb in length were all from phages that were predicted to replicate in Bacteroidetes, as do Lak phages[9], and all cluster within the Mahaphage clade. Overall, phages that grouped together phylogenetically are predicted to replicate in bacteria of the same phylum (Fig. 2).

Metabolism, transcription and translation

The phage genomes encode proteins that are predicted to localize to the bacterial membrane or cell surface. These may affect the susceptibility of the host to infection by other phages (Supplementary Table 5 and Supplementary Information). We identified almost all of the previously reported categories of genes that have been suggested to augment host metabolism (Supplementary Information). Many phages have genes involved in the de novo biosynthesis of purines and pyrimidines, and the interconversion of nucleic and ribonucleic acids and nucleotide phosphorylation states. These gene sets are intriguingly similar to those of bacteria with very small cells and putative symbiotic lifestyles[10] (Supplementary Table 5).

Notably, many phages have genes with predicted functions in transcription and translation (Supplementary Table 6). Complete phage genomes encode up to 67 tRNAs, with sequences that are distinct from those of their hosts (Supplementary Table 7). Generally, the number of tRNAs per genome increases with genome length (Fig. 1) (Spearman's $\rho = 0.61$, $P = 4.5 \times 10^{-22}$, $n = 201$). Huge phages have up to 15 tRNA synthetases per genome (Supplementary Table 7), which are also distinct from but related to those of their hosts (Extended Data Fig. 7a and Supplementary Information). Phages may use these proteins to charge their own tRNA variants with host-derived amino acids. A subset of genomes has genes for tRNA modification and ligation of tRNAs cleaved by host defenses.

Many phages carry genes that are implicated in the interception and redirection of host translation. These genes include the initiation factors IF1 and IF3, as well as ribosomal proteins S4, S1, S21 and L7/L12 (ribosomal proteins were only recently reported in phages[17] (Fig. 3)). Both rpS1 and rpS21 are important for translation initiation in bacteria[18–20], making them likely to be useful for the hijacking of host ribosomes. Further analysis of rpS21 proteins revealed N-terminal extensions that were rich in basic and aromatic residues important for RNA binding. We predict that these phage ribosomal proteins substitute for host proteins[17], and their extensions assist in competitive ribosome binding or preferential initiation of phage mRNAs.

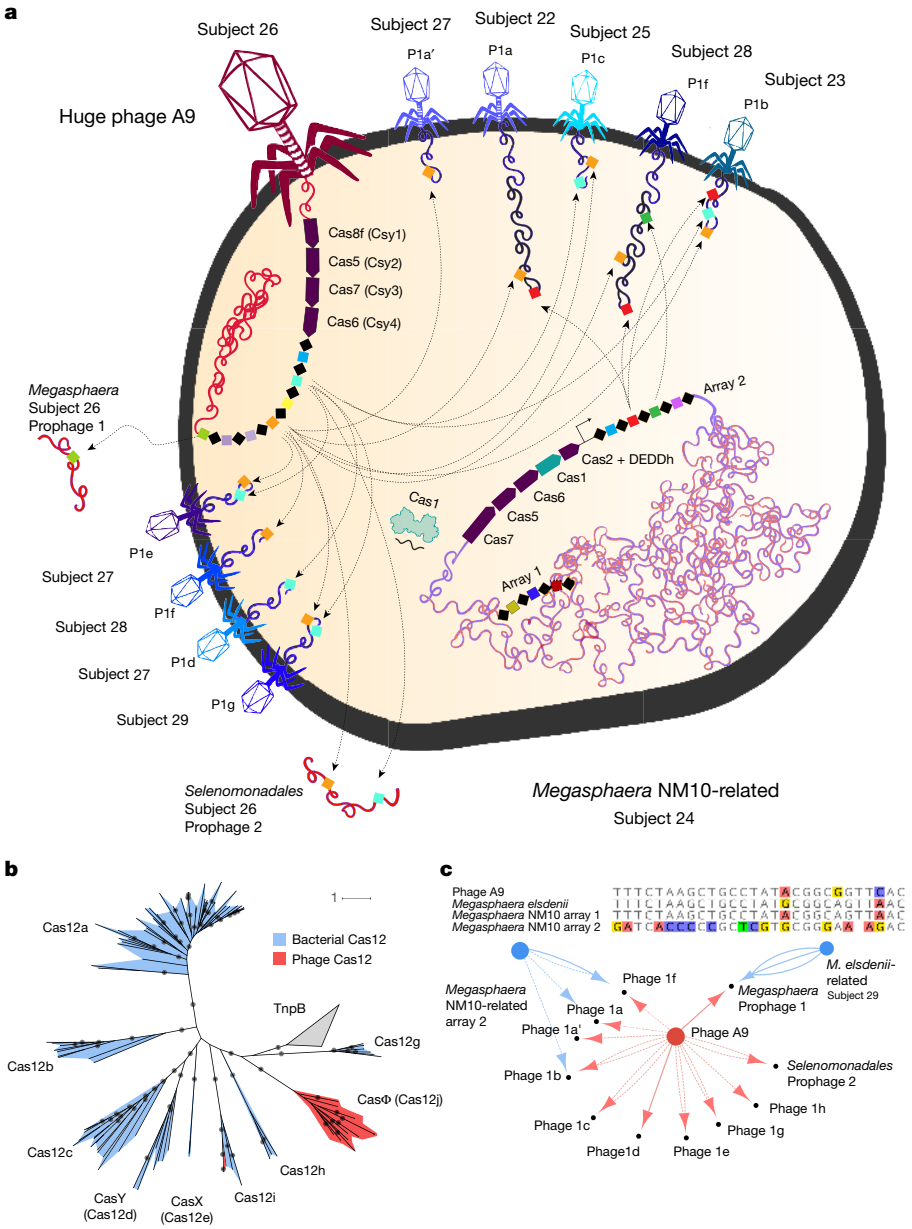


Fig. 4 | Phage and bacterial CRISPR-interaction dynamics. a, Cell diagram of bacterium–phage and phage–phage interactions that involve CRISPR targeting during superinfection. Arrows indicate CRISPR–Cas targeting of the prophage and phage genomes. Phage names indicate related groups delineated by whole-genome alignment. We only included CRISPR interactions from samples of subjects of the same human cohort. **b**, Maximum likelihood phylogenetic tree of Cas12 subtypes a–i. Phage-encoded Cas12i and CasΦ, the new effector, are outlined in red, with bacteria-encoded proteins in blue. Bootstrap values >90 are shown on the branches (circles). Cas14 and type V-U trees are provided separately (Supplementary Fig. 11). Scale bars indicates the number of substitutions per site. **c**, Top, alignment of the consensus repeats from the A9 phage array and predicted host bacterial arrays. Bottom, interaction network showing the targeting of bacteria-encoded (blue) and phage-encoded (red) CRISPR spacers. The number of edges indicate the number of spacers from the array with targets to the smaller node. Solid edges denote spacer targets with no or one mismatch, and dashed edges denote two to three mismatches (to account for degeneration in old-end phage spacers, diversity in different subjects or phage mutation to avoid targeting).

Because rpS1 is often studied in the context of Shine–Dalgarno sequence recognition by the ribosome[19,20], we predicted the ribosomal binding sites for each phage genome (Methods). Whereas most phages have canonical Shine–Dalgarno sequences, huge phages from this study that carry possible rpS1s rarely have identifiable Shine–Dalgarno sequences (Supplementary Information and Supplementary Table 8). It is difficult to confirm 'true' rpS1 proteins owing to the ubiquity of the S1 domain, but this correlation with non-canonical Shine–Dalgarno sequences suggests a role in translation initiation, either on or off the ribosome.

Although assuming control of initiation may be the most logical step for the redirection of host translation by the phage, improving the efficiency of elongation and termination is necessary for robust infection and replication. Accordingly, we found many genes associated with the later steps of translation in phage genomes. These include elongation factors G, Tu and Ts, rpL7/12 and the processing enzyme peptide deformylase (Fig. 3), which has previously been reported in phage genomes[21]. We hypothesize that phage-encoded elongation factors maintain the overall translation efficiency during infection, much like the previously predicted role of peptide deformylase in

sustaining translation of the necessary photosynthetic proteins of the host[21]. Translation termination factors are also represented in our huge phage genomes, including release factor 1 and 2, ribosome recycling factor, as well as transfer messenger RNAs (tmRNAs) and small protein B (SmpB), which rescue ribosomes stalled on damaged transcripts and trigger the degradation of aberrant proteins. These tmRNAs are also used by phages to sense the physiological state of host cells and can induce lysis when the number of stalled ribosomes in the host is high[22]. Notably, some large putative plasmids have analogous suites of translationrelevant genes (Supplementary Table 5).

CRISPR–Cas-mediated interactions

We identified most major types of CRISPR–Cas systems in phages, including Cas9-based type II, the recently described type V-I[23], new variants of the type V-U systems[24] and new subtypes of the type V-F system[25] (Extended Data Fig. 8). The class II systems (types II and V) have not previously been reported in phages. Most phage effector nucleases (for interference) have conserved catalytic residues, implying that they are functional.

In contrast to the well-described case of a phage with a CRISPR system[26], almost all phage CRISPR systems lack spacer acquisition machinery (Cas1, Cas2 and Cas4) and many lack recognizable genes for interference (Extended Data Fig. 9 and Supplementary Table 1). For example, two related phages have a type I-C variant system that lacks Cas1 and Cas2 and have a helicase protein instead of Cas3. These phages also have a second system that contains a new candidate type V effector protein, CasΦ (Cas12j), which is approximately 750 amino acids in length (Fig. 4 and Supplementary Table 1), which occurs proximal to CRISPR arrays.

In some cases, phages that lack genes for interference and spacer integration have similar CRISPR repeats as their hosts (Fig. 4c) and may therefore use the Cas proteins of the host. Alternatively, systems that lack an effector nuclease may repress the transcription of the target sequences without cleavage[27,28]. Additionally, spacer-repeat guide RNAs may have an RNA-interference-like mechanism to silence host CRISPR systems or nucleic acids to which they can hybridize. The phage-encoded CRISPR arrays are often compact (median, six repeats per array) (Extended Data Fig. 10). This range is substantially smaller than typically found in prokaryotic genomes (mean of 41 repeats for class I systems)[29]. Some phage spacers target core structural and regulatory genes of other phages (Fig. 4c and Supplementary Table 10). Thus, phages apparently augment the immune arsenal of their hosts to prevent infection by competing phages.

Some phage-encoded CRISPR loci have spacers that target bacteria in the same sample or in a sample from the same study. We suppose that the targeted bacteria are the hosts for these phages, an inference supported by other host prediction analyses (Supplementary Table 4). Some loci with bacterial chromosome-targeting spacers encode Cas proteins that could cleave the host chromosome, whereas others do not. The targeting of host genes could disable or alter their regulation, which may be advantageous during the phage infection cycle. Some phage CRISPR spacers target bacterial intergenic regions, possibly interfering with genome regulation by blocking promoters or silencing non-coding RNAs.

Notable examples of CRISPR targeting of bacterial chromosomes involve transcription and translation genes. For instance, one phage targets a σ^{70} transcription factor gene in the genome of its host and encodes its own σ^{70} (Supplementary Information). Some huge phage genomes encode anti-sigma factor-like proteins (AsiA), consistent with previous reports of σ^{70} hijacking by phages with AsiA[30]. In another example, a phage spacer targets the host glycyl tRNA synthetase, but the Cas14 effector lacks one of the required catalytic residues for cleavage, suggesting a role in repression (as a 'dCas14'), rather than in cleavage (Supplementary Information).

Notably, we found no evidence of host-encoded spacers that target any CRISPR-bearing phages. However, phage CRISPR targeting of other phages that are also targeted by bacterial CRISPR (Fig. 4c) suggested phage–host associations that were broadly confirmed by the phage taxonomic profile (Supplementary Table 4).

Some large *Pseudomonas*-infecting phages encode anti-CRISPRs[31,32] (Acrs) and proteins that assemble a nucleus-like compartment that segregates their replicating genomes from host-defence and other bacterial systems[33]. We identified proteins encoded in huge phage genomes that cluster with AcrVA5, AcrVA2, AcrIIA7 and AcrIIA11 and may function as Acrs. We also identified tubulin homologues (PhuZ) and proteins (Supplementary Information) that create a proteinaceous phage 'nucleus'[34]. The phage nucleus was recently shown to protect the phage genome against host defence by physically blocking degradation by CRISPR–Cas systems[35].

Conclusions

We show that phages with huge genomes are widespread across Earth's ecosystems. We manually completed 35 genomes, distinguishing them from prophages, providing accurate genome lengths and complete inventories of genes, including those encoded in complex repeat regions that break automated assemblies. Even closely related phages have diversified across habitats. Host and phage migration could transfer genes relevant to medicine and agriculture (for example, genes that affect pathogenicity and antibiotic resistance) (Supplementary Information). Additional mechanisms that are relevant to medical applications involve the direct or indirect activation of immune responses. For example, some phages directly stimulate IFNγ through a TLR9-dependent pathway and exacerbate colitis[36]. Huge phages may represent a reservoir of novel nucleic acid manipulation tools with applications in genome editing and might be harnessed to improve human and animal health. For instance, huge phages equipped with CRISPR–Cas systems might be tamed and used to modulate the functions of the bacterial microbiome or eliminate unwanted bacteria.

The huge phages comprise extensive clades, suggesting that a gene inventory comparable in size to those of many symbiotic bacteria is a conserved strategy for phage survival. Overall, their genes appear to redirect the protein production capacity of the host to favour phage genes by first intercepting the earliest steps of translation and subsequently ensuring the efficient production of proteins. These inferences are aligned with findings for some eukaryotic viruses, which control every phase of protein synthesis[37]. Some phages acquired CRISPR–Cas systems with unusual compositions that may function to control host genes and eliminate competing phages.

More broadly, huge phages represent little-known biology, the platforms for which are distinct from those of small phages and partially analogous to those of symbiotic bacteria, blurring the distinctions between life and non-life. Given phylogenetic evidence for large radiations of huge phages, we wonder whether they are ancient and arose simultaneously with free-living cells, their symbionts and other phages from a pre-life (protogenote) state[38] rather than appearing more recently through episodes of genome expansion.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2007-4.

1. Yuan, Y. & Gao, M. Jumbo bacteriophages: an overview. *Front. Microbiol.* **8**, 403 (2017).
2. Breitbart, M., Bonnain, C., Malki, K. & Sawaya, N. A. Phage puppet masters of the marine microbial realm. *Nat. Microbiol.* **3**, 754–766 (2018).
3. Rascovan, N., Duraisamy, R. & Desnues, C. Metagenomics and the human virome in asymptomatic individuals. *Annu. Rev. Microbiol.* **70**, 125–141 (2016).
4. Emerson, J. B. et al. Host-linked soil viral ecology along a permafrost thaw gradient. *Nat. Microbiol.* **3**, 870–880 (2018).

5. Balcazar, J. L. Bacteriophages as vehicles for antibiotic resistance genes in the environment. *PLoS Pathog.* **10**, e1004219 (2014).
6. Penadés, J. R., Chen, J., Quiles-Puchalt, N., Carpena, N. & Novick, R. P. Bacteriophage-mediated spread of bacterial virulence genes. *Curr. Opin. Microbiol.* **23**, 171–178 (2015).
7. Brown-Jaque, M. et al. Detection of bacteriophage particles containing antibiotic resistance genes in the sputum of cystic fibrosis patients. *Front. Microbiol.* **9**, 856 (2018).
8. Shkoporov, A. N. & Hill, C. Bacteriophages of the human gut: the "known unknown" of the microbiome. *Cell Host Microbe* **25**, 195–209 (2019).
9. Devoto, A. E. et al. Megaphages infect *Prevotella* and variants are widespread in gut microbiomes. *Nat. Microbiol.* **4**, 693–700 (2019).
10. Castelle, C. J. et al. Biosynthetic capacity, metabolic variety and unusual biology in the CPR and DPANN radiations. *Nat. Rev. Microbiol.* **16**, 629–645 (2018).
11. Pérez-Brocal, V. et al. A small microbial genome: the end of a long symbiotic relationship? *Science* **314**, 312–313 (2006).
12. Nakabachi, A. et al. The 160-kilobase genome of the bacterial endosymbiont *Carsonella*. *Science* **314**, 267 (2006).
13. Lobry, J. R. Asymmetric substitution patterns in the two DNA strands of bacteria. *Mol. Biol. Evol.* **13**, 660–665 (1996).
14. Paez-Espino, D. et al. Uncovering Earth's virome. *Nature* **536**, 425–430 (2016).
15. Paez-Espino, D. et al. IMG/VR: a database of cultured and uncultured DNA viruses and retroviruses. *Nucleic Acids Res.* **45**, D457–D465 (2017).
16. Ivanova, N. N. et al. Stop codon reassignments in the wild. *Science* **344**, 909–913 (2014).
17. Mizuno, C. M. et al. Numerous cultivated and uncultivated viruses encode ribosomal proteins. *Nat. Commun.* **10**, 752 (2019).
18. van Duin, J. & Wijnands, R. The function of ribosomal protein S21 in protein synthesis. *Eur. J. Biochem.* **118**, 615–619 (1981).
19. Farwell, M. A., Roberts, M. W. & Rabinowitz, J. C. The effect of ribosomal protein S1 from *Escherichia coli* and *Micrococcus luteus* on protein synthesis in vitro by *E. coli* and *Bacillus subtilis*. *Mol. Microbiol.* **6**, 3375–3383 (1992).
20. Sørensen, M. A., Fricke, J. & Pedersen, S. Ribosomal protein S1 is required for translation of most, if not all, natural mRNAs in *Escherichia coli in vivo*. *J. Mol. Biol.* **280**, 561–569 (1998).
21. Frank, J. A. et al. Structure and function of a cyanophage-encoded peptide deformylase. *ISME J.* **7**, 1150–1160 (2013).
22. Janssen, B. D. & Hayes, C. S. The tmRNA ribosome-rescue system. *Adv. Protein Chem. Struct. Biol.* **86**, 151–191 (2012).
23. Yan, W. X. et al. Functionally diverse type V CRISPR–Cas systems. *Science* **363**, 88–91 (2019).
24. Shmakov, S. et al. Diversity and evolution of class 2 CRISPR–Cas systems. *Nat. Rev. Microbiol.* **15**, 169–182 (2017).
25. Harrington, L. B. et al. Programmed DNA destruction by miniature CRISPR–Cas14 enzymes. *Science* **362**, 839–842 (2018).
26. Seed, K. D., Lazinski, D. W., Calderwood, S. B. & Camilli, A. A bacteriophage encodes its own CRISPR/Cas adaptive response to evade host innate immunity. *Nature* **494**, 489–491 (2013).
27. Luo, M. L., Mullis, A. S., Leenay, R. T. & Beisel, C. L. Repurposing endogenous type I CRISPR–Cas systems for programmable gene repression. *Nucleic Acids Res.* **43**, 674–681 (2015).
28. Stachler, A.-E. & Marchfelder, A. Gene repression in Haloarchaea using the CRISPR (clustered regularly interspaced short palindromic repeats)–Cas I-B system. *J. Biol. Chem.* **291**, 15226–15242 (2016).
29. Toms, A. & Barrangou, R. On the global CRISPR array behavior in class I systems. *Biol. Direct* **12**, 20 (2017).
30. Brown, K. L. & Hughes, K. T. The role of anti-sigma factors in gene regulation. *Mol. Microbiol.* **16**, 397–404 (1995).
31. Bondy-Denomy, J. et al. Multiple mechanisms for CRISPR–Cas inhibition by anti-CRISPR proteins. *Nature* **526**, 136–139 (2015).
32. Pawluk, A. et al. Inactivation of CRISPR–Cas systems by anti-CRISPR proteins in diverse bacterial species. *Nat. Microbiol.* **1**, 16085 (2016).
33. Chaikeeratisak, V. et al. Assembly of a nucleus-like structure during viral replication in bacteria. *Science* **355**, 194–197 (2017).
34. Chaikeeratisak, V. et al. The phage nucleus and tubulin spindle are conserved among large *Pseudomonas* phages. *Cell Rep.* **20**, 1563–1571 (2017).
35. Mendoza, S. D. et al. A bacteriophage nucleus-like compartment shields DNA from CRISPR nucleases. *Nature* **577**, 244–248 (2020).
36. Gogokhia, L. et al. Expansion of bacteriophages is linked to aggravated intestinal inflammation and colitis. *Cell Host Microbe* **25**, 285–299 (2019).
37. Jaafar, Z. A. & Kieft, J. S. Viral RNA structure-based strategies to manipulate translation. *Nat. Rev. Microbiol.* **17**, 110–123 (2019).
38. Woese, C. The universal ancestor. *Proc. Natl Acad. Sci. USA* **95**, 6854–6859 (1998).
39. Subramanian, A. R. Structure and functions of ribosomal protein S1. *Prog. Nucleic Acid Res. Mol. Biol.* **28**, 101–142 (1983).
40. Loveland, A. B. & Korostelev, A. A. Structural dynamics of protein S1 on the 70S ribosome visualized by ensemble cryo-EM. *Methods* **137**, 55–66 (2018).
41. Pettersen, E. F. et al. UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).

Methods

Phage- and plasmid-genome identification

Datasets generated in the current study, those from previous research conducted by our team, the *Tara* Oceans microbiomes[42] and the Global Oceans Virome[43] were searched for sequence assemblies that could have derived from phages with genomes of more than 200 kb in length. Read assembly, gene prediction and initial gene annotation followed standard, previously reported methods[44–48].

Phage candidates were initially found by retrieving sequences that were not assigned to a genome and had no clear taxonomic profile at the domain level. Taxonomic profiles were determined through a voting scheme, in which the winning taxonomy had to have more than 50% votes for each taxonomic rank on the basis of protein annotations in the UniProt and ggKbase (https://ggkbase.berkeley.edu/) databases[49]. Phages were further narrowed down by identifying sequences with a high number of hypothetical protein annotations and/or the presence of phage-specific genes, such as capsid, tail, terminase, spike, holin, portal and baseplate. All candidate phage sequences were checked throughout to distinguish putative prophages from phages. Prophages were identified on the basis of a clear transition into the host genome with a high fraction of confident functional predictions, often associated with core metabolic functions and much higher similarity to bacterial genomes. Plasmids were distinguished from phages on the basis of matches to plasmid partitioning and conjugative transfer genes. Those that did not have phage-specific genes were assigned using phylogenetic tree placement using *recA*, *polA*, *polB*, *dnaE* and the DNA sliding clamp loader gene. Phages and placement assignments were further verified using a network of protein clustering with proteins from RefSeq prokaryotic viruses and 400 randomly sampled plasmids of more than 200 kb using vContact2[50] (Extended Data Fig. 6).

Phage- and plasmid-genome manual curation

All classified scaffolds were tested for end overlaps indicative of circularization. Assembled sequences that could be perfectly circularized were considered potentially complete. Erroneous concatenated sequence assemblies were initially flagged by searching for direct repeats of more than 5 kb using Vmatch[51]. Potentially concatenated sequence assemblies were manually checked for multiple large repeating sequences using the dotplot and RepeatFinder features in Geneious v.9. Sequences were corrected and removed from further analysis if the corrected length was more than 200 kb.

A subset of the phage sequences was selected for manual curation, with the goal of finishing (replacing all Ns at scaffolding gaps or local misassemblies by the correct nucleotide sequences and circularization). Curation generally followed previously described methods[9]. In brief, reads from the appropriate dataset were mapped using Bowtie2 v.2.3.4.1[52] to the de novo assembled sequences. Unplaced mate pairs of mapped reads were retained with shrinksam (https://github.com/bcthomas/shrinksam). Mappings were manually checked throughout to identify local misassemblies using Geneious v.9. N-filled gaps or misassembly corrections made use of unplaced paired reads, in some cases using reads relocated from sites to which they were mismapped. In such cases, mismappings were identified on the basis of much larger than expected paired read distances, high polymorphism densities, backwards mapping of one read pair or any combination of these. Similarly, ends were extended using unplaced or incorrectly placed paired reads until circularization could be established. In some cases, extended ends were used to recruit new scaffolds that were then added to the assembly. The accuracy of all extensions and local assembly changes were verified in a subsequent phase of read mapping. In many cases, assemblies were terminated or internally corrupted by the presence of repeated sequences. In these cases, blocks of repeated sequences as well as unique flanking sequences were identified. Reads were then manually relocated, respecting paired-read placement rules and unique flanking sequences. After gap closure, circularization and verification of accuracy throughout, end overlap was eliminated, genes were predicted and the start moved to an intergenic region, which was—in some cases—suspected to be origin on the basis of a combination of coverage trends and GC skew[53]. Finally, the sequences were checked to identify any repeated sequences that could have led to an incorrect path choice because the repeated regions were larger than the distance spanned by paired reads. This step also ruled out artefactual long phage sequences generated by end-to-end repeats of smaller phages, which occur in previously described datasets[9].

Structural and functional annotations

After the identification and curation of phage genomes, coding sequences and Shine–Dalgarno ribosomal binding site motifs were predicted with Prodigal using genetic code 11 (-m -g 11 -p single). The resulting coding sequences were annotated as previously described by searching UniProt, UniRef100 and KEGG[54]. Functional annotations were further assigned by searching proteins in PFAM r32[55], TIGRFAMS r15[56], Virus Orthologous Groups (VOG) r90 (http://vogdb.org/) and Prokaryotic Virus Orthologous Groups[57] (pVOG). tRNAs were identified with tRNAscan-s.e. v.2.0[58] using the bacterial model. tmRNAs were assigned using ARAGORN v.1.2.38[59] with the genetic code of bacteria and plant chloroplasts.

Clustering of the coding sequences into families was achieved using a two-step procedure. A first protein clustering was done using the fast and sensitive protein-sequence searching software MMseqs[60]. An all-versus-all sequences search was performed using an *E*-value cut-off of 1×10^{-3}, sensitivity of 7.5 and coverage of 0.5. A sequence similarity network was built on the basis of the pairwise similarities and the greedy set cover algorithm from MMseqs was performed to define protein subclusters. The resulting subclusters were defined as subfamilies. To test for distant homology, we grouped subfamilies into protein families using a comparison of hidden Markov models (HMMs). The proteins of each subfamily with at least two protein members were aligned using the result2msa parameter of MMseqs, and HMM profiles were built using the HHpred[61] suite from the multiple sequence alignments. The subfamilies were then compared to each other using HHblits from the HHpred suite (with parameters -v 0 -p 50 -z 4 -Z 32000 -B 0 -b 0). For subfamilies with probability scores of at least 95% and coverage at least 0.50, a similarity score (probability × coverage) was used as weight of the input network in the final clustering using the Markov clustering algorithm[62], with 2.0 as the inflation parameter. These clusters were defined as the protein families. Protein sequences were functionally annotated on the basis of their best hmmsearch match (v.3.1) (*E*-value cut-off 1×10^{-3}) against an HMM database constructed on the basis of orthologous groups defined by the KEGG database[63] (downloaded on 10 June 2015). Domains were predicted using the same hmmsearch procedure against the PFAM r31 database[55]. The domain architecture of each protein sequence was predicted using the DAMA software[64] (default parameters). SIGNALP[65] (v.4.1) (parameters, -f short -t gram+) and PSORT[66] v.3 (parameters, --long --positive) were used to predict the putative cellular localization of the proteins. Prediction of transmembrane helices in proteins was performed using TMHMM[67] (v.2.0) (default parameters). Hairpins (palindromes, based on identical overlapping repeats in the forward and reverse directions) were identified using the Geneious Repeat Finder and located across the dataset using Vmatch[51]. Repeats of more than 25 bp with 100% similarity were tabulated.

Reference genomes for size comparisons

RefSeq r92 genomes were recovered using the NCBI Virus portal and selecting only complete dsDNA genomes with bacterial hosts. Genomes from a previously published study[14] were downloaded from IMG/VR and only sequence assemblies that were labelled 'circular' with predicted bacterial hosts were retained. Given the presence of sequences in IMG/

VR that were based on erroneous concatenations, we only considered sequences from this source that were more than 200 kb; however, a subset of these was removed as artefactual sequences.

Alternative genetic codes

In cases in which the gene prediction using the standard bacterial code (code 11) resulted in seemingly anomalously low coding densities, potential alternative genetic codes were investigated. In addition to making a prediction using the fast and accurate genetic code inference and logo[68] (FACIL) web server, we identified genes with well-defined functions (for example, polymerase or nuclease) and determined the stop codons terminating genes that were shorter than expected. We then repredicted genes using GLIMMER3 v.1.5[69] and Prodigal with TAG not interpreted as a stop codon. Other combinations of repurposed stop codons were evaluated and candidate codes (for example, code 6, with only one stop codon) were ruled out owing to unlikely gene-fusion predictions.

Large terminase subunit and MCP phylogenetic analyses

The phylogenetic tree of the large terminase subunit was constructed by recovering large terminases from the aforementioned protein-clustering and annotation pipeline. The coding sequences that matched with >30 bitscore to PFAM, TIGRFAMS, VOG and pVOG were retained. Any coding sequence that had a hit to large terminase, regardless of bitscore, was searched using HHblits[70] against the uniclust30_2018_08 database. The resulting alignment was then further searched against the PDB70 database. Remaining coding sequences that clustered in protein families with a large terminase HMM were also included after manual verification. Detected large terminases were manually verified using the HHPred[70] and jPred[71] webservers. Large terminases from the >200-kb phage genomes[14] and all >200-kb complete dsDNA phage genomes from RefSeq r92 were also included by protein family clustering with the phage-coding sequences from this study. The resulting terminases were clustered at 95% amino acid identity to reduce redundancy using CD-HIT[72]. Smaller phage genomes were included by searching the resulting coding sequences set against the full RefSeq protein database and retaining the top 10 best hits. Those hits that had no large terminase match against PFAM, TIGRFAMS, VOG or pVOG were removed from further consideration and the remaining set was clustered at 90% amino acid identity. The final set of large terminase coding sequences that were more than 100 amino acids in length were aligned using MAFFT[73] v.7.407 (--localpair --maxiterate 1000) and poorly aligned sequences were removed and the resulting set was realigned. The phylogenetic tree was inferred using IQTREE v.1.6.6 using automatic model selection[74]. The phylogenetic tree of MCP genes was constructed by retrieving all MCPs annotated by combining the PFAM annotations of protein families and direct annotations by PFAM, TIGRFAMS, VOG and pVOG. Reference MCP gene sequences were collected using the same strategy and sources as for the large terminase subunit tree. The resulting set was further screened by searching against PFAM, TIGRFAMS, VOG and pVOG and removing matches that had no large terminase match regardless of bitscore. The final set of MCP sequences were aligned with MAFFT(--localpair --maxiterate 1000) and the phylogenetic tree was constructed using IQTREE with automatic model selection and 1,000 bootstrap replicates.

Whole-genome scale clustering

To identify phage genomes that were closely related at the whole-genome level, we compared sequences using whole-genome alignments. The goal of this analysis was to further corroborate the identified phylogenetic clades and test for the presence of very similar phages in different habitats and environments. Genomes grouped together in the primary clusters from dRep v.2[75] were evaluated for genome alignment using Mauve[76] within Geneious v.9.

CRISPR–Cas locus and target detection

Phage- and host-encoded CRISPR loci (repeats and spacers) were identified using a combination of MinCED (https://github.com/ctSkennerton/minced) and CRISPRDetect[77]. A custom database of Cas genes was built by collecting Cas gene sequences from previous studies[23,25,78–82] and built with MAFFT (--localpair --maxiterate 1000) and hmmbuild. The coding sequences from this study were searched against the HMM database using hmmsearch with $E < 1 \times 10^{-5}$. Matches were checked using a combination of hmmscan and BLAST searches against the NCBI nr database and manually verified by identifying colocated CRISPR arrays and Cas genes. Spacers extracted from between repeats of the CRISPR locus were compared to sequences assemblies from the same site using BLASTN-short[83]. Matches with alignment length >24 bp and ≤1 mismatch were retained and targets were classified as bacteria, phage or other. CRISPR arrays that had ≤1 mismatch, were further searched for more spacer matches in the target sequence by finding more hits with ≤3 mismatches.

Host identification

The phylum affiliations of bacterial hosts for phage and plasmid-like sequences were predicted by considering the UniProt taxonomic profiles of every coding sequence for each phage genome. The phylum level matches for each phage genome were summed and the phylum with the most hits was considered as the potential host phylum. However, only cases in which this phylum that had 3× as many counts as the next most-counted phylum were assigned as the tentative phage host phylum. Phage hosts were further assigned and verified using the CRISPR-targeting strategy describe above with the phage and plasmid-like genomes as targets. CRISPR arrays were predicted on all sequence assemblies from the same site that each phage genome was reconstructed. Sequence assemblies containing spacers with a match of length >24 bp and ≤1 mismatch were used to infer phage–host relationships. In all cases, the predicted host phylum based on taxonomic profiling and CRISPR targeting were in complete agreement. Similarly, the phyla of hosts were predicted on the basis of phylogenetic analysis of phage genes also found in host genomes (for example, involved in translation and nucleotide reactions). Inferences based on computed taxonomic profiles and phylogenetic trees were also in complete agreement.

Phage-encoded tRNA synthetase trees

Phylogenetic trees were constructed for phage-encoded tRNA synthetase, ribosomal and initiation factor protein sequences using a set of the closest reference sequences from NCBI and bacterial genomes from the current study. The tRNA synthetases were identified on the basis of annotation of genes using the standard ggKbase pipeline (see above), and confirmed by HMMs with datasets from TIGRFAMs. For each type of tRNA synthetase, references were selected by comparing all of the corresponding genes of this type against the NCBI nr database using DIAMOND v.0.9.24[84], their top 100 hits were clustered by CD-HIT using a 90% similarity threshold[72]. The phylogenetic tree of each tRNA synthetase was constructed using RAxML v.8.0.26[85] with the PROTGAMMALG model.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

GenBank files for all genomes are provided as Supplementary Information. Sequence reads and genomes have been deposited at the European Nucleotide Archive (ENA) under project PRJEB35371. Genomes have been deposited at ENA under accessions ERS4026114–ERS4026474.

Reads are available at ENA under accessions ERS4025670–ERS4025731. Read accessions and genome accessions for each phage genome are included in Supplementary Table 1.

Code availability

The custom code used to analyse the genomes is available at http://www.github.com/rohansachdeva/assembly_repeats.

42. Karsenti, E. et al. A holistic approach to marine eco-systems biology. *PLoS Biol*. **9**, e1001177 (2011).
43. Roux, S. et al. Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature* **537**, 689–693 (2016).
44. Peng, Y., Leung, H. C. M., Yiu, S. M. & Chin, F. Y. L. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **28**, 1420–1428 (2012).
45. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. metaSPAdes: a new versatile metagenomic assembler. *Genome Res*. **27**, 824–834 (2017).
46. Edgar, R. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461 (2010).
47. Joshi, N. A. & Fass, J. N. Sickle: a sliding-window, adaptive, quality-based trimming tool for FastQ files. v.1.33 https://github.com/najoshi/sickle (2011).
48. Bushnell, B. BBMap short read aligner. https://sourceforge.net/projects/bbmap/ (2019).
49. Raveh-Sadka, T. et al. Gut bacteria are rarely shared by co-hospitalized premature infants, regardless of necrotizing enterocolitis development. *eLife* **4**, e05477 (2015).
50. Bolduc, B. et al. vConTACT: an iVirus tool to classify double-stranded DNA viruses that infect *Archaea* and *Bacteria*. *PeerJ* **5**, e3243 (2017).
51. Kurtz, S. The Vmatch large scale sequence analysis software. http://www.vmatch.de/ (2017).
52. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
53. Brown, C. T., Olm, M. R., Thomas, B. C. & Banfield, J. F. Measurement of bacterial replication rates in microbial communities. *Nat. Biotechnol*. **34**, 1256–1263 (2016).
54. Wrighton, K. C. et al. Metabolic interdependencies between phylogenetically novel fermenters and respiratory organisms in an unconfined aquifer. *ISME J*. **8**, 1452–1463 (2014).
55. Finn, R. D. et al. Pfam: the protein families database. *Nucleic Acids Res*. **42**, D222–D230 (2014).
56. Haft, D. H. et al. TIGRFAMs and genome properties in 2013. *Nucleic Acids Res*. **41**, D387–D395 (2013).
57. Grazziotin, A. L., Koonin, E. V. & Kristensen, D. M. Prokaryotic Virus Orthologous Groups (pVOGs): a resource for comparative genomics and protein family annotation. *Nucleic Acids Res*. **45**, D491–D498 (2017).]
58. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res*. **25**, 955–964 (1997).
59. Laslett, D. & Canback, B. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Res*. **32**, 11–16 (2004).
60. Hauser, M., Steinegger, M. & Söding, J. MMseqs software suite for fast and deep clustering and searching of large protein sequence sets. *Bioinformatics* **32**, 1323–1330 (2016).
61. Remmert, M., Biegert, A., Hauser, A. & Söding, J. HHblits: lightning-fast iterative protein sequence searching by HMM–HMM alignment. *Nat. Methods* **9**, 173–175 (2012).
62. Enright, A. J., Van Dongen, S. & Ouzounis, C. A. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res*. **30**, 1575–1584 (2002).
63. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res*. **44**, D457–D462 (2016).
64. Bernardes, J. S., Vieira, F. R. J., Zaverucha, G. & Carbone, A. A multi-objective optimization approach accurately resolves protein domain architectures. *Bioinformatics* **32**, 345–353 (2016).
65. Petersen, T. N., Brunak, S., von Heijne, G. & Nielsen, H. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat. Methods* **8**, 785–786 (2011).
66. Peabody, M. A., Laird, M. R., Vlasschaert, C., Lo, R. & Brinkman, F. S. L. PSORTdb: expanding the bacteria and archaea protein subcellular localization database to better reflect diversity in cell envelope structures. *Nucleic Acids Res*. **44**, D663–D668 (2016).
67. Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. L. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol*. **305**, 567–580 (2001).
68. Dutilh, B. E. et al. FACIL: fast and accurate genetic code inference and logo. *Bioinformatics* **27**, 1929–1933 (2011).
69. Delcher, A. L., Harmon, D., Kasif, S., White, O. & Salzberg, S. L. Improved microbial gene identification with GLIMMER. *Nucleic Acids Res*. **27**, 4636–4641 (1999).
70. Steinegger, M., Meier, A. & Biegert, A. HH-suite3 for fast remote homology detection and deep protein annotation. *Bioinformatics* **20**, 473 (2019).
71. Cole, C., Barber, J. D. & Barton, G. J. The Jpred 3 secondary structure prediction server. *Nucleic Acids Res*. **36**, W197–W201 (2008).
72. Huang, Y., Niu, B., Gao, Y., Fu, L. & Li, W. CD-HIT suite: a web server for clustering and comparing biological sequences. *Bioinformatics* **26**, 680–682 (2010).
73. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol*. **30**, 772–780 (2013).
74. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol*. **32**, 268–274 (2015).
75. Olm, M. R., Brown, C. T., Brooks, B. & Banfield, J. F. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J*. **11**, 2864–2868 (2017).
76. Darling, A. C. E., Mau, B., Blattner, F. R. & Perna, N. T. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res*. **14**, 1394–1403 (2004).
77. Biswas, A., Staals, R. H. J., Morales, S. E., Fineran, P. C. & Brown, C. M. CRISPRDetect: a flexible algorithm to define CRISPR arrays. *BMC Genomics* **17**, 356 (2016).
78. Burstein, D. et al. New CRISPR–Cas systems from uncultivated microbes. *Nature* **542**, 237–241 (2017).
79. Makarova, K. S. et al. An updated evolutionary classification of CRISPR–Cas systems. *Nat. Rev. Microbiol*. **13**, 722–736 (2015).
80. Shmakov, S. et al. Discovery and functional characterization of diverse class 2 CRISPR–Cas systems. *Mol. Cell* **60**, 385–397 (2015).
81. Yan, W. X. et al. Cas13d is a compact RNA-targeting type VI CRISPR effector positively modulated by a WYL-domain-containing accessory protein. *Mol. Cell* **70**, 327–339 (2018).
82. Smargon, A. A. et al. Cas13b is a Type VI-B CRISPR-associated RNA-guided RNAse differentially regulated by accessory proteins Csx27 and Csx28. *Mol. Cell* **65**, 618–630 (2017).
83. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol*. **215**, 403–410 (1990).
84. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).
85. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
86. Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P.-L. & Ideker, T. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* **27**, 431–432 (2011)

Acknowledgements Funding for this project was provided by the National Institutes of Health (NIH) under awards RAI092531A and R01-GM109454 and the Alfred P. Sloan Foundation under grant APSF-2012-10-05 to J.F.B. and M.M.; National Science Foundation (NSF) Sustainable Chemistry grant 1349278 to J.F.B., NSF Graduate Research Fellowships to B.A.-S. (DGE 1752814) and M.R.O. (DGE 1106400); the Paul Allen Foundation Frontiers Group; Chan Zuckerberg Biohub; Innovative Genomics Institute; The Novo Nordisk Foundation (NNF16OC0021856) to P.M.; NASA 13R-0043 and CA AES to K.R.A. and K.F.; German Science Foundation postdoctoral scholarship to A.J.P. (DFG PR 1603/1-1); Camille & Henry Dreyfus Postdoctoral Fellowship in Environmental Chemistry to C.H.; Watershed Function Scientific Focus Area funded by the US Department of Energy, Office of Science, Office of Biological and Environmental Research (DE-AC02-05CH11231) to A.L., P.M.-C. and A.T.; NSF grants GRT00048468 and 1342701 to K.C.W. and M.B., and CHE-1740549 to J.H.D.C.; the Ministry of Economy, Trade and Industry of Japan to Y.A.; the March of Dimes Prematurity Research Center at Stanford University School of Medicine, the Thomas C. and Joan M. Merigan Endowment at Stanford University to D.A.R.; the National Research Foundation of South Africa (UID 64877) to S.T.L.H. and NSF CZO funding. We acknowledge the scientists who generated the public database sequences and thank J. Tung, E. Archie, F. Aarestrup and R. Kruger for contributing data and P. Pausch and G. Knott for discussions. The phage icon used in the graphics was created by Two Photon Art and the Innovative Genomics Institute.

Author contributions Analyses were conducted primarily by B.A.-S., R.S., L.-X.C., F.W. and J.F.B. Specifically, phylogenetic and gene inventory analyses were performed by B.A.-S. and R.S., tRNA synthetase analysis was led by L.-X.C., ribosome analyses were led by F.W. and B.A.-S. with input from J.H.D.C., and genome analysis scripts were written by R.S. R.M. provided code for protein clustering. Size-distribution and tRNA analysis was conducted by R.S. Genome curation was performed by J.F.B., P.M., L.-X.C., A.D. and C.S. B.A.-S. led the CRISPR–Cas analyses. J.A.D. provided support and scientific input. S.L. provided bioinformatics support. L.-X.C., P.M., A.D., C.J.C., M.R.O., K.B.-G., Y.A., C.H., B.B., A.T., A.L., P.M.-C., C.S., D.S.A.G., M.A.B., A.S., A.L.J., T.C.N., R. Kantor, R. Keren, K.R.L., I.F.F., K.F., R.A., K.A., J.Z., A.J.P., M.E.P., S.G.T., W.-J.L., K.W., S.H., M.M., D.A.R., A.-C.L., L.W. and J.M.S. contributed data. B.A.-S., R.S. and J.F.B. wrote the manuscript, with input from all authors. The study was conceived by J.F.B., with input from P.M. and A.D.

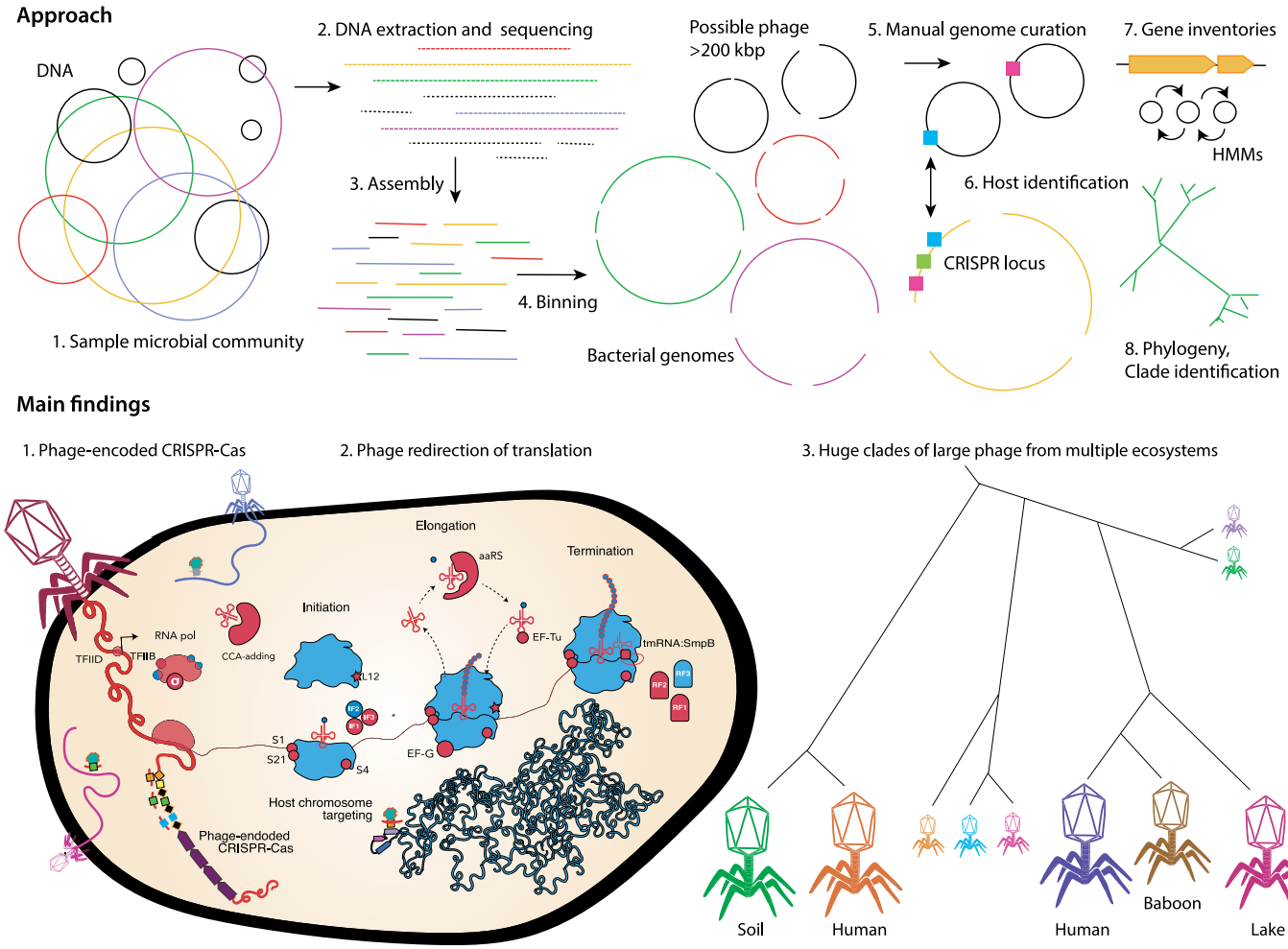
Competing interests The Regents of the University of California have patents pending for CRISPR technologies on which the authors are inventors. J.A.D. is a co-founder of Caribou Biosciences, Editas Medicine, Intellia Therapeutics, Scribe Therapeutics and Mammoth Biosciences, a scientific advisory board member of Caribou Biosciences, Intellia Therapeutics, eFFECTOR Therapeutics, Scribe Therapeutics, Synthego, Mammoth Biosciences and Inari, and is a Director at Johnson & Johnson and has sponsored research projects by Biogen and Pfizer. J.F.B. is a founder of Metagenomi.

Additional information

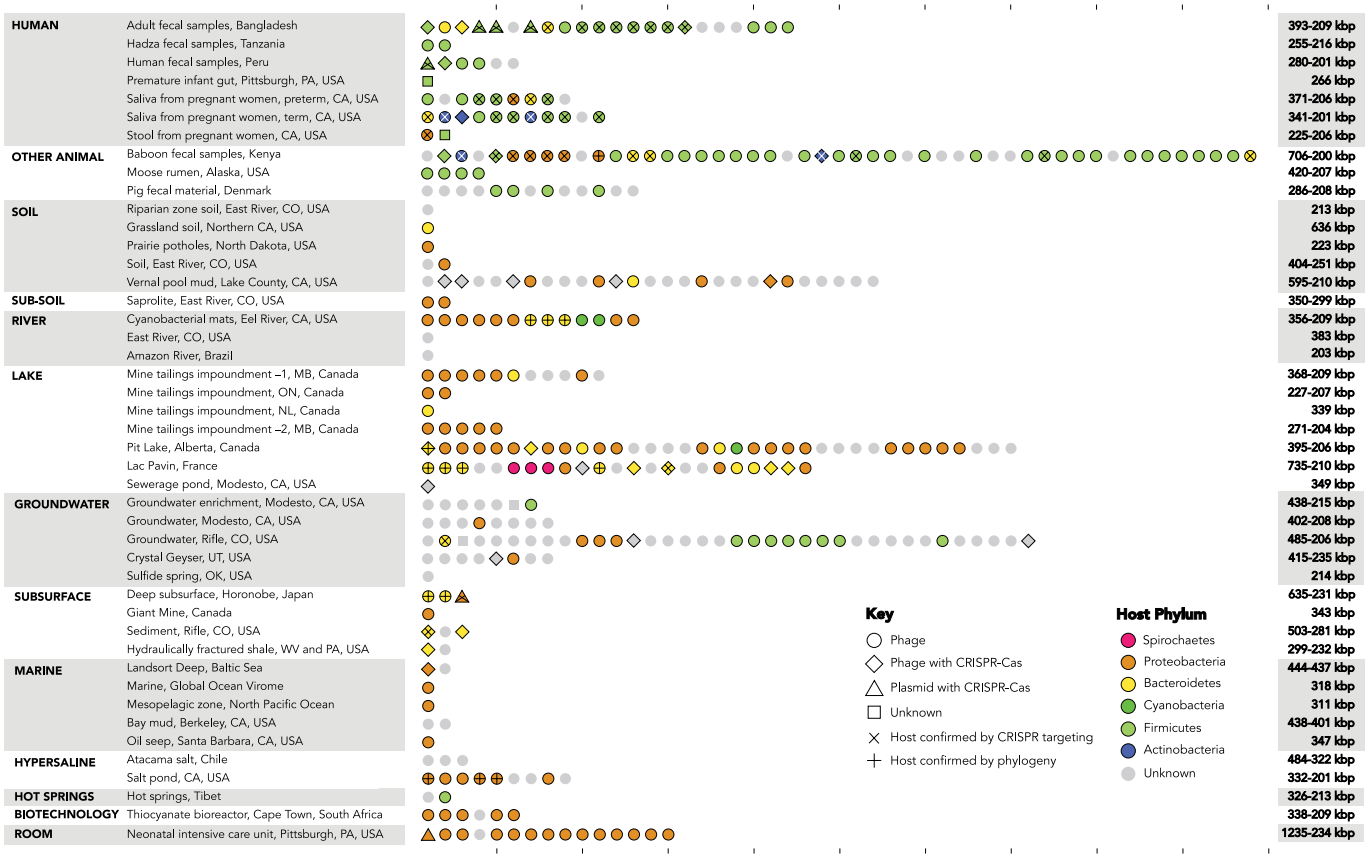
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2007-4.

Correspondence and requests for materials should be addressed to J.F.B.

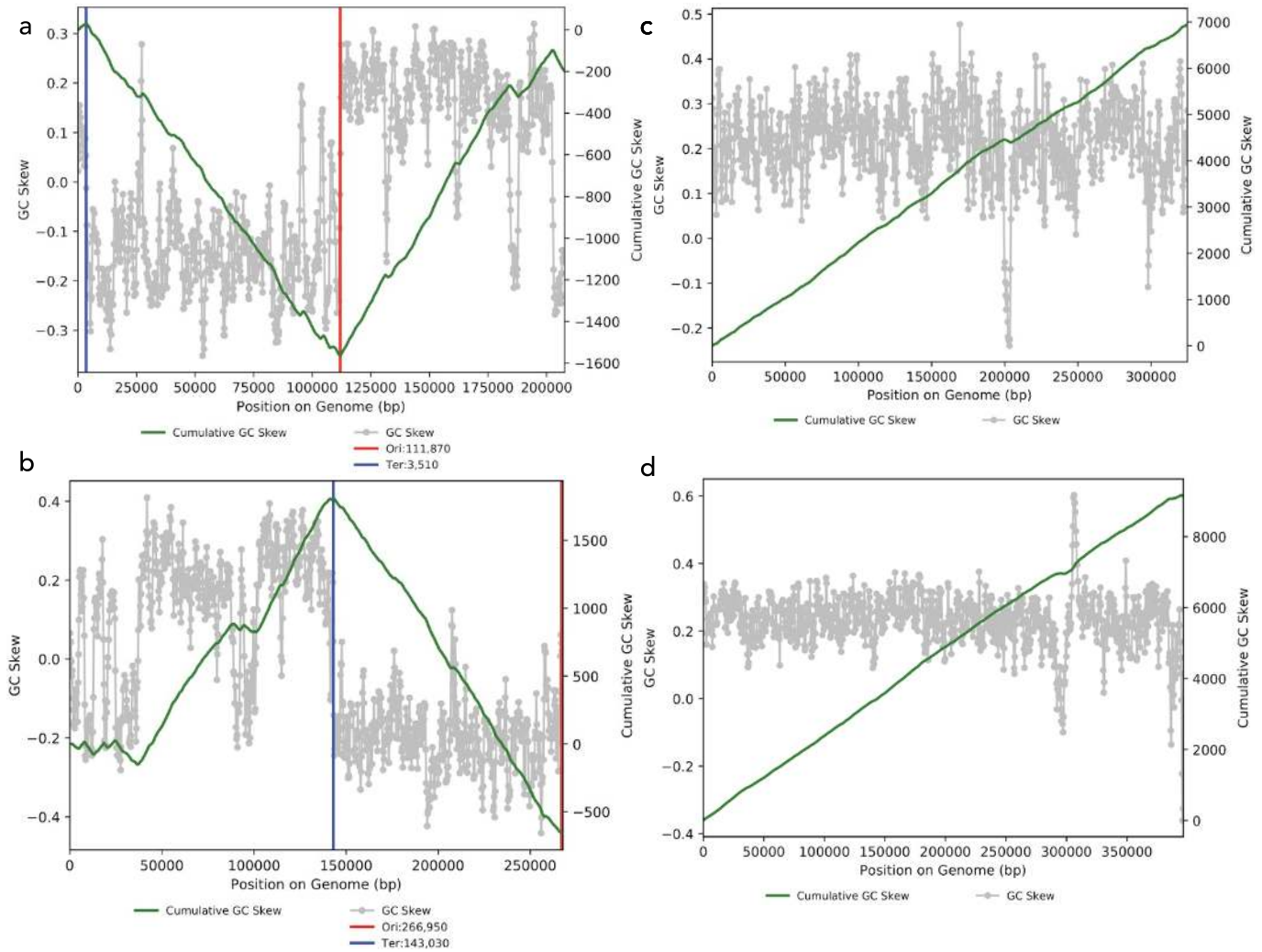
Peer review information *Nature* thanks Robert Edwards and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.



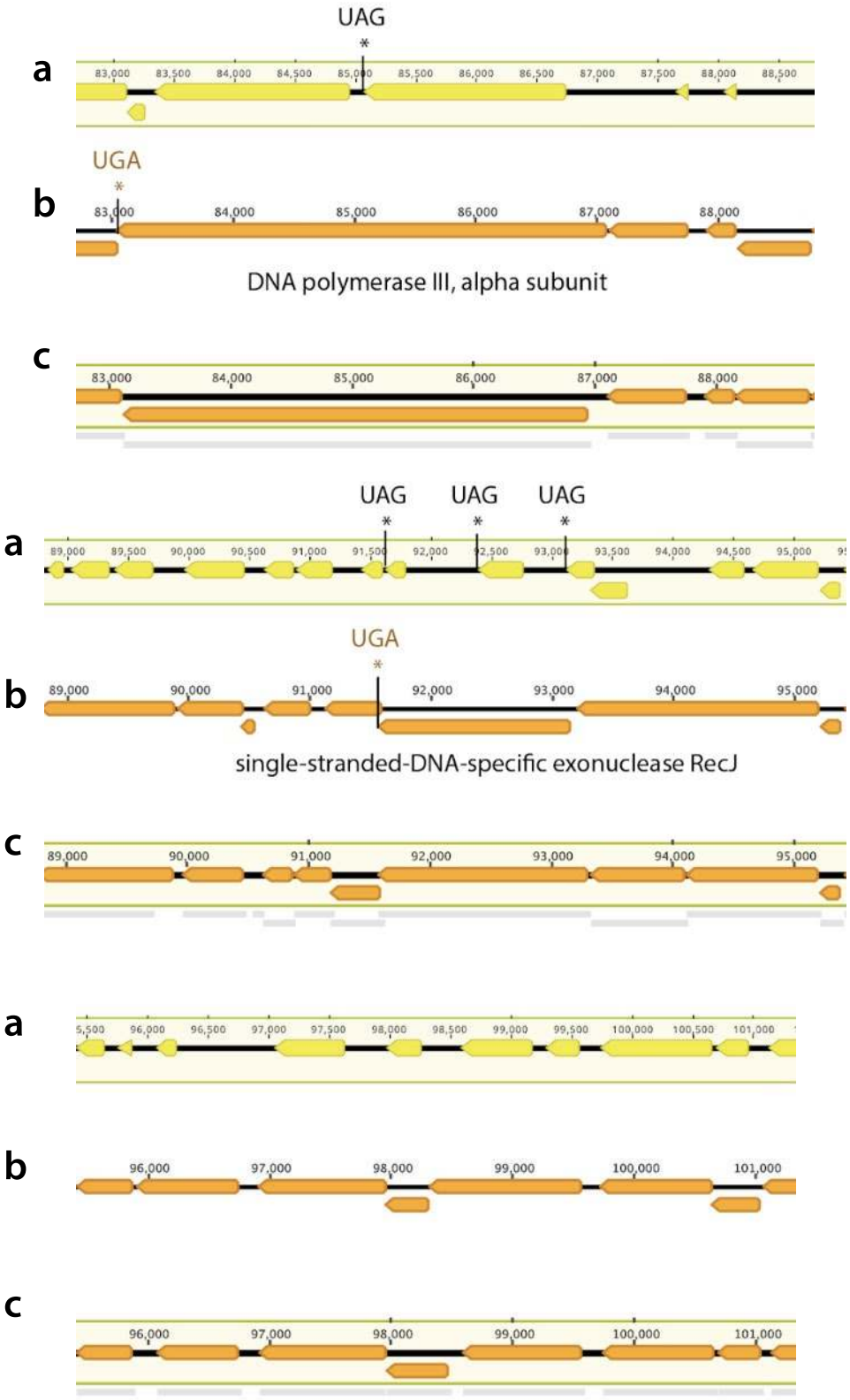
Extended Data Fig. 1 | Graphical abstract describing the approach and main findings of this study. aaRS, aminoacyl-tRNA synthetase; CCA-adding, tRNA nucleotidyltransferase; EF, elongation factor; IF, initiation factor; PDF, peptide deformylase; QueC/D/F, queuosine synthesis and tRNA modification; RF, release factor; RNA Pol, RNA polymerase; RRF, ribosome recycling factor; TFIIB, transcription factor II B.



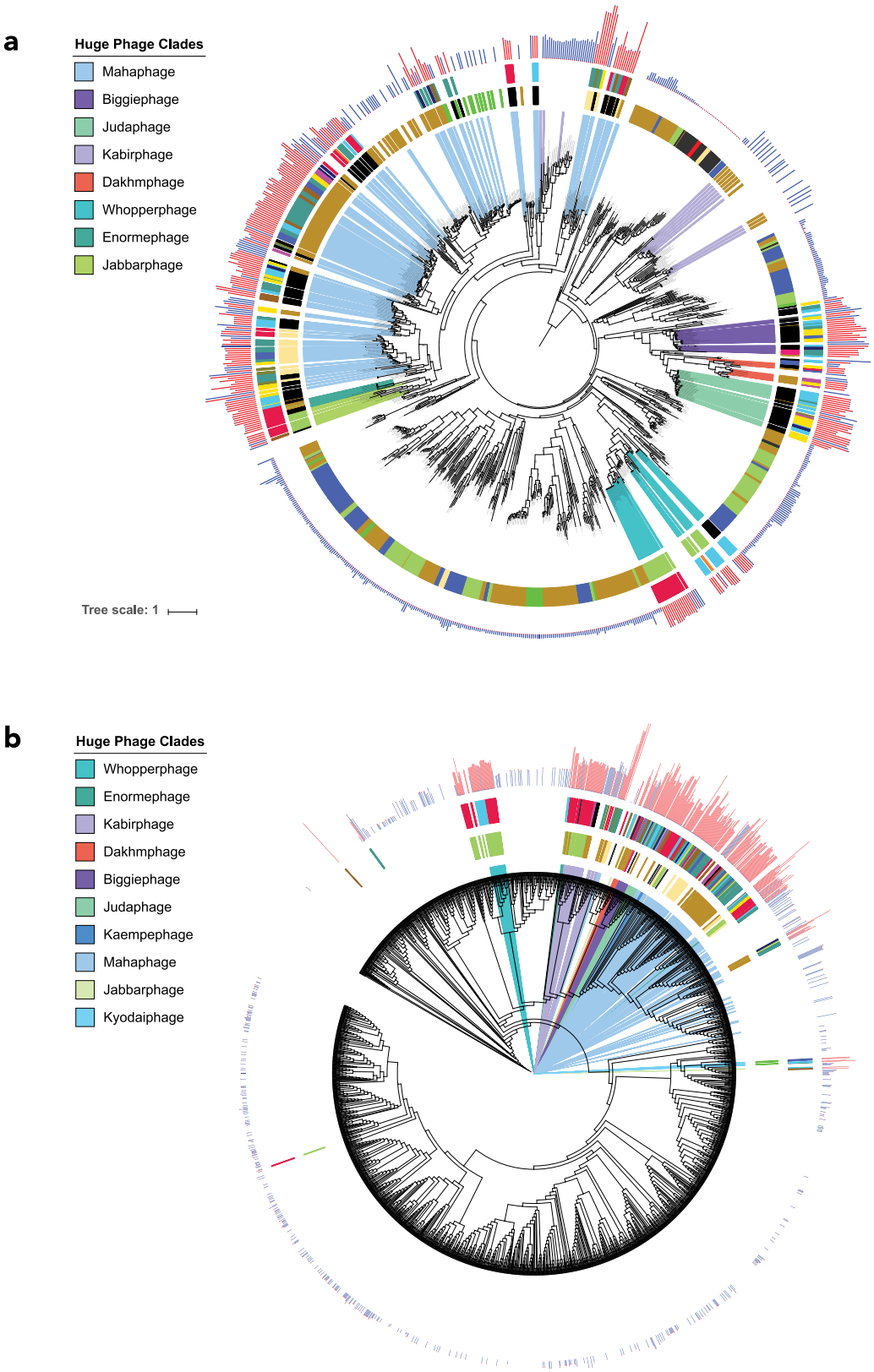
Extended Data Fig. 2 | Ecosystems with phage genomes and plasmid-like sequences of more than 200 kb. Genomes grouped by sampling-site type. Each box represents a phage genome or plasmid-like sequence, and boxes are horizontally arranged in order of decreasing genome size. The size range for each site type is listed to the right. Colours indicate putative host phylum on the basis of genome taxonomic profile, with confirmation by CRISPR spacer targeting (X) or rpS21 (+).



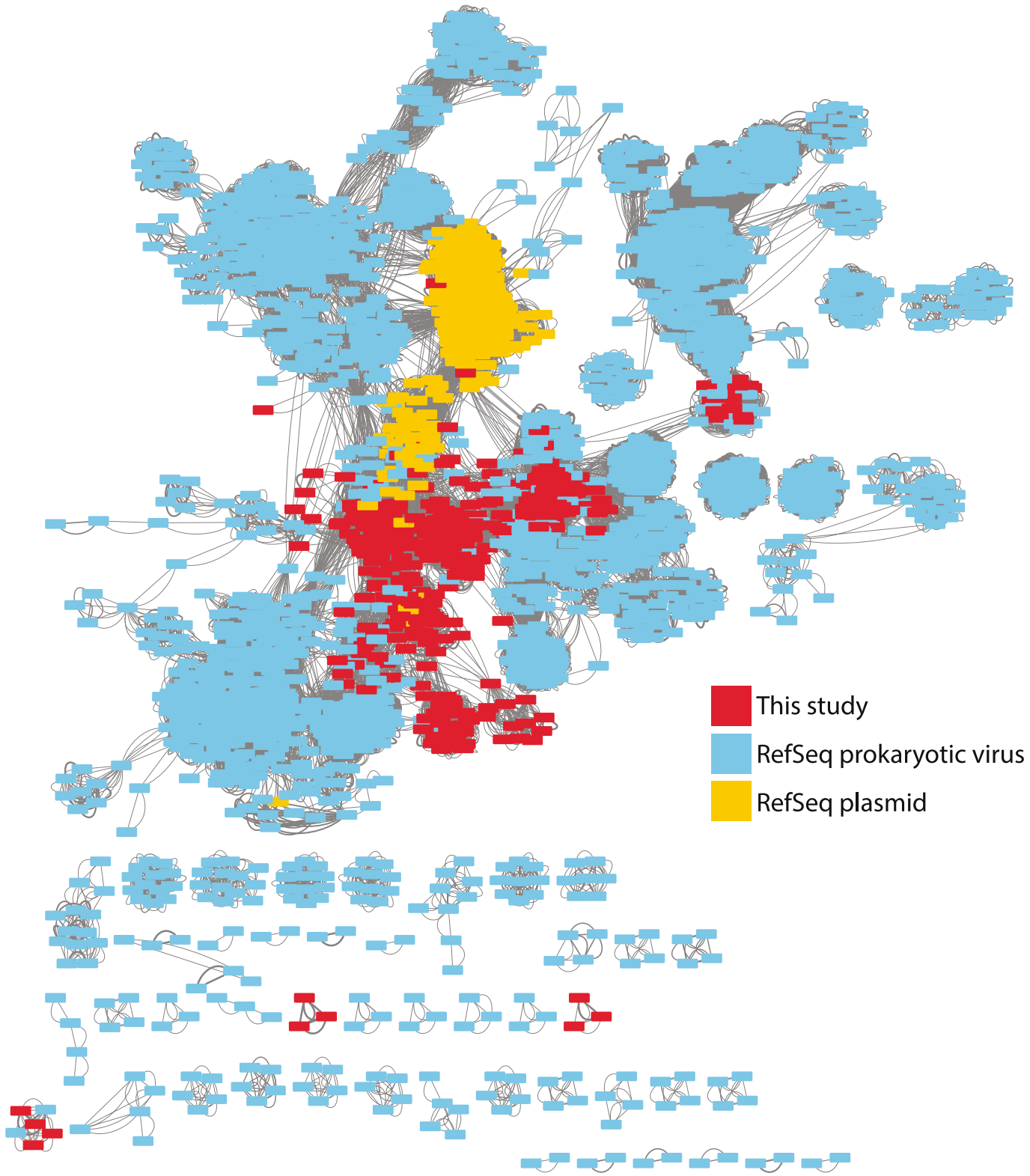
Extended Data Fig. 3 | Examples of phage genomes that display GC skew indicative of bidirectional replication. a, **b**, Example phage genomes with GC skew patterns that are strongly indicative of bidirectional replication (origin-to-terminus) that is typically found in bacteria (however, the origin may not correspond to the start of the genome). **c**, **d**, Phage genomes with GC skew patterns that are suggestive of unidirectional patterns.



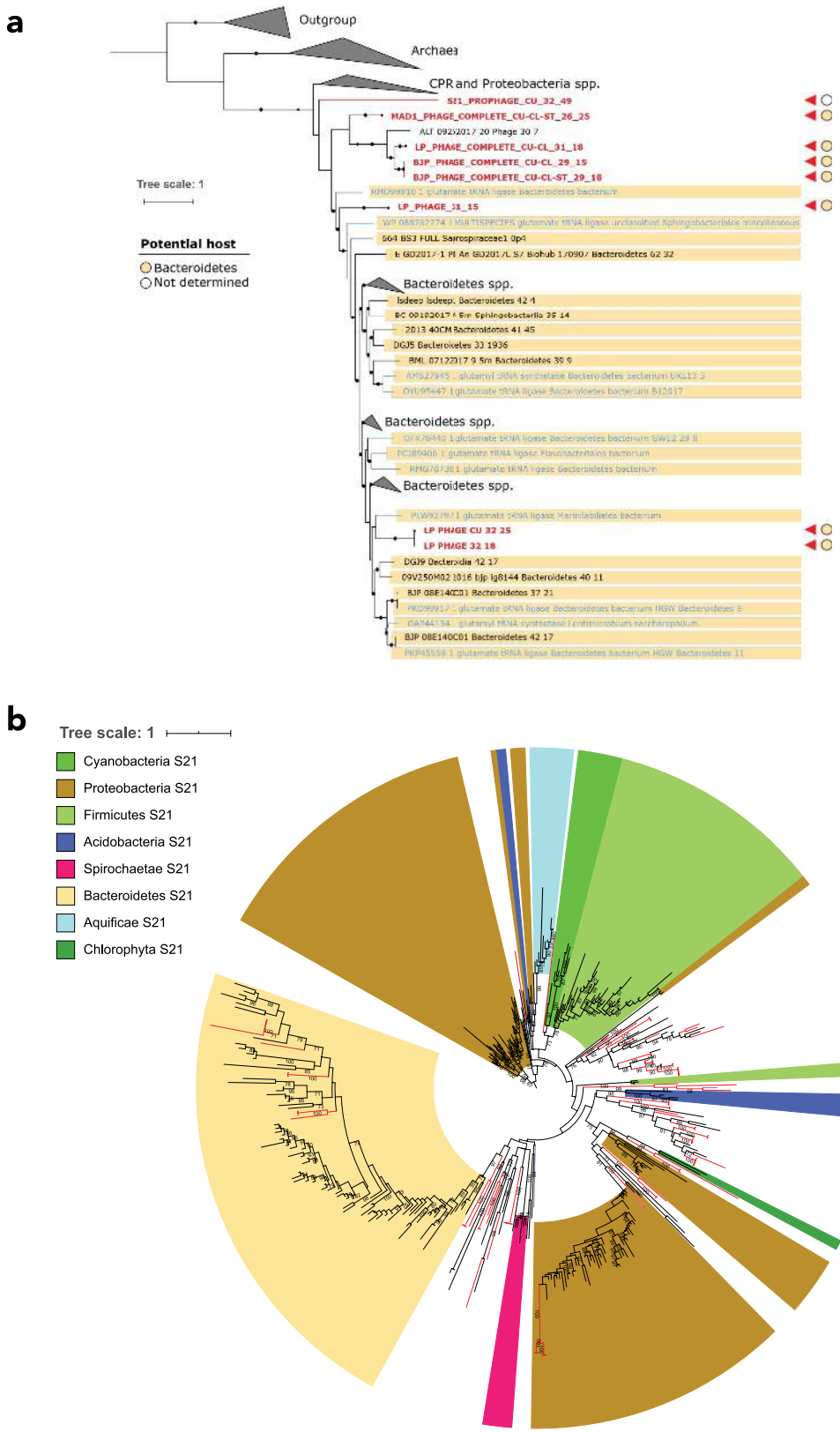
Extended Data Fig. 4 | Example of the alternative coding of phages. Comparisons of gene predictions for a region with genes of clearly predicted function in M05_PHAGE_COMPLETE_32_3. **a**, The standard (code 11) genetic code. **b**, Both TAG and TAA repurposed (code 6). **c**, TAG repurposed (code 16). Overall, analysis of well-annotated genes supported code 16 as the best choice (TAG to X, as X could not be clearly resolved on the basis of sequence alignments with related proteins).



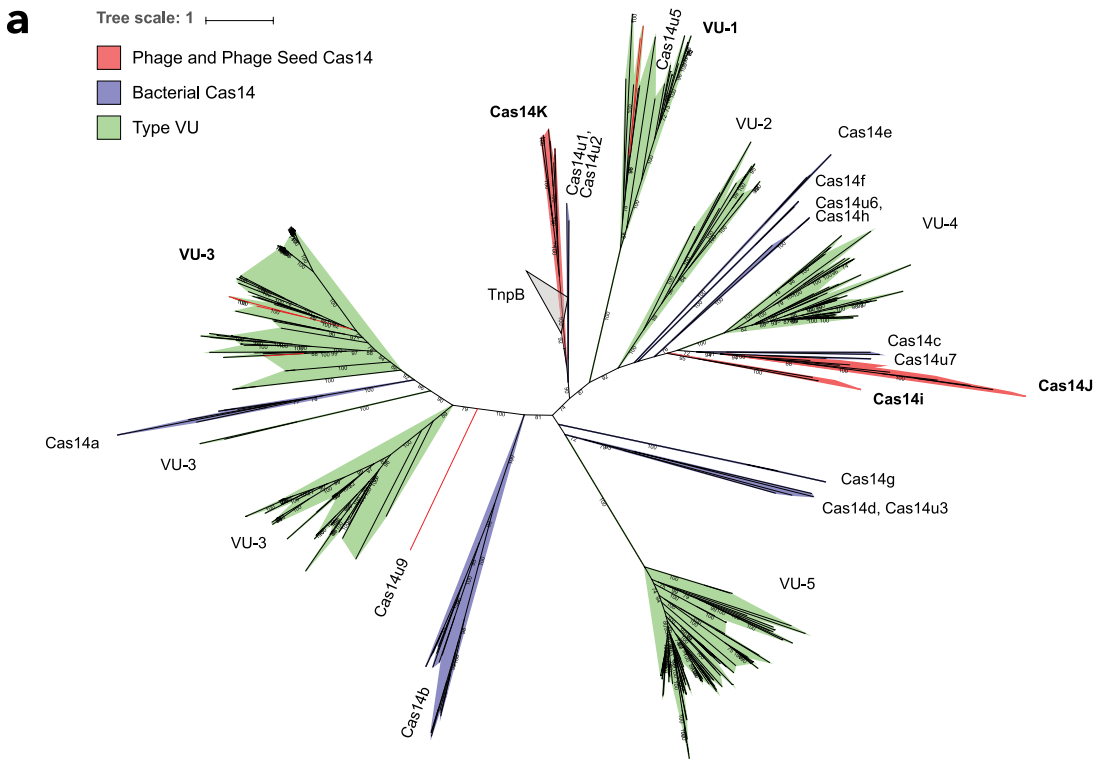
Extended Data Fig. 5 | Phylogenetic and protein-cluster relationships between phages. a, The phylogenetic tree of phages based on the MCPs. The outer ring shows genome length; bars in red indicate genomes reconstructed and reported in this study and bars in blue indicate database genomes. The next ring indicates the environment of origin. The inner ring indicates the phylum of the host (black indicates unknown). Superimposed colours indicate named clades that consist of huge phages that were identified in the terminase tree. Colours are as in Fig. 2. **b**, Hierarchical clustering dendrogram of phage genomes based on the Jaccard distance between the presence or absence profiles of protein families, performed using an average linkage method. The outermost ring shows phage genome length, the next ring shows the environment of origin, then predicted phylum affiliation of bacterial hosts. Superimposed colours indicate named clades that consist of huge phages that were identified in the terminase tree. Colours are as in Fig. 2. The clustering supports the phylogenetic analyses shown in **a** and Fig. 2.



Extended Data Fig. 6 | Protein-clustering network for phages and plasmids. Network analysis using vContact2 and Cytoscape[86] based on the number of shared protein clusters between the genomes in this study, RefSeq prokaryotic virus genomes and 400 randomly sampled plasmid sequences from RefSeq. Each node represents a genome and each edge is the hypergeometric similarity (>30) between genomes based on shared protein clusters. This analysis was used to help to distinguish between the classification of genomes as phage, plasmid or unknown.

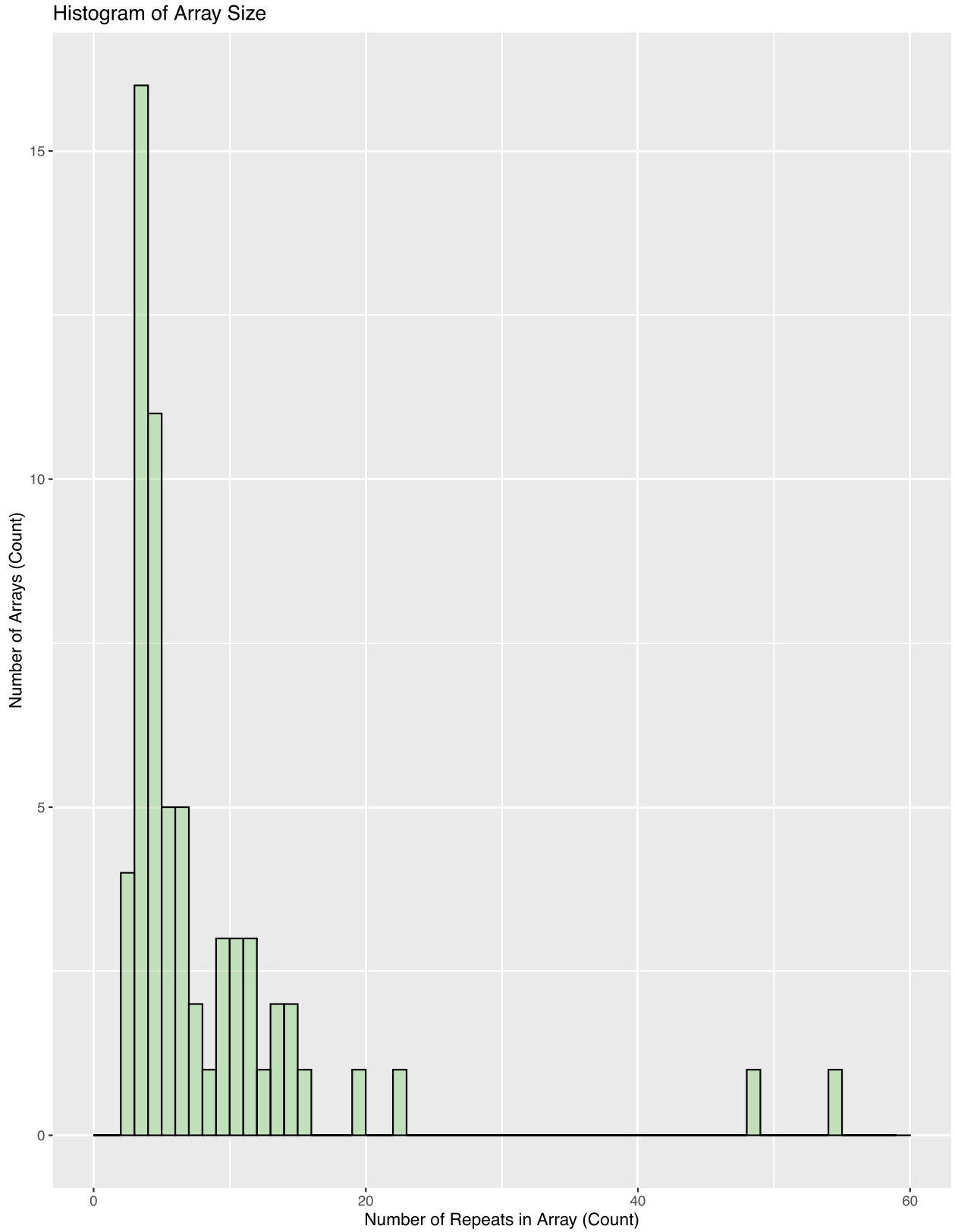


Extended Data Fig. 7 | Phylogenetic analysis of tRNA synthetase. **a**, Aminoacyl tRNA synthetases were detected in many huge phages reported in this study (Supplementary Table 6). The phylogenetic subtree for glutamate-tRNA synthetase sequences from phages (red text and small triangles) that place within or close to sequences from Bacteroidetes hosts is shown as an example. Bacterial sequences from public databases are indicated by black text and those from metagenomes from which huge phage genomes were reconstructed are indicated by blue text. Coloured circles indicate the predicted phylum of the bacterial host for each phage. **b**, Phylogenetic tree of phage-encoded ribosomal protein S21 and the top RefSeq hits for each protein, constructed using IQTREE. Sequences from this study are indicated by red branches.



Extended Data Fig. 8 | Phylogenetic trees of Cas14, CRISPR–Cas type V-U and Cas9. a, Phylogenetic tree for Cas14 and type V-U. **b**, Phylogenetic tree for Cas9. Sequences from this study are indicated by red branches.

Extended Data Fig. 9 | Variant type I CRISPR–Cas system and Cas4-like proteins found in the genomes of huge phages. a, Locus architecture for type-I variant CRISPR phages. An interesting type-I system identified in huge phages lacks Cas6 but has Cas5, which is most similar to the Cas5d protein from type I-C, in which Cas5d acts as the pre-crRNA endonuclease (a role commonly reserved for Cas6). The proposed active site residues of Cas5d are to some extent different in the Cas5 of this system, although this may still confer processing activity, as this change is also observed in other Cas6 homologues. **b**, Phylogenetic tree of superfamily 1–6 helicases, including Cas3 and the unidentified helicase in the type I-C variant system. Sequences from this study are indicated by red branches. **c**, Phylogenetic tree of Cas4, Cas4-like proteins from the phage and plasmid genomes reported here, and the top 50 RefSeq hits to the Cas4-like proteins. Cas4-like genes from this study are denoted by red circles.



Extended Data Fig. 10 | Distribution of phage- and plasmid-encoded CRISPR array sizes. The indicated count is the number of recovered repeats.

nature research

Corresponding author(s): Jillian F Banfield

Last updated by author(s): Nov 13, 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Geneious v9.1.8 (Licensed, paid version used in this study, free versions available) BBmap v37.5 IDBA_UD v1.1.1 Bowtie2 v2.3.4.1 MEGAHIT v1.1.3 SPAdes v3.11.1
Data analysis	vContact2 Vmatch v2.3 Prodigal v2.6.3 tRNAscan-SE v2.0 ARAGORN v1.2.38 MMseqs2 Version: 9f493f538d28b1412a2d124614e9d6ee27a55f45 HHsuite v3.0.3 CD-HIT v4.6.8 SignalP v4.1 DAMA v1.0 PSORT v3.0 TMHMM v2.0 MAFFT v7.407 DIAMOND v0.9.24 RAxML v8.0.26 IQTREE v1.6.6 HMMER v3.1b2 GLIMMER3 v1.5

MinCED v0.2.0
https://github.com/rohansachdeva/assembly_repeats v0

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Accession numbers for each genome and reads are provided in the data availability statement and in Table S1. Genbank files for each genome are also provided as supplementary data.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	The sample size was chosen to provide a high breadth of ecosystem coverage for the recovery of huge phage genomes. Accordingly, there were no statistical methods to determine sample sizes. Data were compiled from multiple sources where virus-like genomes could be found. Data were compiled from multiple sources wherever virus-like genomes could be found. Samples from each sampling site is listed in Table 1 and Table S1
Data exclusions	IMG/VR phage data was excluded because of the prevalence of artifactual concatenated viral sequence assemblies. This was pre-established based on criteria in Devoto et al. 2019.
Replication	Near identical phage genomes were recovered from multiple independent samples, verifying sequence assembly and genome reconstruction. Host identification was verified by a combination of CRISPR targeting, phylogenetic analysis of ribosomal proteins, and phylum-level taxonomic profiles. Annotations were verified across multiple databases.
Randomization	Randomization is not applicable to this because the reported study is a survey of huge phage genomes across many ecosystems and not dependent on experimental outcomes.
Blinding	Blinding was not performed because it was not applicable to this study. This study was a survey of huge phage genomes across global populations, and was not dependent on trial outcomes

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging