

Long-read metagenomics reveals phage dynamics in the human gut microbiome

<https://doi.org/10.1038/s41586-025-09786-2>

Received: 9 May 2025

Accepted: 21 October 2025

Published online: 26 November 2025

Open access

 Check for updates

Jakob Wirbel^{1,6}, Angela S. Hickey^{2,6}, Daniel Chang², Nora J. Enright³, Mai Dvorak⁴, Rachael B. Chanin¹, Danica T. Schmidtke⁵ & Ami S. Bhatt^{1,2,✉}

Gut bacteriophages profoundly impact microbial ecology and health^{1–3}; yet, they are understudied. Using deep long-read bulk metagenomic sequencing, we tracked prophage integration dynamics in stool samples from six healthy individuals, spanning a 2-year timescale. Although most prophages remained stably integrated into their hosts, approximately 5% of phages were dynamically gained or lost from persistent bacterial hosts. Within a sample, we found that bacterial hosts with and without a given prophage coexisted simultaneously. Furthermore, phage induction, when detected, occurred predominantly at low levels (1–3× coverage compared to the host region), in line with theoretical expectations⁴. We identified multiple instances of integration of the same phage into bacteria of different taxonomic families, challenging the dogma that phages are specific to a host of a given species or strain⁵. Finally, we describe a new class of ‘IScream phages’, which co-opt bacterial IS30 transposases to mediate their mobilization, representing a previously unrecognized form of phage domestication of selfish bacterial elements. Taken together, these findings illuminate fundamental aspects of phage–bacterial dynamics in the human gut microbiome and expand our understanding of the evolutionary mechanisms that drive horizontal gene transfer and microbial genome plasticity.

Bacteriophages are the most abundant biological entity on Earth, playing crucial roles in shaping microbial communities. Most phages are either lytic or are integrated into their bacterial host’s DNA (prophages), forming a bacterial lysogen. In-depth studies of several model integrated phages (such as λ and Mu) have built the framework for our understanding of lysogens and established foundational principles in molecular biology¹. Recent advances in metagenomic sequencing and improved analytical tools have shown that lysogens are much more diverse, prevalent and abundant in human microbiomes than previously recognized^{6–10}.

Most human gut phages are prophages¹¹, although alternative lifestyles, such as phage plasmids or carrier states, also exist^{12–14}. Typically, prophages insert their genome into the host chromosome using dedicated integrases of three different classes: tyrosine recombinases, small and large serine recombinases and DDE recombinases. By integrating into the host chromosome, phages ensure replication and vertical transfer within their hosts when conditions are not optimal for lytic replication. Hosts may derive a fitness benefit from prophage-encoded genes, such as antibiotic resistance or virulence factors^{2,3}, but run the risk that prophages may re-enter the lytic cycle, killing their hosts. Prophages can accumulate mutations over time, leading to decay of mobilization machinery¹⁵, whereas other prophage genes may be adopted and maintained by their bacterial host¹⁶. Similarly, prophages may incorrectly package host genes upon induction, thereby mediating horizontal gene transfer between related bacterial

strains or species through processes such as generalized and specialized transduction¹⁷.

Prophages can profoundly impact their bacterial host and the microbial community at large. Yet, it remains challenging to study phages and hosts within the same sample. Distinguishing between phage and bacterial regions of the chromosome has been difficult, prompting researchers to use virus-like particle (VLP) sequencing to enrich for viral sequences¹⁸. Using this method, studies have suggested that phages in the gut are relatively stable, persisting within their human superhost over 1 year^{19,20}. VLP-based studies have informed our knowledge of phage abundance and stability in the human gut microbiome, but this technique has two notable limitations: focusing on VLPs (1) overlooks prophages that are not actively producing phage particles at the time of sampling; and (2) there is little direct evidence of which prokaryotic host(s) each phage infects. Bulk metagenomics, in which viral and prokaryotic genomes are sequenced simultaneously, slightly addresses these challenges, given new phage prediction tools^{21–24}. However, most bulk studies are performed with short-read sequencing, which cannot resolve repetitive elements, posing many challenges when studying phages derived from complex communities. For example, phage genomes can have large regions of similarity (phage genomic mosaicism²⁵) leading to fragmented assemblies and incomplete genomes. Similarly, if a given prophage is integrated into multiple hosts, short-read assembly would be unable to resolve its hosts, complicating further investigation.

¹Division of Hematology, Department of Medicine, Stanford University, Stanford, CA, USA. ²Department of Genetics, Stanford University, Stanford, CA, USA. ³Department of Bioengineering, Stanford University, Stanford, CA, USA. ⁴Department of Biology, Stanford University, Stanford, CA, USA. ⁵Department of Microbiology and Immunology, Stanford University, Stanford, CA, USA.

⁶These authors contributed equally: Jakob Wirbel, Angela S. Hickey. ✉e-mail: asbhatt@stanford.edu

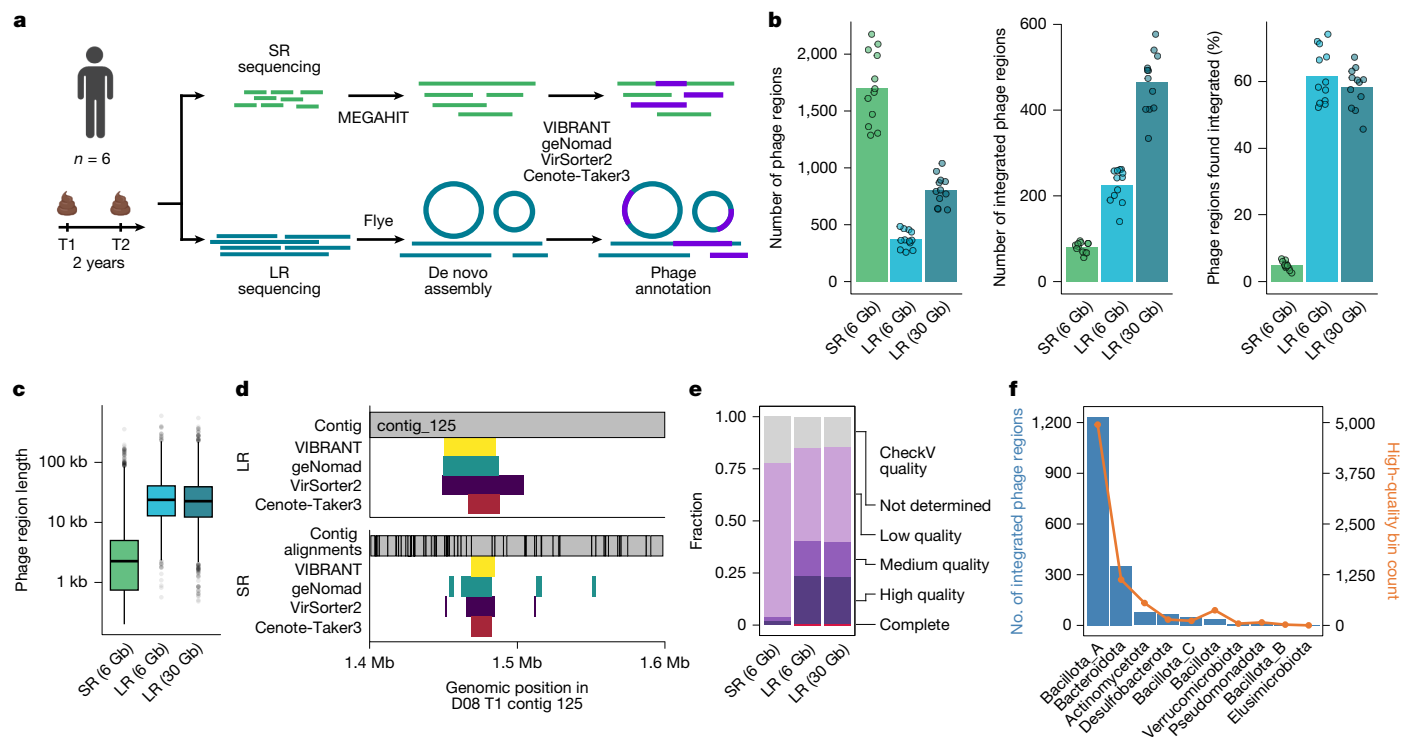


Fig. 1 | Improved detection of integrated prophages with long-read sequencing. **a**, Schematic of the analysis workflow. DNA from stool samples collected approximately 2 years apart from six healthy individuals was sequenced using both short-read (Illumina) and long-read (ONT) platforms. MEGAHIT and metaFlye were used to assemble the short and long reads, respectively. Phages were then predicted from both assemblies using VIBRANT, geNomad, VirSorter2 and Cenote-Taker3. LR, long read; SR, short read. **b**, Number of phage regions, number of integrated phage regions and percentage of phage regions found to be integrated, as predicted by geNomad for the assemblies from short-read (6 Gb; $n = 12$), subsampled long-read (6 Gb; $n = 12$) and deep long-read (30 Gb; $n = 12$) sequencing. Bars represent the mean across all samples, and points indicate values for individual samples. **c**, Lengths of phage regions are shown for short-read (6 Gb; $n = 20,467$), subsampled long-read (6 Gb; $n = 4,442$) and deep long-read (30 Gb; $n = 9,652$)

Long-read metagenomic sequencing can address these challenges by yielding more contiguous assemblies, potentially resolving individual phage genomes and the hosts of integrated prophages. Previous studies have demonstrated advantages of long-read sequencing for the study of phages. One study used long-read VLP sequencing to resolve more complete viral genomes and identify structural variations (SVs) within phages^{6,26}. Another study reported improved prophage and CRISPR spacer assembly with long-read sequencing, capturing individualized phage populations that were stable over a 10-day period²⁷. Although this has enabled host prediction from complex communities, suggesting extremely broad host range for some phages^{6,7,28}, CRISPR spacers are typically very short (20–50 nucleotides), which may lead to an overestimation of host range²⁹. Accurate assembly of prophages into their native host through long-read assembly would provide direct evidence of infection, enabling the study of specific phage–host relationships.

Here we sought to examine the relationship between prophages and their hosts in the human gut using long-read sequencing. To do so, we generated a deep, long-read, longitudinal metagenomic dataset from stool samples collected from six individuals over a 2-year period. This enabled us to study the dynamics of integrated prophages in multiple individuals across longer timescales than have been previously reported. Although we found a small fraction of phages to be gained

or lost over 2 years, most phages seem to be stably integrated into their bacterial hosts. Prophages appear to induce at predominantly low levels, with some phages and their hosts existing in mixed populations (induced phage, integrated phage and host without phage) at the same time. A small number of phages assemble into multiple and taxonomically diverse host contexts, providing strong evidence of a broad host range. Unexpectedly, we also identified a group of related prophages (IScream phages), which do not encode their own canonical integrase but instead have probably domesticated bacterial insertion sequence (IS) elements for their integration and excision machinery. In this study, we demonstrated how longitudinal long-read metagenomics can help elucidate diverse aspects of prophage biology in human microbiomes.

Longitudinal long-read metagenomics

To learn more about the biology of integrated phages and their activity over time, we collected stool samples from six healthy adults at two time points, spaced 2 years apart (T1 and T2). We then generated long-read metagenomic DNA sequencing data from these samples on the Oxford Nanopore Technologies (ONT) platform to a depth of approximately 30 billion bases (Gb) (Fig. 1a, Extended Data Fig. 1 and Supplementary Table 1). All samples were also sequenced using short-read shotgun

sequencing using the Illumina platform to a depth of 6 Gb. To enable comparison between short-read and long-read sequencing without the bias of different sequencing depths, we subsampled our long-read data to the mean of the short-read data (Methods). Following quality control and host-read removal, short reads were assembled using MEGAHIT³⁰ and long reads using metaFlye³¹. Both assemblies were subsequently binned into metagenome-assembled genomes (MAGs) (Methods). As expected, long-read assemblies exhibited higher contiguity with a higher mean contig N50 (255.5 kb for long reads versus 7.8 kb for short reads) and fewer contigs within corresponding high-quality bins ($P = 2.2 \times 10^{-16}$; paired Wilcoxon test; Extended Data Fig. 1). By contrast to gene-calling challenges from previous ONT chemistries³², the average length of predicted genes was similar across both types of assemblies (Extended Data Fig. 1). This demonstrates that ONT reads generated with the latest chemistry (R10.4; median read quality of approximately Q20) can yield accurate assemblies without the need for short-read polishing³³. Taken together, our long-read assemblies exhibited much higher contiguity than the short-read-based assemblies, without sacrificing quality.

Prophages in long-read assemblies

To compare phages in both short-read and long-read assemblies, we used the computational phage prediction tool geNomad²³, which combines alignment-free and gene-based models for viral prediction. Across all samples, geNomad found more phage regions in the short-read assemblies than in the downsampled long-read assemblies. However, the percentage of integrated phages was lower (approximately 5% in short-read assemblies versus approximately 60% in long-read assemblies; Fig. 1b). Finding a majority fraction of phages in the long-read assemblies to be integrated is consistent with theoretical expectations¹¹. These results were recapitulated using phage annotations from three other commonly used prediction tools: VIBRANT²¹, VirSorter2 (ref. 22) and Cenote-Taker3 (ref. 24) (Extended Data Fig. 2).

The average length of predicted phage regions was much smaller in short-read than in long-read assemblies (Fig. 1c); therefore, we considered that a single phage region predicted in the long-read assemblies might be represented by multiple smaller phage regions in the short-read assembly. We mapped the short-read contigs to the long-read assemblies and compared phage annotations across sequencing methods. Figure 1d shows the phage predictions and alignments of short-read contigs against a single long-read contig containing a 50-kb region that is predicted by geNomad, VIBRANT, VirSorter2 and Cenote-Taker3 to be an integrated prophage. This region recruits alignments from multiple phage contigs in the short-read assembly but also from contigs not predicted to be phage, indicating likely mis-annotations (Extended Data Fig. 3). Overall, the majority (82%) of long-read integrated phages were found to be fragmented in the corresponding short-read assemblies (Extended Data Fig. 3 and Methods). Consistent with this, a higher fraction of long-read phages was classified as medium or higher quality by CheckV³⁴, which assesses genome completeness and the presence of phage hallmark genes (Fig. 1e). These results demonstrate the superiority of long-read metagenomics for the accurate and complete assembly of integrated phages.

In the following sections, we focus on deep (30 Gb) versus downsampled (6 Gb) long-read sequencing assemblies as we detected more phage regions with the same quality at this depth (Fig. 1b,c,e).

Long-read sequencing allowed us to assemble the majority of phages as integrated elements into their bacterial hosts; therefore, we could directly determine the taxonomy of the bacterial hosts of each of the integrated phages from our assemblies. To do so, we used taxonomic assignments of the genes in the surrounding host region (Methods). We found a high concordance of host assignment between this approach and existing host prediction methods³⁵ (approximately 90% agreement up to the family level; Extended Data Fig. 4). To ensure accurate

host identification, we focused on phages with agreement between gene-based and high-quality binning host taxonomy. The majority of integrated phages were found in the phyla Bacillota A and Bacteroidota, which reflect the number of high-quality bins present across samples (Fig. 1f). In summary, long-read metagenomic sequencing and assembly improved the detection of integrated prophages and their hosts.

Most prophage induction rates are low

With samples collected from the same individuals over a 2-year interval, we aimed to explore whether phages were acquired or lost independently of their bacterial hosts during this period. In this analysis, we also investigated whether phages could be found in different genetic loci within the same or in different hosts. To compare identical phages over time, we clustered all integrated phages with sufficient coverage (median coverage higher than ten) from the same individual, taking their surrounding host region into account (Extended Data Fig. 5 and Methods). We classified phage clusters on the basis of their assembly: overall, 35% of phage clusters were difficult to classify because of fragmented assemblies or had phages assembled at the ends of contigs, and 13% of the clustered phage genomes were assembled within their host genome in a single time point only (Fig. 2a). The remaining 52% of clustered phage genomes were found in hosts that were assembled in both time points. Of those, 90% of the clustered phage genomes were found in the same position within their bacterial hosts (host context) over the 2-year period. More rarely (approximately 5%), clustered phage genomes were found to be dynamic across the 2 years, meaning phages were lost or gained while the host was present at both time points. Examples of lost or gained phages between time points in an otherwise stable bacterial host are provided as sequencing coverage plots (Fig. 2b,c). Some of these dynamic phages could be the result of a new infection or loss events within the same strain. However, we also found dynamic phages within bacterial species, in which different strains with low shared average nucleotide identity (ANI) (less than 99%; Extended Data Fig. 5) exist at the two time points. In these cases, bacterial strain replacement is probably causing the detection of a 'dynamic' phage.

On the basis of the observed drops in read coverage for dynamic phages, we reasoned that read-level evidence could reveal the exact boundaries of integrated phages. To quantify this systematically, we used Sniffles2 (ref. 36) to identify SVs from reads mapped across time points (Methods). With the exact boundaries predicted by Sniffles2, we could quantify the prediction error for all included phage prediction tools (Extended Data Fig. 6). We observed deletion SVs in dynamic phages and in those assembled as stably integrated. These results indicate the presence of both hosts with an integrated phage (lysogens) and hosts without an integrated phage (naïve hosts) at the same time. In total, we found that naïve hosts coexisted with lysogens in approximately 7% of cases, made apparent by incomplete drops in coverage (Fig. 2d). We then quantified the proportion of reads supporting the presence of naïve hosts and observed a wide distribution of values for this fraction, indicating that prophage prevalence within a population can be highly heterogeneous (Fig. 2d). In a subset of SV-overlapping phages (approximately 40%), we could detect the deletion SV (evidence for the naïve host) only through reads from the other time point, meaning that the proportion of naïve hosts within the population changed over time. Using the exact boundaries detected this way, we often observed a small fraction of reads supporting the existence of naïve hosts in the original time point, despite falling below Sniffles2 detection thresholds (Fig. 2d). This indicates that heterogeneous prophage prevalence, in which lysogens and naïve bacterial hosts can coexist, may be more common than we can detect.

In addition to the detection of naïve hosts, SV calling also resulted in the detection of circular phage genomes. Circularization of phage genomes can occur during prophage induction, allowing us to measure

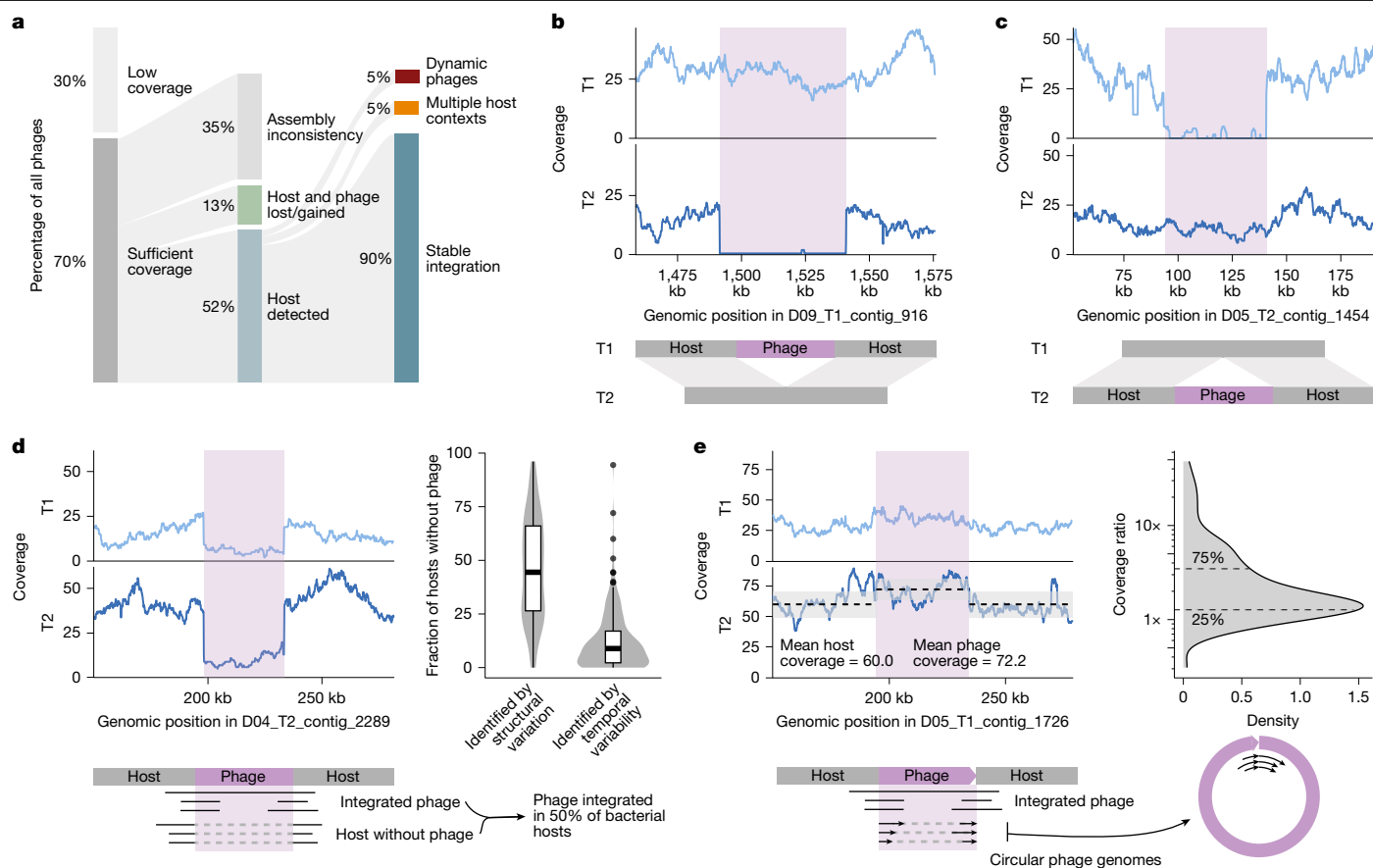


Fig. 2 | Phage dynamics and population heterogeneity in the gut. **a**, Alluvial (flow) plot showing the percentage of phage clusters and their classification (Extended Data Fig. 5) across our classification pipeline. **b**, Coverage plot supporting the loss of an integrated phage in T2 for individual D09. **c**, Coverage plot supporting the gain of an integrated phage in T2 into a host that is already present in T1 for individual D05. **d**, Left, coverage plot showing population heterogeneity in terms of integration for a prophage in individual D04. The fraction of hosts without phage was calculated based on the reads supporting the integrated phage or the host without phage (see schematic below). Right, this fraction is shown as a box plot for all phages with SV evidence in the same

time point ($n = 360$) and for all phages with SV evidence in the other time point ($n = 209$) (identified by temporal variability). Box plots show the IQRs as boxes, with the median as a black horizontal line, whiskers extending up to the most extreme points within 1.5-fold IQR and outliers represented as dots. **e**, Left, coverage plot for a phage with SV evidence for circular phage genomes (see schematic below). The mean coverage of the phage and host region is indicated by dashed black lines (the grey area shows the mean $\pm$ 1 s.d.). Right, the coverage ratio between the phage and the surrounding host region is shown as a density plot, with 25th and 75th percentiles indicated by dashed black lines.

the induction rate of some integrated phages in situ. In most cases, the presence of circular phage genomes was not concomitant with an appreciable increase in genome coverage compared to the surrounding host region (Fig. 2e). Instead, we observed the majority of induced phages to be present at 1–3 $\times$ coverage compared to the surrounding host region. This is consistent with low-level phage induction as opposed to large lytic replication bursts (greater than 10 $\times$), the latter of which seems rare in the gut⁴. Some phages with evidence of circular genomes exhibited coverage ratios lower than 1 $\times$ compared to the host region, which could be explained by the simultaneous presence of lysogens, naive hosts and induced circular phage genomes (Extended Data Fig. 6).

Overall, our data indicate that integrated phages are relatively stable in their hosts over a 2-year time frame, with few cases of new phage integration or loss. Additionally, both phage integration and phage induction are subject to substantial population-level heterogeneity in the gut, with low average rates of phage induction.

Evidence for broad prophage host range

In addition to detecting dynamic phages, we were able to find a small percentage of phage clusters (approximately 5%) that were assembled

in multiple host contexts within and across time points (Fig. 2a and Extended Data Fig. 7). Although some well-studied phages (P1 (ref. 37), PR5 (ref. 38) and PRD1 (ref. 39)) have been described to infect bacteria of different genera, computational predictions for gut phages have suggested that even broader host ranges may exist^{7–9}. These predictions typically rely on CRISPR spacer analysis, which may overestimate the true host range owing to the short spacer length and modularity of phage genomes.

Only a few phages within the same individual were present in multiple host contexts (median $n = 11$); therefore, we expanded our analysis to assess the host range of closely related phages. We used standard clustering cutoffs³⁴ for viral species of 95% identity and 85% coverage to cluster all phages across all individuals. Using our high-confidence host identification approach, we determined the broadest host taxonomic level that was shared within a phage cluster. As expected, the majority of phages were restricted to the species level (78%), with only under 20% being restricted to the genus level (Fig. 3c). We found evidence for a small number of phages that demonstrate broader host range; 11 phage species were restricted to the family level, and eight were restricted to the order level, most of which belong to the order Bacteroidales (Supplementary Table 2). A single example of a phage species restricted to the class level was insufficiently supported by read

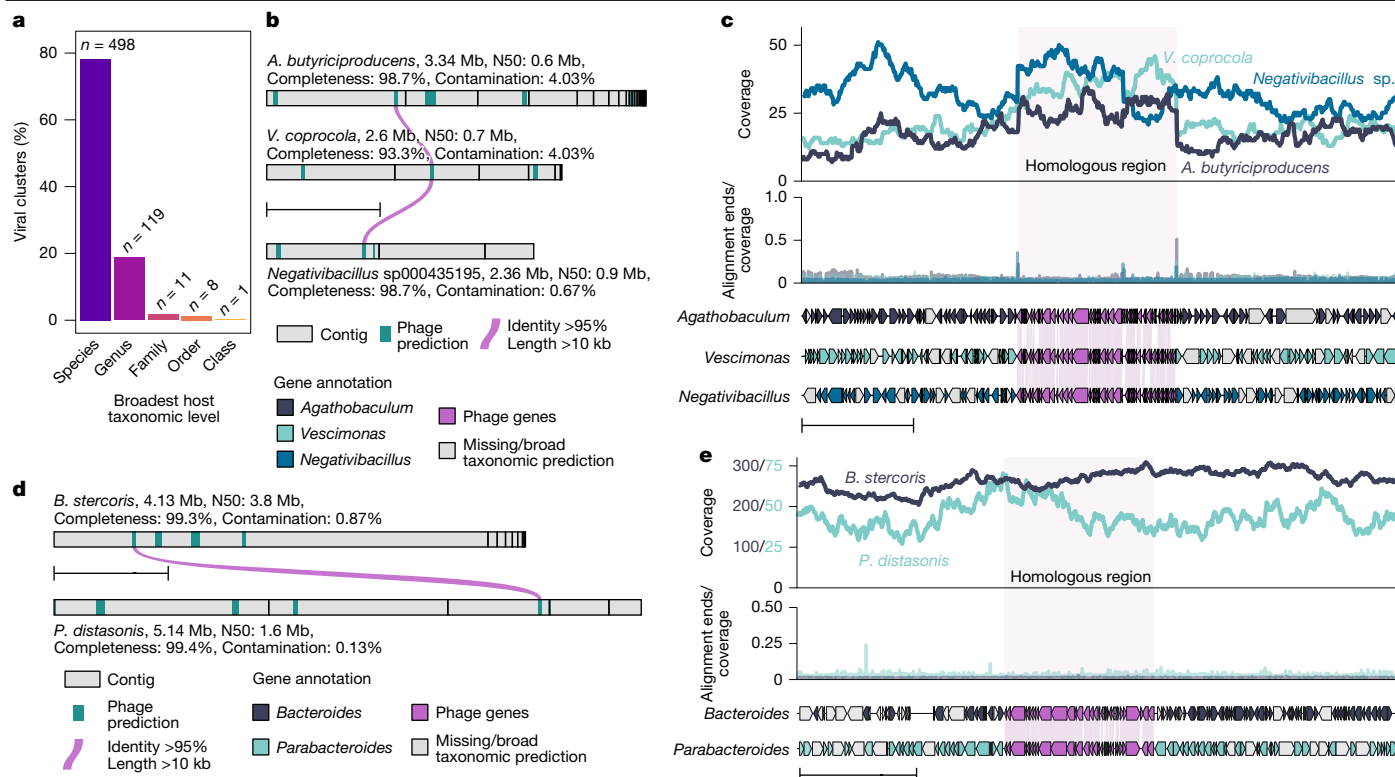


Fig. 3 | Long-read assemblies provide evidence for broad host range for integrated phages. **a**, Percentage of all viral clusters with members that are restricted to the species, genus, family, order or class taxonomic level. The number of clusters is noted above each bar. **b**, Chart representing high-quality MAGs for three different species, annotated with phage predictions from geNomad. A single genomic region (greater than 10 kb) is shared between all three genomes with more than 95% nucleotide identity, annotated by a pink link. **c**, Coverage and synteny plot for the annotated region from **b**. The top panel shows read coverage, the middle panel shows the number of alignment

ends divided by coverage and the bottom panel shows the genes (as arrows) in the three genomes. Genes are coloured according to their taxonomic predictions. **d**, Chart representing two high-quality MAGs, with a single region shared between them. **e**, Coverage and synteny plot for the annotated region from **d**. The top panel shows read coverage, the middle panel shows the number of alignment ends divided by coverage and the bottom panel shows the genes (as arrows) in the three genomes. Genes are coloured according to their taxonomic prediction. The shaded regions in **c** and **e** indicate amino acid similarity greater than 80%. Scale bars, 1 Mb (**b,d**), 25 kb (**c,e**).

coverage (Extended Data Fig. 8). One phage species was integrated into hosts annotated as *Vescimonas coprocola* (family Oscillospiraceae), *Negativibacillus* sp. (family Ruminococcaceae) and *Agathobaculum butyriciproducens* (family Butyricicoccaceae) (Fig. 3b). The integrated prophage was the only region with more than 95% nucleotide identity and a length greater than 10 kb that was shared across the three high-quality genome bins. The taxonomic annotation of each host was also supported by individual gene predictions consistent across the host contig harbouring the phage (Fig. 3c). Although we observed strong evidence of some phages integrated into different bacterial families, we wondered whether some of these cases might be the result of mis-assemblies. One pattern that is typically observed in a mis-assembly is that the majority of read alignments end abruptly at the point of mis-assembly, as opposed to alignments spanning the junction⁴⁰. To rule out mis-assembly of phage integration, we evaluated the alignment ends normalized to coverage across the junction of the host and phage genome (Fig. 3c). For this specific example, we observed a moderate concentration of alignment ends piling up at the boundaries of the phage region. These peaks resulted from reads supporting the presence of circular phage genomes, indicating that this phage species may be capable of independent replication across disparate bacterial families (Supplementary Table 2).

We also detected another phage species that was well supported, by high coverage and the lack of alignment end accumulation, to have bacterial hosts in organisms from different families but in the same order: *Parabacteroides distasonis* (family Tannerellaceae) and *Bacteroides stercoris* (family Bacteroidaceae) (Fig. 3d,e). We manually inspected

the other order-restricted phage species (Extended Data Fig. 8). In four cases, the phage regions were annotated at the beginning of a contig, preventing the validation of phage integration by having bacterial host genomic fragments on both ends of the phage genome. However, an orthogonal assembler (myloasm⁴¹) confirmed the integration of the phages into the correct context in two of the four cases (Extended Data Fig. 8 and Supplementary Table 2). The final two phage species were part of a larger region shared between Tannerellaceae and Bacteroidaceae and potentially involved in recombination (Extended Data Fig. 8).

Taken together, these examples provide strong assembly-level evidence for a broad host range of some phages, mostly infecting hosts of the order Bacteroidales.

IScream phages

Prophages typically carry enzymes to enable their genomic integration into host DNA (through integrase). When examining the set of prophages with exact genome boundaries from SV evidence ($n = 569$), we found tyrosine and serine integrases to be most common (Fig. 4a). DDE-type integrases, similar to the integrase used by phage Mu⁴², were found predominantly in phages with another integrase, suggesting that these DDE enzymes represent bacterial ISs (mobile, selfish genetic elements⁴³ similar to the integrase of Mu) instead of genuine phage mobilization machinery.

Among 101 phages lacking an identified integrase, we found 22 that contained IS30 family elements on both ends of the prophage genome. This organization is reminiscent of composite transposons,

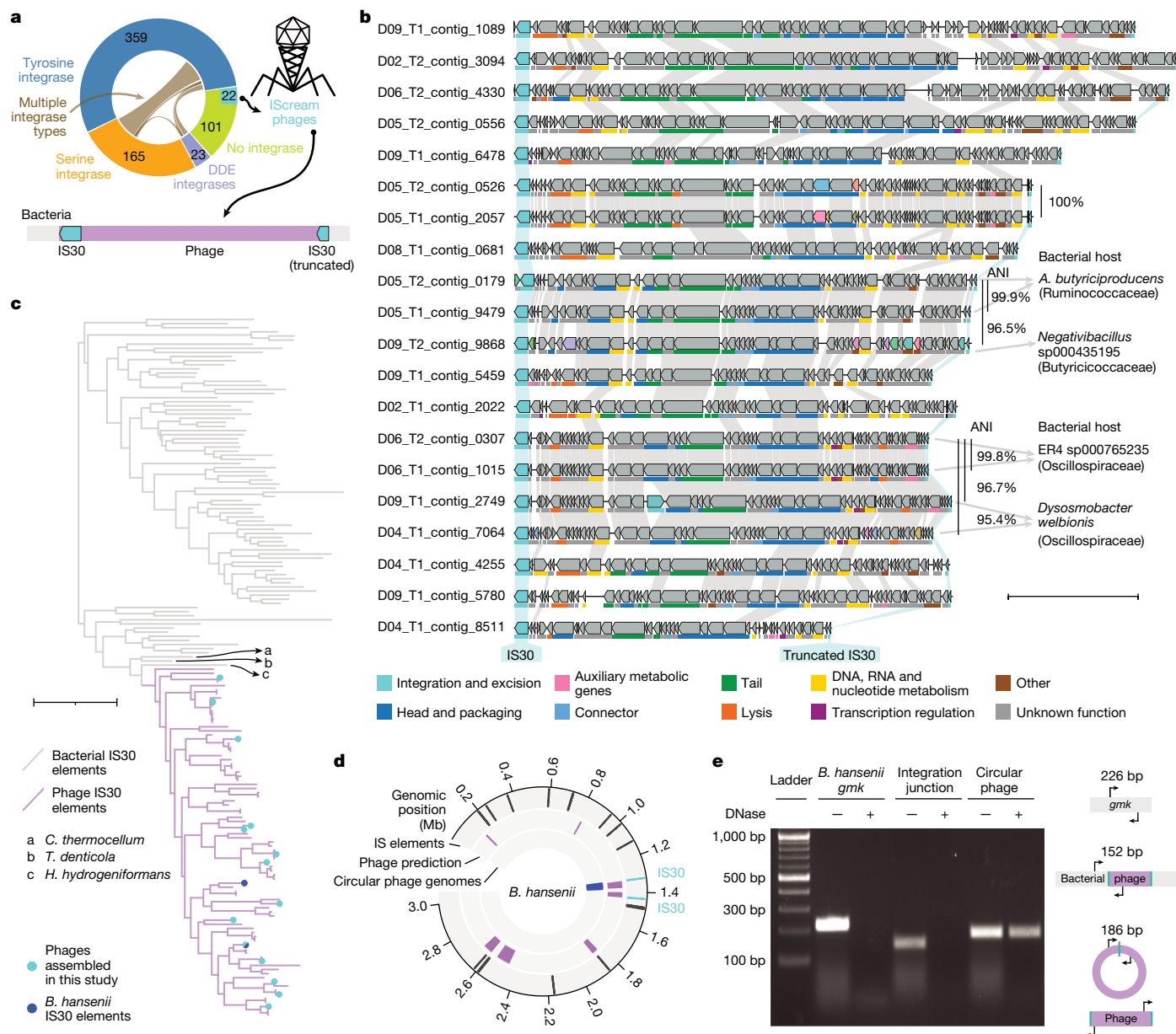


Fig. 4 | Description of the new IScream phage group. **a**, Doughnut plot showing the number of phages with different integrase enzymes. Only phages with SV evidence were included. Of phages without annotated integrases, 22 contained genes annotated as IS30 transposases at both ends (schematic shown below). We named these phages IScream phages. **b**, Genome organization of IScream phages (length > 10 kb) assembled here, visualized with LoVis4u. Gene functionality (annotated as coloured bars below genes) was inferred using pharokka; truncated IS30 elements were manually confirmed. Genes are connected across genomes if the predicted proteins have higher than 25% amino acid identity. Two clusters of closely related IScream phages are annotated with their ANI and the taxonomic classification of their assembled host context on the right side. **c**, Tree showing the relationship between bona fide bacterial IS30 transposases from the ISfinder database (in grey) and IScream phage outward-directed IS30 (purple) identified in the MGv catalogue.

a type of mobile genetic element in which two IS elements flank a gene cassette, encoding, for example, proteins conferring antibiotic resistance⁴³. We considered that these phages are mobilized similar to composite transposons through the IS30 within their genomes. Other phages, such as Mu, are known to be mobilized through a transposase⁴². Here we describe a new group of phages that probably co-opted bacterial IS30 elements for their mobilization. Because

All IS30 transposases were clustered at 70% amino acid similarity before multiple sequence alignment and tree construction. Clusters containing an IS30 from any of the IScream phages assembled here are annotated by cyan dots. **d**, Circos plot for the *B. henselii* ATCC 27552 genome. The outer ring indicates the location of the IS elements, the middle ring shows the predicted integrated phages and the innermost ring shows the location of the circular phage genomes detected by the presence of SV. **e**, Image of a 2% agarose gel showing various PCR products from the *B. henselii* culture with and without DNase treatment. The primer location and expected product size are schematically annotated on the right side. The experiment was repeated three independent times, yielding similar results. Scale bars, 10 kb (**b**), 1 (**c**). Illustration in **a** adapted from SVG Repo (<https://www.svgrepo.com/>) under a CC0 1.0 Universal Public Domain licence.

these phage genomes are ‘sandwiched’ by IS30 elements, we named them IScream phages.

To explore the IScream phages in more detail, we first focused on genome organization and observed high synteny across the full-length IS30-bound phage genomes (Fig. 4b). Genes for core phage lifestyle functions, such as structural proteins or host lysis, seem to be present in all IScream phages (Extended Data Fig. 9). Two highly similar clusters

of IScream phages (ANI > 95%) are present across several individuals, integrated into different bacterial host contexts. In fact, one of the order-restricted broad host range phage clusters within the class Clostridia was identified as an IScream phage (Fig. 4b and Supplementary Table 2). Because we observed only a small number of IScream phages in our data, we screened the Metagenomic Gut Virus (MGV) catalogue⁸ and found 1,780 potential IS30-bound phages, revealing that these phages are abundant and prevalent gut residents (Extended Data Fig. 9).

Focusing more on the IS30 transposase proteins potentially used for phage integration, we clustered all IS30 proteins in the IScream phages assembled here. We observed that the IS30 proteins directed outwards of the integrated phage genome to be relatively conserved (mean amino acid identity = 52%), whereas the IS30 proteins on the other side were highly variable in length (183–4,175 nucleotides) and typically truncated or fused to other protein domains, for example, domains involved in conjugation (TraX) or defense against restriction (DarA). These deprecated IS30 proteins also typically lack a classical IS30 catalytic domain (Extended Data Fig. 10), suggesting that the outward-directed IS30 functionally catalyses phage mobilization. This is again similar to composite transposons because one of the IS elements in typical composite transposons can lose its catalytic activity and decay over time⁴⁴.

To explore the potential evolutionary history of IS30 domestication by phages, we clustered the full length, outwardly directed IS30 proteins across the phages identified in our dataset, all IS-bound phages in MGV and all bacterial IS30 elements annotated in the ISfinder database⁴⁵. We found all phage-derived IS30 proteins to form a sub-clade most closely related to the IS30 elements in *Clostridium thermocellum*, *Treponema denticola* and *Halanaerobium hydrogeniformans* (Fig. 4c). We additionally found IS30 proteins on phages in meso-American paleofeces⁴⁶ that clustered together with other phage IS30 proteins (Extended Data Fig. 10). This suggests that the IScream phages in our data and from a paleofeces sample resulted from a single domestication event that potentially occurred in a host related to one of the three species mentioned above.

To gather further evidence that the IS-bound phages represent bona fide phages that can form virions rather than cryptic prophages or other selfish non-phage elements, we analysed publicly available VLP-enriched sequencing data from neonatal gut samples⁴⁷ using Phanta, a virus-inclusive read-level profiling method⁴⁸. We found higher abundance of some MGV-derived IScream phages in VLP rather than in metagenomic shotgun (metaG) sequencing (Extended Data Fig. 9), suggesting viral particle production. We additionally screened *Clostridia* genomes (Methods) and found two phages flanked by IS30 elements in *Blautia hansenii* American Type Culture Collection (ATCC) 27552 (Fig. 4d). We obtained and cultured this strain of *B. hansenii* and performed long-read sequencing of the overnight culture. Consistent with our previous results showing a low level of induction for integrated prophages, we observed circular phage genomes in our overnight culture, indicating spontaneous induction of the phage (Fig. 4d and Extended Data Fig. 10). Phages were isolated by means of polyethylene glycol (PEG) precipitation (Methods). Using polymerase chain reaction (PCR), we found the expected circular phage genome in the PEG precipitate (Fig. 4e). This circular version of the genome, but not the integrated version or the bacterial genome, was protected from DNase treatment, consistent with the presence of phage particles in the culture, in which the capsid protects the phage genome from DNase exposure (Fig. 4e). To visualize the potential phage particles, we used transmission electron microscopy to image lysate from the *B. hansenii* culture (Methods). We identified phage-like particles (Extended Data Fig. 10), which we suspect represent the IScream phage in question because our long-read sequencing data indicated that this phage was the only predicted phage region within *B. hansenii* showing evidence of spontaneous induction (Extended Data Fig. 10).

Together, these findings indicate that IScream phages have co-opted bacterial IS30 transposases for phage mobilization while retaining phage activity, highlighting a previously overlooked group of gut-resident phages.

Discussion

Decades of phage research have yielded fundamental insights into molecular biology, genetics and evolution. Many studies focused on a select set of culturable phages have informed the ‘central tenets’ of phage biology^{1,49}. Although valuable, it is unclear how generalizable these core principles are. Here we generated a deep long-read longitudinal metagenomic dataset to derive a more complete understanding of prophage–host interactions within the human gut microbiome. Our findings revealed that (1) prophages are usually stable within their hosts over a 2-year time frame; (2) lysogens and naive hosts can coexist within a population at varying proportions; (3) phage induction, when observed, occurs at predominantly low levels; (4) phages can integrate into bacterial hosts of different families; and (5) some prophages might have domesticated ISs as integrases for integration and excision.

Previous studies using short-read VLP sequencing^{19,20} or bulk metagenomic sequencing over shorter time frames (10 days²⁷) have reported generally high temporal stability and individuality of the virome. Using bulk long-read metagenomic sequencing and a longer sampling period of 2 years, we observed most prophages to be stably integrated, consistent with previous stability estimates. Only rarely do we infer possible new phage infection events. For example, we observed an identical phage present in two different strains of *Alisipites putredinis* at T1 and T2 from the same individual (Extended Data Fig. 7). Furthermore, for stably integrated prophages, we commonly observed population heterogeneity in terms of integration, in which lysogens and naive hosts coexist.

In a small subset of phages, we detected phage induction through read-level evidence. Within this population, we found that low-level induction is more common than large burst events, consistent with short-read-based surveys and ecological models^{4,50}. In some cases, we found examples in which lysogens, naive hosts and induced circular phage genomes coexist (Extended Data Fig. 6). This could reflect strain-level variation that limits the ability of phages to infect the entire host population, or geographic separation of subpopulations. Mechanisms such as phase variation of surface structures have also been reported to modulate phage susceptibility, providing a regenerating susceptible subpopulation^{13,51}. Because the spatial distribution of induction in the population is unknown, it is difficult to determine if induction is a sporadic or somehow coordinated event⁵². Taken together, these findings are consistent with a previously proposed ecological model, in which phages spread slowly through a susceptible bacterial population¹³, potentially constrained by spatial separation⁵³. In addition to exploring the presence of a susceptible subpopulation of hosts, low-level phage induction might serve as a successful strategy for prophages to prevent mutation accumulation and deterioration into cryptic prophages¹⁵.

Most cultured phages are thought to have a narrow host range, as determined by plaque-based assays⁵. Although some phages can infect hosts within the same genus^{37–39}, some are restricted to the species or strain level⁵⁴. Here we provide direct evidence supporting the existence of phages with broad lysogenic host range, as we found them integrated into bacterial hosts from different families. A recent study by Bignaud et al.⁵⁵, which used metagenomic Hi-C data, supports these findings and demonstrated that broad host range is more common than previously recognized. Long reads may allow us to investigate this at scale in future studies. Lysogenic host range does not necessarily imply productive host range (the ability to produce infective particles from several hosts)⁵ because we found evidence for circular phage genomes in a single case only (Fig. 3b). Many potential determinants may contribute to

broad host range, including prophage-encoded diversity-generating retroelements⁵⁶, inversions⁵⁷ and polyG tracts⁵⁸, although they have not yet been associated with host range of this breadth. The ability of these phages to infect several gut residents may have implications for phage therapy that is being explored as a promising alternative to antibiotics⁵⁹. Although many of these approaches leverage the narrow host range of cultured phages, evaluating host range through isolation and culturing may limit our ability to detect its breadth⁶⁰. The long-read metagenomic analysis we presented here allowed us to better capture the landscape of the host range of gut resident phages.

Finally, we describe the new group of IScream phages, which probably use IS30 transposases of bacterial origin for their mobilization. A previous study had suggested that the distinction between site-specific recombination and transposases might not be well defined⁶¹. Kiss et al.⁶¹ created a synthetic system in which they replaced the integrase of phage λ with an *Escherichia coli* IS30 and observed that the IS30 transposase was sufficient to create a functional phage. Although IS30 transposases use DDE chemistry similar to that of the transposable phage Mu⁴², here we provide evidence that IS30 may act as recombination machinery for natural phages. We propose that this IS30 was domesticated from a host IS30 element from the class Clostridia. Although it is known that bacteria can domesticate the genes of phages¹⁶, these findings provide an example for how phages might have domesticated originally selfish genetic elements for their own purpose.

Although we were able to expand our understanding of prophage dynamics in the human gut microbiome through long-read sequencing, this study has several limitations. First, our samples were limited to six individuals from a shared geographic area; therefore, our findings may not be generalizable. Second, detection of prophages and their hosts is necessarily linked to sequencing depth. Despite our deep sequencing, we have not yet exhausted the diversity of the microbiome and were unable to detect low-abundance organisms. Third, although de novo assembly allows for reference-free investigation of microbial communities, this approach is not free from errors⁴⁰ and potentially collapses population heterogeneity within a sample. Careful analyses of read-level evidence are therefore needed to support assembly-level claims and quantify the presence of mixed populations. Wherever possible, we have used direct read alignment-based approaches to orthogonally validate results derived from assemblies. Finally, phage annotation, host taxonomic classification and SV calling are subject to error. For example, although dynamic phages are enriched for SV calls (37% versus 11% for stably integrated phages), not all of them were found by Sniffles2 owing to internal filtering steps to ensure specificity.

Conclusion

Integrated prophages play fundamental roles in the human gut. They are prevalent and abundant entities that outnumber lytic phages and represent untapped potential for molecular tool development or therapeutic interventions beyond antibiotics. In this study, we revealed key aspects of prophage biology through long-read sequencing, highlighting phage integration dynamics, population heterogeneity, host range and a new group of IS-bound phages. We anticipate that future studies will further elucidate the impact of phage-mediated horizontal gene transfer in the gut¹⁷, characterize the specificity of phage integrases for biotechnological applications and continue to explore how the evolutionary flux between phages and bacteria may lead to genomic innovation.

Online content

Any methods, additional references, Nature Portfolio 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/s41586-025-09786-2>.

- Salmond, G. P. C. & Fineran, P. C. A century of the phage: past, present and future. *Nat. Rev. Microbiol.* **13**, 777–786 (2015).
- Howard-Varona, C., Hargreaves, K. R., Abedon, S. T. & Sullivan, M. B. Lysogeny in nature: mechanisms, impact and ecology of temperate phages. *ISME J.* **11**, 1511–1520 (2017).
- Faruque, S. M. & Mekalanos, J. J. Phage-bacterial interactions in the evolution of toxigenic *Vibrio cholerae*. *Virulence* **3**, 556–565 (2012).
- Lopez, J. A. et al. Abundance measurements reveal the balance between lysis and lysogeny in the human gut microbiome. *Curr. Biol.* **35**, 2282–2294 (2025).
- Hyman, P. & Abedon, S. Bacteriophage host range and bacterial resistance. *Adv. Appl. Microbiol.* **70**, 217–248 (2010).
- Chen, J. et al. Efficient recovery of complete gut viral genomes by combined short- and long-read sequencing. *Adv. Sci.* **11**, e2305818 (2024).
- Camarillo-Guerrero, L. F., Almeida, A., Rangel-Pineros, G., Finn, R. D. & Lawley, T. D. Massive expansion of human gut bacteriophage diversity. *Cell* **184**, 1098–1109 (2021).
- Nayfach, S. et al. Metagenomic compendium of 189,680 DNA viruses from the human gut microbiome. *Nat. Microbiol.* **6**, 960–970 (2021).
- Tisza, M. J. & Buck, C. B. A catalog of tens of thousands of viruses from human metagenomes reveals hidden associations with chronic diseases. *Proc. Natl Acad. Sci. USA* **118**, e2023202118 (2021).
- Gregory, A. C. et al. The Gut Virome Database reveals age-dependent patterns of virome diversity in the human gut. *Cell Host Microbe* **28**, 724–740 (2020).
- Guerin, E. & Hill, C. Shining light on human gut bacteriophages. *Front. Cell. Infect. Microbiol.* **10**, 481 (2020).
- Guerin, E. et al. Isolation and characterisation of Φ crAss002, a crAss-like phage from the human gut that infects *Bacteroides xylanisolvens*. *Microbiome* **9**, 89 (2021).
- Shkoporov, A. N. et al. Long-term persistence of crAss-like phage crAss001 is associated with phase variation in *Bacteroides intestinalis*. *BMC Biol.* **19**, 163 (2021).
- Schmidtke, D. T. et al. The prototypic crAssphage is a linear phage-plasmid. *Cell Host Microbe* **33**, 1347–1362 (2025).
- Brissow, H., Canchaya, C. & Hardt, W.-D. Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion. *Microbiol. Mol. Biol. Rev.* **68**, 560–602 (2004).
- Bobay, L.-M., Touchon, M. & Rocha, E. P. C. Pervasive domestication of defective prophages by bacteria. *Proc. Natl Acad. Sci. USA* **111**, 12127–12132 (2014).
- Borodovich, T. et al. Large scale capsid-mediated mobilisation of bacterial genomic DNA in the gut microbiome. Preprint at *bioRxiv* <https://doi.org/10.1101/2024.11.15.623857> (2024).
- Reyes, A., Semenkovich, N. P., Whiteson, K., Rohwer, F. & Gordon, J. I. Going viral: next-generation sequencing applied to phage populations in the human gut. *Nat. Rev. Microbiol.* **10**, 607–617 (2012).
- Minot, S. et al. Rapid evolution of the human gut virome. *Proc. Natl Acad. Sci. USA* **110**, 12450–12455 (2013).
- Shkoporov, A. N. et al. The human gut virome is highly diverse, stable, and individual specific. *Cell Host Microbe* **26**, 527–541 (2019).
- Kieft, K., Zhou, Z. & Anantharaman, K. VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. *Microbiome* **8**, 90 (2020).
- Guo, J. et al. VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. *Microbiome* **9**, 37 (2021).
- Camargo, A. P. et al. Identification of mobile genetic elements with geNomad. *Nat. Biotechnol.* **42**, 1303–1312 (2024).
- Tisza, M. J., Petrosino, J. F. & Cregeen, S. J. J. Cenote-Taker 3 for fast and accurate virus discovery and annotation of the virome. Preprint at *bioRxiv* <https://doi.org/10.1101/2025.08.20.671380> (2025).
- Hatfull, G. F. & Hendrix, R. W. Bacteriophages and their genomes. *Curr. Opin. Virol.* **1**, 298–303 (2011).
- Lai, S., Wang, H., Bork, P., Chen, W.-H. & Zhao, X.-M. Long-read sequencing reveals extensive gut phageome structural variations driven by genetic exchange with bacterial hosts. *Sci. Adv.* **10**, eadn3316 (2024).
- Chen, L. et al. Short- and long-read metagenomics expand individualized structural variations in gut microbiomes. *Nat. Commun.* **13**, 3175 (2022).
- Páez-Espino, D. et al. Uncovering Earth's virome. *Nature* **536**, 425–430 (2016).
- Marbouty, M., Thierry, A., Millot, G. A. & Koszul, R. MetaHiC phage-bacteria infection network reveals active cycling phages of the healthy human gut. *eLife* **10**, e06008 (2021).
- Li, D., Liu, C.-M., Luo, R., Sadakane, K. & Lam, T.-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674–1676 (2015).
- Kolmogorov, M. et al. metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nat. Methods* **17**, 1103–1110 (2020).
- Watson, M. & Warr, A. Errors in long-read assemblies can critically affect protein prediction. *Nat. Biotechnol.* **37**, 124–126 (2019).
- Sereika, M. et al. Oxford Nanopore R10.4 long-read sequencing enables the generation of near-finished bacterial genomes from pure cultures and metagenomes without short-read or reference polishing. *Nat. Methods* **19**, 823–826 (2022).
- Nayfach, S. et al. CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nat. Biotechnol.* **39**, 578–585 (2021).
- Roux, S. et al. iPHoP: an integrated machine learning framework to maximize host prediction for metagenome-derived viruses of archaea and bacteria. *PLoS Biol.* **21**, e3002083 (2023).
- Smolka, M. et al. Detection of mosaic and population-level structural variants with Sniffles2. *Nat. Biotechnol.* **42**, 1571–1580 (2024).
- Murooka, Y. & Harada, T. Expansion of the host range of coliphage P1 and gene transfer from enteric bacteria to other gram-negative bacteria. *Appl. Environ. Microbiol.* **38**, 754–757 (1979).
- Wong, F. H. & Bryan, L. E. Characteristics of PR5, a lipid-containing plasmid-dependent phage. *Can. J. Microbiol.* **24**, 875–882 (1978).

39. Olsen, R. H., Siak, J. S. & Gray, R. H. Characteristics of PRD1, a plasmid-dependent broad host range DNA bacteriophage. *J. Virol.* **14**, 689–699 (1974).
40. Trigodet, F., Sachdeva, R., Banfield, J. F. & Eren, A. M. Assemblies of long-read metagenomes suffer from diverse errors. Preprint at *bioRxiv* <https://doi.org/10.1101/2025.04.22.649783> (2025).
41. Shaw, J. & Li, H. myloasm. *GitHub* <https://myloasm-docs.github.io/> (2025).
42. Harshey, R. M. Transposable phage Mu. *Microbiol. Spectr.* <https://doi.org/10.1128/microbiolspec.MDNA3-0007-2014> (2014).
43. Siguier, P., Goubeyre, E., Varani, A., Ton-Hoang, B. & Chandler, M. Everyman's guide to bacterial insertion sequences. *Microbiol. Spectr.* <https://doi.org/10.1128/microbiolspec.mdna3-0030-2014> (2015).
44. Wagner, A. Cooperation is fleeting in the world of transposable elements. *PLoS Comput. Biol.* **2**, e162 (2006).
45. Siguier, P., Perochon, J., Lestrade, L., Mahillon, J. & Chandler, M. ISfinder: the reference centre for bacterial insertion sequences. *Nucleic Acids Res.* **34**, D32–D36 (2006).
46. Wibowo, M. C. et al. Reconstruction of ancient microbial genomes from the human gut. *Nature* **594**, 234–239 (2021).
47. Liang, G. et al. The stepwise assembly of the neonatal virome is modulated by breastfeeding. *Nature* **581**, 470–474 (2020).
48. Pinto, Y., Chakraborty, M., Jain, N. & Bhatt, A. S. Phage-inclusive profiling of human gut microbiomes with Phanta. *Nat. Biotechnol.* **42**, 651–662 (2024).
49. Ellis, E. & Delbrück, M. The growth of bacteriophage. *J. Gen. Physiol.* **22**, 365–384 (1939).
50. Sutcliffe, S. G., Reyes, A. & Maurice, C. F. Bacteriophages playing nice: lysogenic bacteriophage replication stable in the human gut microbiota. *iScience* **26**, 106007 (2023).
51. Porter, N. T. et al. Phase-variable capsular polysaccharides and lipoproteins modify bacteriophage susceptibility in *Bacteroides thetaiotaomicron*. *Nat. Microbiol.* **5**, 1170–1181 (2020).
52. Nanda, A. M., Thormann, K. & Frunzke, J. Impact of spontaneous prophage induction on the fitness of bacterial populations and host-microbe interactions. *J. Bacteriol.* **197**, 410–419 (2015).
53. Berngruber, T. W., Lion, S. & Gandon, S. Spatial structure, transmission modes and the evolution of viral exploitation strategies. *PLoS Pathog.* **11**, e1004810 (2015).
54. Piel, D. et al. Phage-host coevolution in natural populations. *Nat. Microbiol.* **7**, 1075–1086 (2022).
55. Bignaud, A. et al. Phages with a broad host range are common across ecosystems. *Nat. Microbiol.* **10**, 2537–2549 (2025).
56. Benler, S. et al. A diversity-generating retroelement encoded by a globally ubiquitous *Bacteroides* phage. *Microbiome* **6**, 191 (2018).
57. van de Putte, P., Cramer, S. & Giphart-Gassler, M. Invertible DNA determines host specificity of bacteriophage Mu. *Nature* **286**, 218–222 (1980).
58. Sørensen, M. C. H. et al. *Campylobacter* phages use hypermutable polyG tracts to create phenotypic diversity and evade bacterial resistance. *Cell Rep.* **35**, 109214 (2021).
59. Gordillo Altamirano, F. L. & Barr, J. J. Phage therapy in the postantibiotic era. *Clin. Microbiol. Rev.* **32**, e00066-18 (2019).
60. Ross, A., Ward, S. & Hyman, P. More is better: selecting for broad host range bacteriophages. *Front. Microbiol.* **7**, 1352 (2016).
61. Kiss, J., Szabó, M. & Olasz, F. Site-specific recombination by the DDE family member mobile element IS30 transposase. *Proc. Natl Acad. Sci. USA* **100**, 15000–15005 (2003).

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



Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025

Methods

Study population and ethics statement

For time point 1, stool samples were collected from six adult volunteers living in the Bay Area, CA, USA. This study involving humans was approved by our institutional internal review board (Stanford IRB 42043; principal investigator: A.S.B.), and informed consent was obtained from all participants. For time point 2, the same individuals were recontacted for a second stool donation, 2 years after the initial one.

The samples for time point 1 were included in the publication of Maghini et al.⁶², and short-read metagenomic sequencing reads are available at the National Center for Biotechnology Information's Sequence Read Archive under the identifier PRJNA940499.

Sample collection and processing

Stool samples were collected without a preservative and stored at -80°C . All DNA extractions were performed using the QIAamp PowerFecal Pro DNA Kit (QIAGEN; 51804) according to the manufacturer's instructions, with the exception of using the EZ-Vac Vacuum Manifold (Zymo Research) instead of centrifugation. DNA concentration was measured using a Qubit 3.0 Fluorometer (Thermo Fisher Scientific) with the dsDNA High Sensitivity kit.

Metagenomic short-read sequencing

The samples from T1 had already been sequenced; therefore, we generated new libraries only for samples from T2. Metagenomic sequencing libraries were pooled, and 2×150 bp reads were generated using the NovaSeq 6000 platform (Illumina; 20012850) to a final depth of 6 Gb per sample.

Metagenomic long-read sequencing

All samples from both time points underwent long-read metagenomic sequencing using the ONT platform. DNA fragment distribution was assessed using a TapeStation (Agilent; G2992AA). Samples with apparent fragmentation were cleaned up using a bead-based protocol before library preparation⁶³.

Libraries were prepared using the Native Barcoding Kit V24 (ONT; SQK-NBD114.24) using 1,000 ng of DNA as input. In the pooling step, four samples were combined, resulting in three total libraries. The libraries were loaded onto PromethION R10.4.1 flow cells (ONT; FLO-PRO114M) and sequenced until exhaustion of the flow cells.

Short-read data processing

For short-read sequencing, all raw reads were processed with our in-house NextFlow pipeline (v.22.10.5; ref. 64; https://github.com/bhattlab/bhattlab_workflows_nf). In short, reads were deduplicated using HTStream SuperDeduper (v.1.3.3), and low-quality bases were trimmed using TrimGalore (v.0.6.7). Reads were then mapped against the human genome (hg38) using bwa (v.0.7.17.31), and all matching reads were discarded. For comparability, T1 samples were downsampled to final library sizes randomly drawn from the distribution of T2 library sizes after preprocessing.

For each sample, metagenomic assembly was performed using MetaHIT (v.1.2.9), and genes were predicted using Bakta (v.1.8.2). Assemblies were binned into draft genomes using MetaBAT (v.2.5), CONCOCT (v.1.1.0) and MaxBin (v.2.2.7), followed by bin consolidation using DAS Tool (v.1.1.6). Bin quality was assessed using CheckM (v.1.2.2), and taxonomic classification was performed using GTDB-Tk (v.2.3.0) using the Genome Taxonomy Database (GTDB) r214.

Long-read data processing

For long-read sequencing, POD5 files were basecalled and demultiplexed using Dorado (v.0.5.3) using the 'super-high accuracy' model (v.3.4.0) to create the final set of FASTQ files. Read quality and

length distribution were assessed using NanoPlot (v.1.41.6) before and after the removal of human reads (read mapping against the human genome (v.38) using minimap2 (v.2.26-r1175)). Metagenomic assembly was performed using metaFlye (v.2.9.2-b1786) using the nano-hq flag to use only reads of quality Q20 or higher for initial assembly. Binning and taxonomic classification of bins were performed as described for short reads. The workflows for long-read data processing are also available at https://github.com/bhattlab/bhattlab_workflows_nf.

To better compare the short and long reads, the long reads were subsampled to the same mean depth as the short reads using a custom script, which randomly selected read IDs from the processed reads up to a specified amount of total sequencing. Seqtk subseq (v.1.4-r130) was used to subsample the original FASTQ files for the selected read IDs. Read quality and length distribution were assessed with NanoPlot (v.1.41.6). Reads were assembled and binned using the same workflow described above.

Phage prediction

To predict phages, we applied geNomad (v.1.7.6; ref. 23), VIBRANT (v.1.2.1; ref. 21), VirSorter2 (v.2.2.4; ref. 22) and Cenote-Taker3 (v.3.4.0; ref. 24) to the short-read and long-read assemblies for each metagenomic sample. For each predicted phage, we used CheckV (v.1.0.1; ref. 34) to assess the quality and to adjust the boundary predictions on the basis of CheckV host trimming. VIBRANT, geNomad, Cenote-Taker3 and CheckV are included in the NextFlow project available at https://github.com/bhattlab/bhattlab_workflows_nf. VirSorter2 was run separately because it relies on Snakemake (v.5.26.0) for execution. All predicted phages, including full phage contigs and predicted prophages, were collated from the four tools.

Comparison across short-read and long-read sequencing

To compare the phage predictions across short-read and long-read sequencing, we mapped short-read contigs against the (subsampled) long-read assembly using blast+ (v.2.2.31), filtering alignments for identity (99%) and query coverage (90%). To determine if an integrated long-read phage was fully covered by short-read contigs, we required a single short-read contig to align to at least 95% of the phage region. To determine the short-read coverage of integrated phages, we mapped the short reads to the long-read assembly using Bowtie 2 (v.2.5.4) and calculated the per-base coverage using SAMtools (v.1.21).

To compare short-read and long-read phage predictions in more detail, we adapted the segment overlap metric (originally developed to measure overlaps between predicted biosynthetic gene clusters⁶⁵) to measure which fraction of long-read phages was covered by the predicted short-read phages and vice versa (Extended Data Fig. 4). This metric calculates recall by considering each long-read phage prediction as a positive instance. A long-read phage was considered a true positive if covered (up to a variable cutoff of $x\%$) by alignments of one or more short-read phage contigs; otherwise, it was classified as a false negative. Recall, defined as true positives over the sum of all positives, quantifies the fraction of long-read phages that are found by short-read sequencing. Precision was defined on the basis of the short-read phages; those overlapping (to at least $x\%$) a long-read phage were considered true positives, whereas those not overlapping were considered false positives. Precision was then calculated as true positives over the sum of true and false positives, quantifying the fraction of short-read phages found by long-read sequencing.

Clustering of phages across time points

Prophages were clustered across time points within each individual by mapping all phages of T1 and T2 against each other using blast+ (v.2.2.31), and genome identity and coverage were computed using the CheckV companion scripts³⁴. To cluster the host regions, we extracted 20-kb regions on either side of integrated prophages, concatenated these phage-surrounding regions and performed the same clustering

analysis with the CheckV companion scripts. Additionally, we mapped the original phage region and their host regions against the assembly of the other time point to prevent erroneous classification of dynamic phages not annotated in the other time point. Finally, we quantified the median read coverage in both time points for all phage regions and filtered out all phages with a median coverage of less than ten in both time points.

Using these clustering results, we iteratively consolidated the clustering in the following way (Extended Data Fig. 5): phages with high identity (98%) and genome coverage (90%) were clustered together. If the hosts clustered together as well, they were classified as stably integrated; otherwise, they were classified as phages found in several host contexts. If two phages failed to cluster together but their host regions did, we relaxed the cutoff for genome coverage to classify them as stably integrated because we observed large SVs between phages integrated into the same host context. For all singleton clusters, we checked whether the host region or the phage itself had a high-identity and high-coverage mapping to the assembly of the other time point. Because we observed many false-positive dynamic phages at contig edges, we discounted singleton phages found at the edges of contigs. If a phage and its host region mapped well to the same contig in the other time point, we classified this phage as stably integrated; otherwise, we classified them as either dynamic phages or singletons lost/gained with their host if we found the host region in the assembly of the other time point.

Strain replacement analysis

To evaluate the relationship between paired MAGs within individuals, we identified all high-quality and high-coverage bins that shared the same taxonomic classification on the basis of GTDB-Tk (v.2.3.0) across time points. High-coverage bins were quantified by calculating the median contig coverage on the basis of metaFlye (v.2.9.2-b1786) output within each bin. The shared ANI of paired bins was calculated using FastANI (v.1.34; ref. 66) and default one-to-one parameters. Mean proportion of strain replacement across individuals was then calculated by setting a minimum shared ANI for the same strain and determining the number of paired MAGs that fall above this value per individual. This was done using an ANI cutoff of all unique ANI values to calculate the strain replacement for different ANI cutoffs.

Identification of SVs overlapping predicted phage regions

To find SVs in our data, we mapped the reads of each time point against the assembly of the other time point within an individual using NGMLR (v.0.2.7; ref. 67). Binary Alignment Map files were sorted with SAMtools (v.1.9) and structural variants were identified using Sniffles2 (v.2.2; ref. 36).

Assignment of host taxonomy for integrated prophages

We used iPHoP (v.1.3.3; ref. 35) to predict hosts for all phages using default parameters and database version iPHoP_db_Aug23_rw. For integrated long-read phages, we annotated their hosts by classifying each annotated gene on a given contig using the mmseqs taxonomy module from MMseqs2 (v.14.7e284; ref. 68). This module provides a taxonomic annotation on the basis of the GTDB (v.214.1; ref. 69) database for each gene. For each contig, we then combined the annotations for all genes not located in predicted phage regions at each taxonomic level. Annotations were accepted if more than 50% of genes agreed, disregarding genes without taxonomic annotation.

The iPHoP, MMseqs2-based and binning taxonomic assignments were evaluated for consensus at each taxonomic level (Extended Data Fig. 4). To determine a subset of integrated phages for which we had high-confidence host assignment, we filtered our results for agreement between high-quality bins (more than 90% completeness and less than 5% contamination; ref. 70) and MMseqs2-based assignment down to the family level.

Clustering of phages on species level for host-range analysis

To evaluate host range within phage species, we clustered all geNomad-annotated phages in all samples at a minimum of 95% ANI and 80% alignment fraction using CheckV (v.1.0.1; ref. 34) supporting scripts. We specified a minimum of 80% query and target coverage (`--min_tcov 80 --min_qcov 80`) because we expect phages to be assembled more contiguously by long reads. The resulting cluster membership information was merged with the host annotations from our MMseqs2 taxonomy approach and bin taxonomy, as described above.

To visualize phages found in several bacterial families, we compared bins using AliTV⁷¹, filtering alignments by length and nucleotide identity. Additionally, we calculated their read coverage (removing reads less than 500 bp in length and with more than 1% mismatches to prevent spurious mappings) and the number of read alignment ends per genomic position. A high number of read alignment ends could point towards potential mis-assemblies, as explored in a previous study⁴⁰.

To gain orthogonal evidence for the integration of the same phage species into distinct bacterial families, we also assembled our data using myloasm (v.0.1.0; ref. 41), an alternative assembler to Flye. We used standard parameters, except for `--min-reads-contig 3`. For all putative phage species in distinct bacterial families, we identified the corresponding myloasm contig and tried to verify the prophage integration (Supplementary Table 2). In the case of two clusters, the phages were found at the edges of contigs, but their integration into the bacterial species could be verified by the broader host context derived from the myloasm assembly (Extended Data Fig. 8).

Gene content for integrated prophages

To assess gene content across all integrated prophages with structural variant evidence, we used pharokka (v.1.7.3) against the pharokka (v.1.4.0) database⁷² with the `--meta`, `--skip-mash` and `--split` flags.

Synteny and gene content visualization

Visualization of all gene content and synteny was done using LoVis4u (v.0.1.4.1; ref. 73). Two of the 22 IScream phages were smaller than 10 kb and therefore were removed from this analysis. For IScream phage synteny visualization, gff files of the identified full-length IScream phages, generated by pharokka annotation (described above), were used as input for LoVis4u visualization with default configuration. For host context visualization, gff files from Bakta were parsed to visualize specified windows and to convert them to a format compatible with LoVis4u. The reformatted gff3 files were used as input for LoVis4u visualization using an updated configuration file specifying `mmseqs_min_seq_id = 0.8`.

Identification of potential integrase enzymes

To identify potential integrase enzymes in our assemblies, we used a set of Pfam hidden Markov models described in an earlier exploration of mobile genetic elements⁷⁴: PFPF07508 for large serine recombinases; PF00239 for small serine recombinases; PF00589 for tyrosine recombinases; and PF00665, PF13333 and PF13683 for DDE recombinases. We used the `hmmsearch` command (with the `--cut_ga` flag) from HMMER (v.3.4; ref. 75) against all predicted proteins. We then annotated each phage by counting which type of potential integrase was present within the phage boundaries using only phages with evidence from SVs.

Identification of IScream phages in MGV and Clostridia genomes

To explore the prevalence and abundance of IScream phages, we searched for potential IScream phages in public datasets. We ran ISEScan (v.1.7.2.3; ref. 76) on all viral genome assemblies from MGV and identified genomes that contained one IS30 element starting less than 1 kb from the beginning of the assembly and one IS30 element ending less than 1 kb from the end of the assembly.

We next searched for existing bacterial isolates containing integrated IScream phages. We downloaded all 2,160 class Clostridia genome assemblies annotated as complete or chromosomal level using the NCBI Datasets tool. We ran ISEScan and geNomad on all assemblies and identified genomes that contained a geNomad-annotated prophage region with one IS30 element starting less than 1 kb from the beginning of the prophage region and one IS30 element ending less than 1 kb from the end of the prophage region, which included *B. hansenii* ATCC 27552 (see Supplementary Table 3 for a complete list of potential IScream phages).

Additionally, we analysed the species-level taxonomic profiles for two public datasets, generated with Phanta (v.1.0; ref. 48), which included a viral database built on representative genomes from MGv (see the original Phanta publication for details about data processing). Each phage species from this database was classified as potential IScream phage if a genome contained in this species bin was identified to be an IScream phage.

The dataset from Liang et al.⁴⁷ included bulk metagenomic sequencing and VLP-enriched sequencing of infants. In this dataset, we quantified the relative taxonomic abundance of IScream-containing phage species in paired bulk and VLP sequencing to identify potential particle formation of IScream phages. Finally, the non-cancer control samples from the dataset of Yachida et al.⁷⁷ was used to calculate prevalence and mean abundance of phage species in healthy individuals.

Clustering of IS30 elements

To cluster IS30 elements across IScream phages, we used the ete3 toolkit (v.3.1.3; ref. 78). First, we built a tree for all IS30 proteins on all IScream phages using the 'standard_fasttree' workflow from ete3, consisting of a multiple sequence alignment with Clustal Omega (v.1.2.4) and tree construction using FastTree (v.2.1.8) (Extended Data Fig. 10). Because this analysis showed a clear separation between outward-directed and inward-directed IS30 proteins, we subsequently focused only on the outward-directed IS30. To gain insights into the evolutionary history of IScream phages, we extracted all outward-directed IS30 proteins from our IScream phages, all potential MGv IScream phages and all bona fide bacterial IS30 elements from the ISfinder database. For the MGv phages, we filtered the outward-directed IS30 proteins to be longer than 600 and shorter than 2,200 nucleotides. All proteins were clustered at 70% amino acid similarity over 80% of the alignment, using MMseqs (v.14.7e284; ref. 79). A tree was then constructed on the cluster representatives using the 'standard_fasttree' workflow in ete3, modified to trim positions in the multiple sequence alignment with more than 90% gaps using trimAl (v.1.4.rev6).

Screening for potential phage IS30 elements in ancient stool metagenomes

To identify potential IScream phages in ancient stool samples, we downloaded the raw data from Wibowo et al.⁴⁶, containing sequencing of desiccated paleofecal samples (1,000–2,000 years old) from the southwestern USA and Mexico. Raw reads were processed and assembled as described above, phages were identified with geNomad, and IS elements were detected with ISEScan. For all contigs that contained IS30 elements within predicted phages, we assessed their DNA damage level with DamageProfiler (v.1.1; ref. 80). To prevent inclusion of modern IScream phage IS30 proteins, we filtered the phage-overlapping IS30 proteins for being on contigs recruiting more than 1,000 reads and showing an estimated 5'C>T damage level or an estimated 3'G>A damage level over 1%, resulting in six potential IScream IS30 open reading frames. Note that these filtering steps are rather specific because many more IS30 proteins are identified on shorter contigs, which are not predicted to be of phage origin. The six ancient IS30 proteins were added to the clustered IS30 proteins from MGv, IScream phages and ISfinder, and the tree was recomputed as described above.

Culturing of the *B. hansenii* IScream phage

B. hansenii (ATCC 27552) was grown anaerobically (90% nitrogen, 5% carbon dioxide and 5% hydrogen) in an anaerobic chamber (Sheldon Manufacturing) in Brain Heart Infusion (Sigma) supplemented (BHIS) with hemin (5 µg ml⁻¹), L-cysteine (1 mg ml⁻¹) and sodium bicarbonate (0.2%).

B. hansenii was grown overnight in pre-reduced BHIS. The culture was pelleted by means of centrifugation (Eppendorf; 5920R) at 4,000g for 15 min, and DNA was extracted using a DNeasy Blood and Tissue Kit (QIAGEN; 69504) following the manufacturer's instructions for Gram-positive bacteria. Bacterial genome sequencing was performed by Plasmidsaurus using ONT with custom analysis and annotation.

PEG precipitation of *B. hansenii* phage particles

B. hansenii was grown overnight in 4 l of BHIS. The supernatant from the culture was harvested by means of centrifugation (Eppendorf; 5920R) at 4,198g for 15 min at 4 °C. Sodium chloride was added to the supernatant to a final concentration of 5 M, and PEG 8000 was added to a final concentration of 10%. This solution was stirred for 30 min at 4 °C to dissolve and then left overnight at 4 °C with no agitation. PEG-precipitated phage particles were harvested by centrifugation at 10,000g for 10 min at 4 °C. The supernatant was removed, and the resulting pellets were air-dried for 3–5 min in an inverted position. The pellets were combined and resuspended in a total of 10 ml of SM buffer (100 mM NaCl, 8 mM MgSO₄·7H₂O and 50 mM Tris-HCl (pH 7.5)). An equal volume of chloroform was added, mixing by inverting and centrifuged at 12,000g for 10 min at 4 °C. The aqueous phase was collected, and chloroform treatment was repeated.

PCR validation of phage particles from *B. hansenii* IScream phage

PCR primers were designed using NCBI Primer-BLAST under default settings. The PCR product size was targeted to be between 150 bp and 250 bp. PCR primers were designed to target (1) *B. hansenii* *gmk* gene; (2) an internal region of the IScream phage; and (3) the upstream junction site of the integrated phage and bacterial chromosome, as well as to span the junction of circularization of the IScream phage region (Fig. 4e; the primers are listed in Supplementary Table 4). PCRs were performed on PEG-precipitated samples directly with and without DNase treatment. For the DNase-treated samples, 50 µl of PEG Prep was treated with 5 µl of 10X TURBO DNase buffer, 1.13 µl of TURBO DNase (Invitrogen; AM2239) and 1 µl of RNase (1 mg ml⁻¹; Invitrogen; AM2270), incubated at 37 °C for 1 h and then heat inactivated at 70 °C for 10 min.

PCRs were performed using Q5 high-fidelity DNA polymerase (annealing temperature of 67 °C, annealing time of 30 s and extension time of 5 s at 72 °C). PCRs were run on a 2% agarose gel. A raw gel image is shown in Supplementary Fig. 1.

Transmission electron microscopy of *B. hansenii* lysate

A 500-ml culture of *B. hansenii* was grown overnight, as described above. The supernatant from the culture was harvested after 24 h by means of centrifugation (Eppendorf; 5920R) at 4,198g for 15 min at 4 °C. The supernatant was then filtered using a 0.22-µm filter (Fisher Scientific; 09-719C). The supernatant (5 µl) was placed on glow-discharged 200-mesh carbon/Formvar-coated Cu grids (FF300-Cu) and allowed to settle for 3 min. Grids were washed by touching the sample side with two drops of water. Three drops of 1% uranyl acetate in double-distilled water were then added, and the third drop was incubated on the grid for 1 min at room temperature. The remainder of the last drop was removed with filter paper, and the grid was allowed to dry. The samples were then observed on the JEOL JEM-1400 transmission electron microscope at 120 kV, and photographs were taken using a Gatan Orius digital camera.

Statistical analysis and visualization

All statistical tests were performed in R (v.4.2.2). Data visualization was performed using ggplot2 (v.3.5.1), which is part of the tidyverse (v.2.0.0) suite of tools⁸¹.

Reporting summary

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

Data availability

The raw sequencing data for all samples sequenced in this study are available from the European Nucleotide Archive under the study identifier PRJEB88320. The short-read sequencing data from T1 had been included in a paper by Maghini et al.⁶² and are available under the identifier PRJNA940499. Data for the MGv catalogue from the publication by Nayfach et al.⁸ are available at <https://portal.nersc.gov/MGv/>. The raw sequencing data of ancient metagenomic samples from the publication by Wibowo et al.⁴⁶ are available under PRJNA561510. The raw data for the Phanta-profiled datasets from Liang et al.⁴⁷ are available under PRJNA524703 and PRJDB4176 from the study by Yachida et al.⁷⁴. Source data are provided with this paper.

Code availability

The source code developed for the reported analysis and data visualization is publicly available at Zenodo (<https://doi.org/10.5281/zenodo.15192469>)⁸² and at GitHub (https://github.com/bhattlab/long_read_benchmark).

62. Maghini, D. G. et al. Quantifying bias introduced by sample collection in relative and absolute microbiome measurements. *Nat. Biotechnol.* **42**, 328–338 (2024).
63. Maghini, D. G., Moss, E. L., Vance, S. E. & Bhatt, A. S. Improved high-molecular-weight DNA extraction, nanopore sequencing and metagenomic assembly from the human gut microbiome. *Nat. Protoc.* **16**, 458–471 (2021).
64. Di Tommaso, P. et al. Nextflow enables reproducible computational workflows. *Nat. Biotechnol.* **35**, 316–319 (2017).
65. Carroll, L. M. et al. Accurate *de novo* identification of biosynthetic gene clusters with GECCO. Preprint at *bioRxiv* <https://doi.org/10.1101/2021.05.03.442509> (2021).
66. Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T. & Aluru, S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat. Commun.* **9**, 5114 (2018).
67. Sedlazeck, F. J. et al. Accurate detection of complex structural variations using single-molecule sequencing. *Nat. Methods* **15**, 461–468 (2018).
68. Mirdita, M., Steinegger, M., Breitwieser, F., Söding, J. & Levy Karin, E. Fast and sensitive taxonomic assignment to metagenomic contigs. *Bioinformatics* **37**, 3029–3031 (2021).
69. Parks, D. H. et al. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* **36**, 996–1004 (2018).
70. Bowers, R. M. et al. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat. Biotechnol.* **35**, 725–731 (2017).

71. Ankenbrand, M. J., Hohlfield, S., Hackl, T. & Förster, F. AliTV—interactive visualization of whole genome comparisons. *PeerJ Comput. Sci.* **3**, e116 (2017).
72. Bouras, G. et al. Pharokka: a fast scalable bacteriophage annotation tool. *Bioinformatics* **39**, btac776 (2023).
73. Egorov, A. A. & Atkinson, G. C. LoVis4u: a locus visualization tool for comparative genomics and coverage profiles. *NAR Genom. Bioinform.* **7**, lqaf009 (2025).
74. Khedkar, S. et al. Landscape of mobile genetic elements and their antibiotic resistance cargo in prokaryotic genomes. *Nucleic Acids Res.* **50**, 3155–3168 (2022).
75. Eddy, S. R. Accelerated profile HMM searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).
76. Xie, Z. & Tang, H. ISEScan: automated identification of insertion sequence elements in prokaryotic genomes. *Bioinformatics* **33**, 3340–3347 (2017).
77. Yachida, S. et al. Metagenomic and metabolomic analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal cancer. *Nat. Med.* **25**, 968–976 (2019).
78. Huerta-Cepas, J., Serra, F. & Bork, P. ETE 3: reconstruction, analysis, and visualization of phylogenomic data. *Mol. Biol. Evol.* **33**, 1635–1638 (2016).
79. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017).
80. Neukamm, J., Peltzer, A. & Nieselt, K. DamageProfiler: fast damage pattern calculation for ancient DNA. *Bioinformatics* **37**, 3652–3653 (2021).
81. Wickham, H. et al. Welcome to the tidyverse. *J. Open Source Softw.* **4**, 1686 (2019).
82. Hickey, A., Wirbel, J. & Bhatt, A. Data for the long-read benchmark project. Zenodo <https://doi.org/10.5281/zenodo.15192469> (2025).

Acknowledgements We thank all members of the Bhatt laboratory for enriching discussions and invaluable feedback, especially D. Maghini and Y. Pinto. We also thank members of the Ashley, Altemose and Good laboratories at Stanford University for help with long-read sequencing and for inspiring discussions. We thank X. Zeng from the Fischbach laboratory for providing the *B. hansenii* ATCC 27552 sample and J. J. Perrino for his help and guidance during transmission electron microscopy imaging. The project described was supported in part by ARRA (award no. 1S1ORR026780-01) from the National Center for Research Resources. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the National Center for Research Resources or the National Institutes of Health. A.S.B. is supported by the Paul Allen Distinguished Investigator Award, and the Bhatt laboratory is supported by a Stand Up To Cancer Grant and NIH R01AI148623, R01AI143757 and U54AG089334 (Human Virome Program centre grant). This research was supported in part by grant NSF PHY-2309135 to the Kavli Institute for Theoretical Physics (KITP). J.W. is a Damon Runyon Quantitative Biology Fellow supported by the Damon Runyon Cancer Research Foundation (DRQ-22-24). A.S.H. and D.T.S. acknowledge support from the National Science Foundation Graduate Research Fellowship Program (DGE-1656518). D.T.S. and M.D. are supported by the Cell and Molecular Biology Training Grant (T32 GM007276). R.B.C. is supported by the A.P. Giannini Foundation. D.C. is supported by NIH T32HG000044. N.J.E. is supported by the Stanford University DARE Fellowship.

Author contributions A.S.B., J.W. and A.S.H. conceived and designed the study. M.D. collected and processed all stool samples. M.D., A.S.H., R.B.C. and J.W. conducted all long-read library preparation and sequencing. Short-read and long-read comparison, phage dynamics and host-range analyses and data visualizations were performed by J.W. and A.S.H. J.W., D.C. and N.J.E. conducted IScream phage analyses and visualization. A.S.H., D.T.S., R.B.C. and D.C. performed the *Blautia* experiments and interpretation. Funding acquisition was done by A.S.B. The original draft was written by J.W., A.S.H. and A.S.B. All authors contributed to the review and editing of this paper.

Competing interests A.S.B. is a founder of Stylus Medicine, serves on the scientific advisory board and is a board observer. She also serves on the Scientific Advisory Board of Caribou Biosciences and Cantata Biosciences. The other authors declare no competing interests.

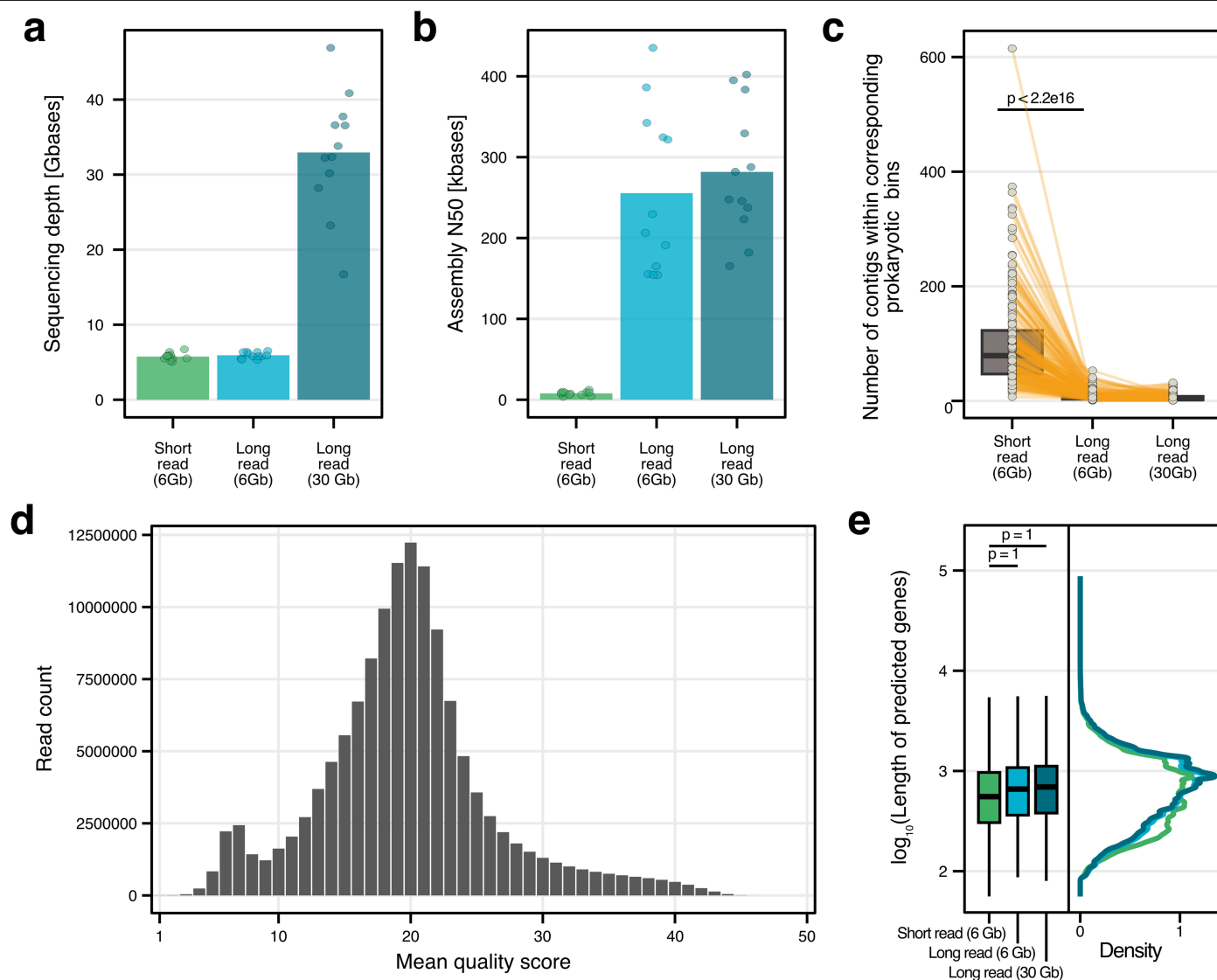
Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41586-025-09786-2>.

Correspondence and requests for materials should be addressed to Ami S. Bhatt.

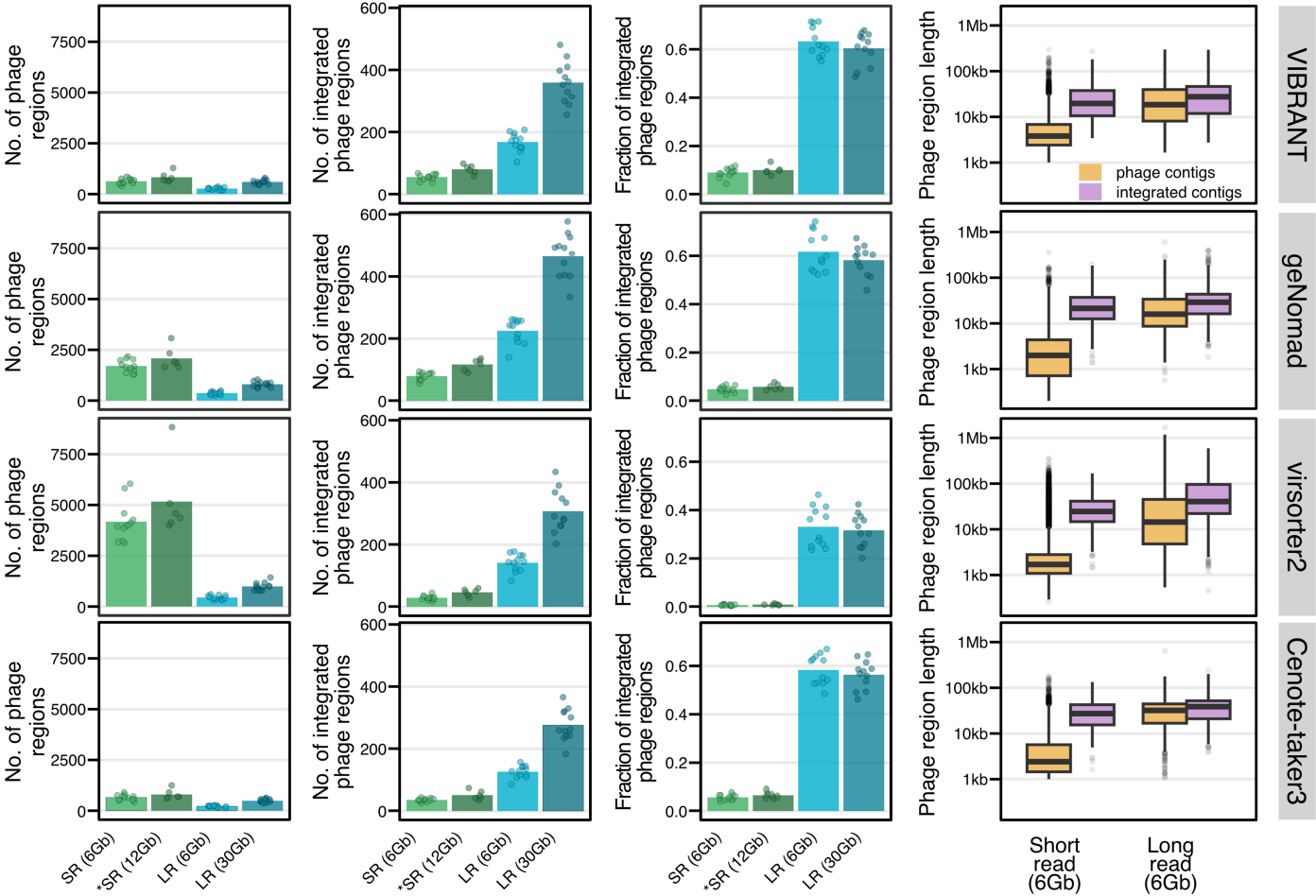
Peer review information *Nature* thanks the anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permissions information is available at <http://www.nature.com/reprints>.



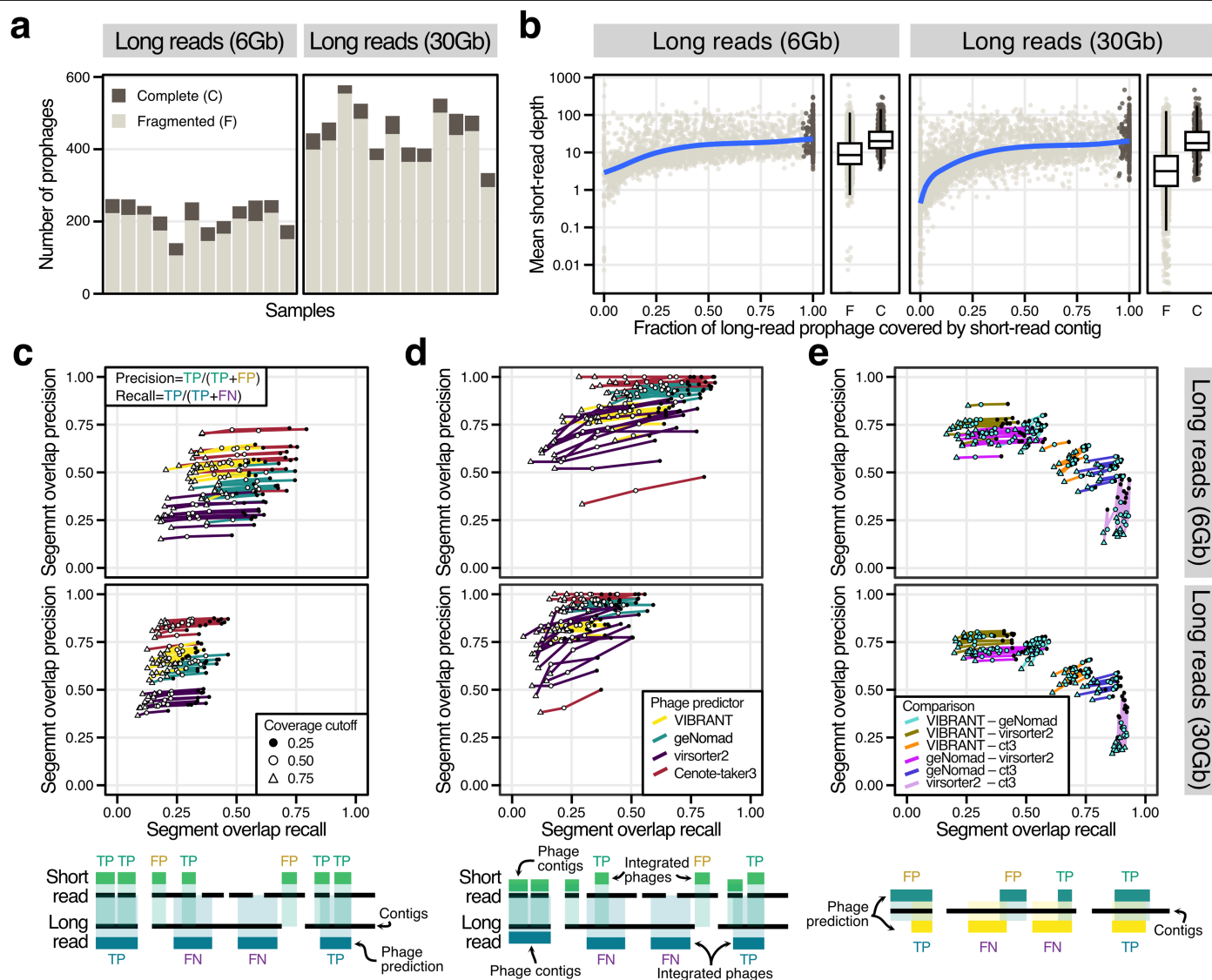
Extended Data Fig. 1 | Comparisons between short-read and long-read sequencing. **a)** Mean sequencing depth of Illumina short-read sequencing, subsampled ONT long-read sequencing, and deep ONT long-read sequencing across all samples ($n = 12$) are shown as bars with points indicating individual samples. **b)** Mean assembly n50 across different sequencing types and depth as bars while points indicate individual samples ($n = 12$). **c)** Comparison of the number of contigs present in assembled bins that were shared between the three sequencing types: Illumina short reads (SR) at a depth -6 Gb, subsampled ONT reads (LR) to -6 Gb, and deep ONT reads (DLR) at -30 Gb, represented as a box plot. Box plots show the interquartile ranges (IQRs) as boxes, with the median as a black horizontal line, whiskers extending up to the most extreme points within 1.5-fold IQR, and all data points are indicated as dots. Corresponding bins were determined by matching bin taxonomic assignment across sequencing approaches. Orange lines track corresponding bins across the sequencing types. A paired Wilcoxon signed rank test was used to compare the number of contigs per bin from short reads at -6 Gb depth and long reads at

-6 Gb depth ($n = 183$, p -value $< 2.2e-16$). **d)** Histogram of mean read quality score from deep long read sequencing samples. **e)** Comparison of the length of predicted genes from assemblies of each sequencing type. A one-sided Wilcoxon rank sum test was used to compare the median length of predicted genes between Illumina short reads and ONT long reads subsampled to the same depth. Results showed no significant difference (SR $n = 3457575$, LR $n = 3408004$, $p = 1$). Similarly, one-sided Wilcoxon rank sum test was used to compare the short read assembly with the deep ONT assembly, similarly showing no significant difference (SR $n = 3457575$, DLR $n = 6671897$, $p = 1$). The Wilcoxon test indicated that the median gene lengths from short-read assembly were not significantly greater than median gene lengths from ONT assemblies. Box plots show the IQRs as boxes, with the median as a black horizontal line and the whiskers extending up to the most extreme points within 1.5-fold IQR. Line plots show overall sample distribution on a density scale of zero to one.



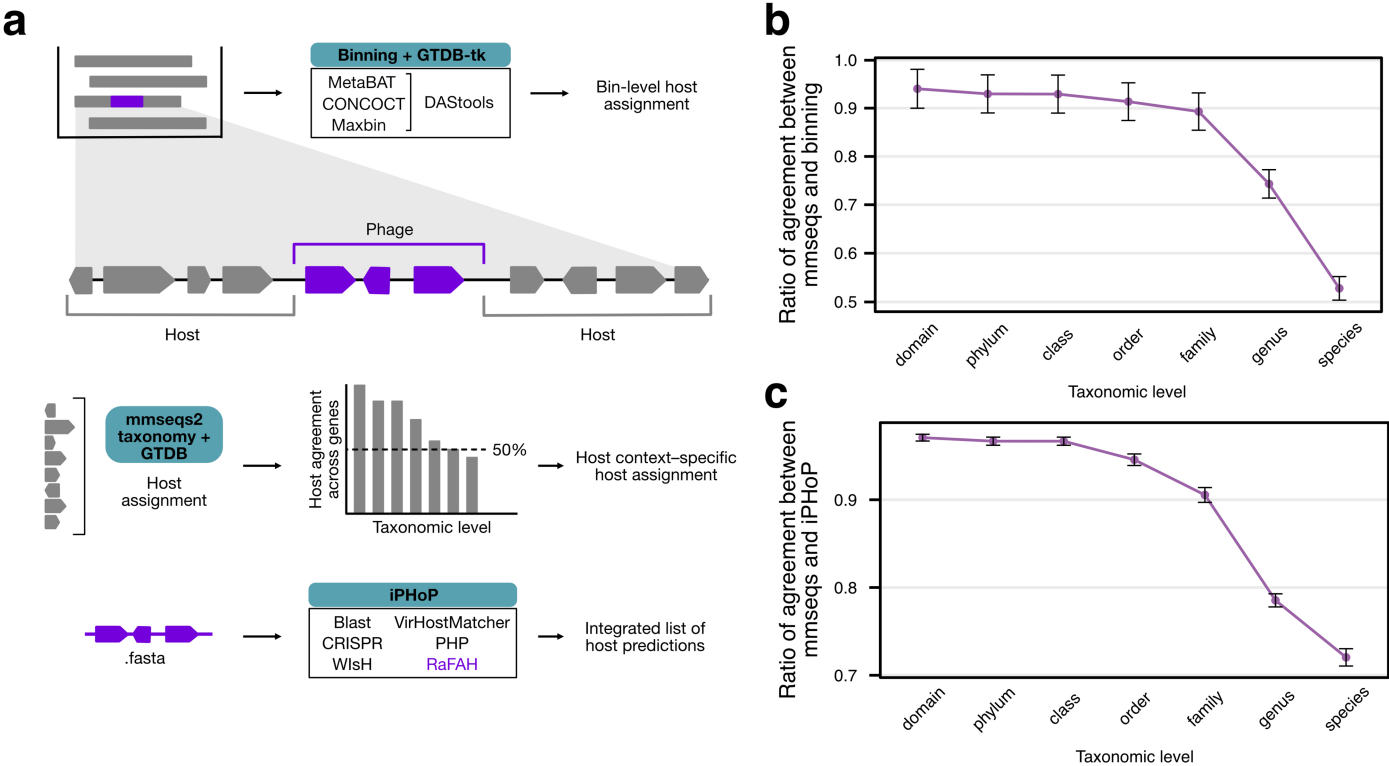
Extended Data Fig. 2 | Comparisons between different phage annotation tools. For each phage annotation tool (VIBRANT, geNomad, virstorter2, and Cenote-taker3), the average total number of phages annotated, average number of integrated phages annotated, and the average fraction of identified phages that were integrated are presented for each sequencing type and depth (Illumina short reads -6 Gb ($n = 12$), Illumina short reads -12 Gb (T1 only, $n = 6$)*, ONT long reads -6 Gb ($n = 12$), and ONT long reads -30 Gb ($n = 12$)). Phage length

was compared between entire phage contigs (orange) and integrated phages (purple) for Illumina short reads at -6 Gb and subsampled ONT long reads at -6 Gb, indicating that short-read sequencing resulted in more fragmented (and therefore shorter) phage genomes. Box plots show the interquartile ranges (IQRs) as boxes, with the median as a black horizontal line, whiskers extending up to the most extreme points within 1.5-fold IQR, and outliers indicated as dots.



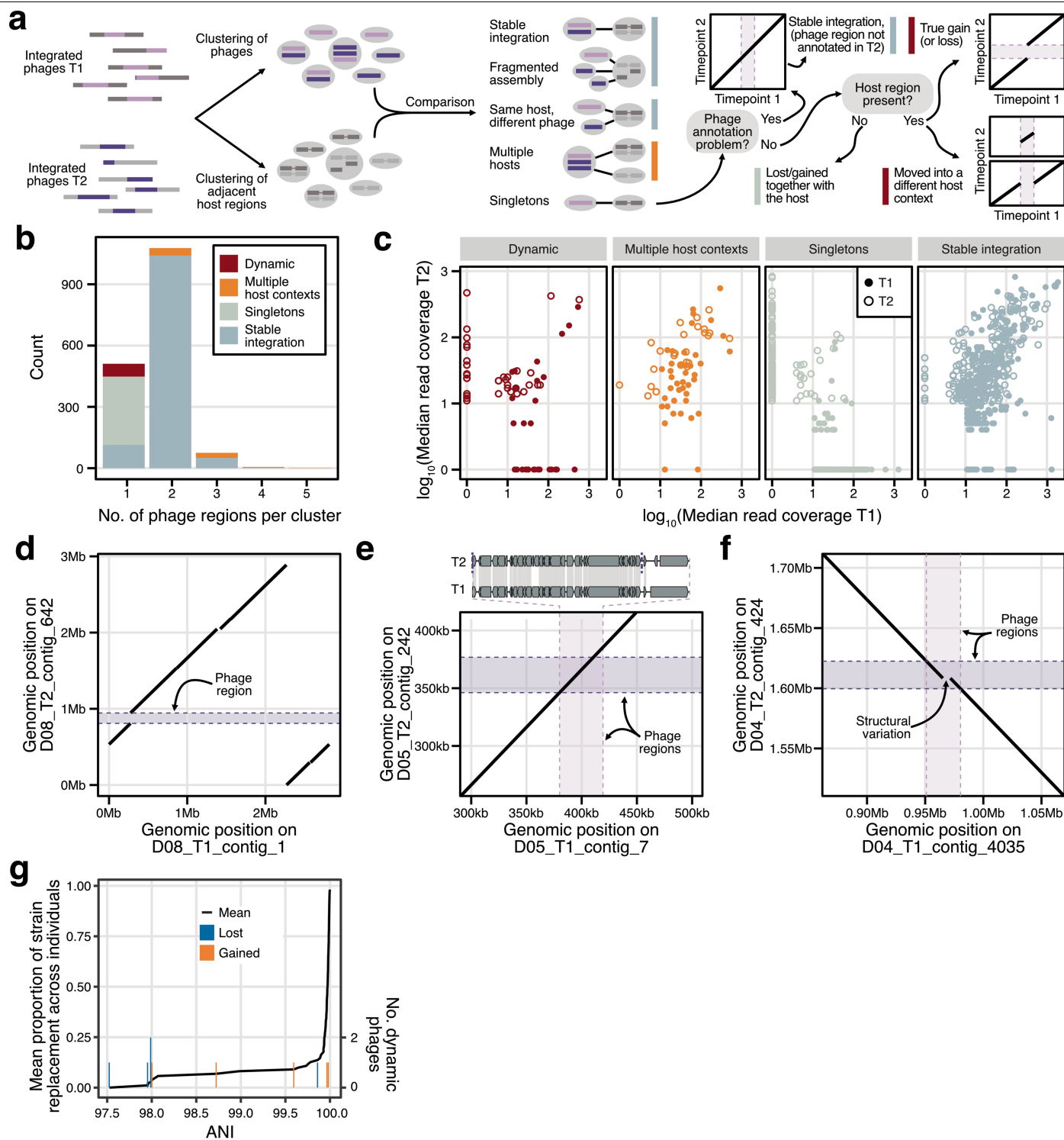
Extended Data Fig. 3 | Short-read phage fragmentation and segment overlap analysis. **a)** Barplot showing the number of integrated phages per sample for downsampled (~6 Gb) and deep (~30 Gb) long-read sequencing. The fill color indicates if the same phage was found to be fragmented in the short-read assembly or found to be completely covered by a single short-read contig. **b)** Short-read depth of all integrated prophages identified with geNomad in the long-read assemblies. Short reads were mapped to the long-read assemblies and the average depth for all regions identified to be phages was calculated. Dots indicate if the phage was found to be fully covered ($n = 467$ for long reads (6 Gb), $n = 499$ for long reads (30Gb)) or fragmented ($n = 2240$ for long reads (6 Gb), $n = 5088$ for long reads (30 Gb)) when mapping short-read contigs to the long-read assemblies. The right-side panel for each plot shows the average short-read depth as box plots. Box plots show the interquartile ranges (IQRs) as boxes, with the median as a black horizontal line and the whiskers extending up to the most extreme points within 1.5-fold IQR. **c)** Segment overlap metric (see Methods and ref Carroll et al.⁶⁵) calculated between short-read phages and long-read phages. Across all predictors, only about 20% of long-read phages were found to be sufficiently covered by short-read phages (segment overlap recall). This fraction increased with a more lenient cutoff for how much of the long-read phage had to be covered to be considered a true positive, indicating that only a small part of the long-read

phage is covered by short-read phage contigs. Similarly, the segment overlap precision (how many short-read phages overlapped long-read phages) was higher for all phage predictors, but decreases with the overall number of phages predicted to be present. Especially for virstorter2, only about 35% of the ~4000 predicted short-read phages per sample overlapped corresponding long-read phage predictions. This indicates that virstorter2 predicts many short-read contigs to be phage, but does not predict the same sequence as phage when assembled in context. See schematic at the bottom of the figure panel for a visual representation of mapping of short-read phages to long-read phages. **d)** Same as panel **c**, but only for integrated phages (disregarding phage contigs). In this analysis, segment overlap precision is ~80%, indicating that most of phages identified to be integrated prophages in short-read assemblies are similarly identified in the long-read assemblies, whereas the coverage of integrated phages in the long-read assemblies remains relatively low (~20%). See schematic at the bottom of the figure panel for a visual representation of mapping of integrated short-read phages to long-read phages. **e)** Segment overlap metric when comparing across phage predictors, in the long-read assemblies only. See schematic at the bottom of the figure panel for a visual representation for how segment overlap and recall were calculated when comparing predictors.



Extended Data Fig. 4 | Host taxonomy assignment schematic and benchmarking. a) Schematic representation of the different approaches to assigning hosts to integrated phages. From the long-read assemblies, we assign bacterial hosts for a given prophages by comparing the bin taxonomic assignment from GTDB-tk (bin membership of the contig containing the integrated prophage, see Methods) and gene-level annotations from mmseqs-taxonomy (see Methods). For this approach, all genes in host regions of the same contig as the integrated prophage are annotated against the GTDB database and a consensus taxonomic assignment is generated by majority rule. The current gold standard for taxonomic prediction of phage hosts is the prediction tool iPhoP, which integrates the predictions from several different approaches.

Host prediction via iPhoP has traditionally been necessary, since phages are often assembled as fragments or without host context with short-read sequencing. For each phage genome, iPhoP generates an integrated list of host predictions. **b)** Mean ratio of the host agreement between the mmseq2-taxonomy based approach and binning taxonomic assignment for integrated prophages across all samples at each taxonomic level. Data are presented as mean values $\pm$ SEM. **c)** Mean ratio of the host agreement between the mmseq2-taxonomy based approach and iPhoP taxonomic assignment for integrated prophages across all samples at each taxonomic level. Since iPhoP outputs a list of possible hosts, here we looked for agreement at each taxonomic level from any potential host. Data are presented as mean values $\pm$ SEM.

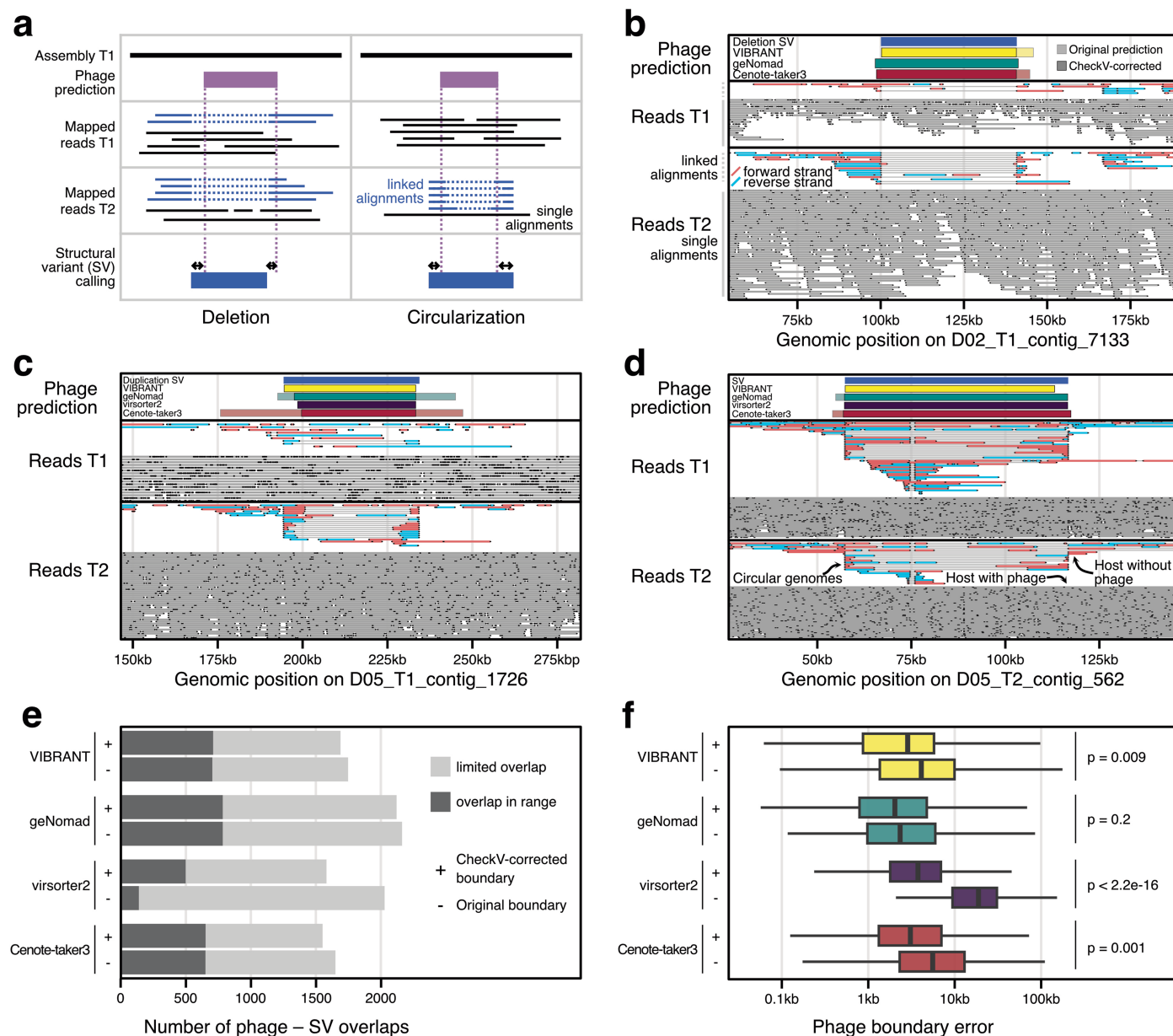


Extended Data Fig. 5 | See next page for caption.

Extended Data Fig. 5 | Clustering of prophages within individuals.

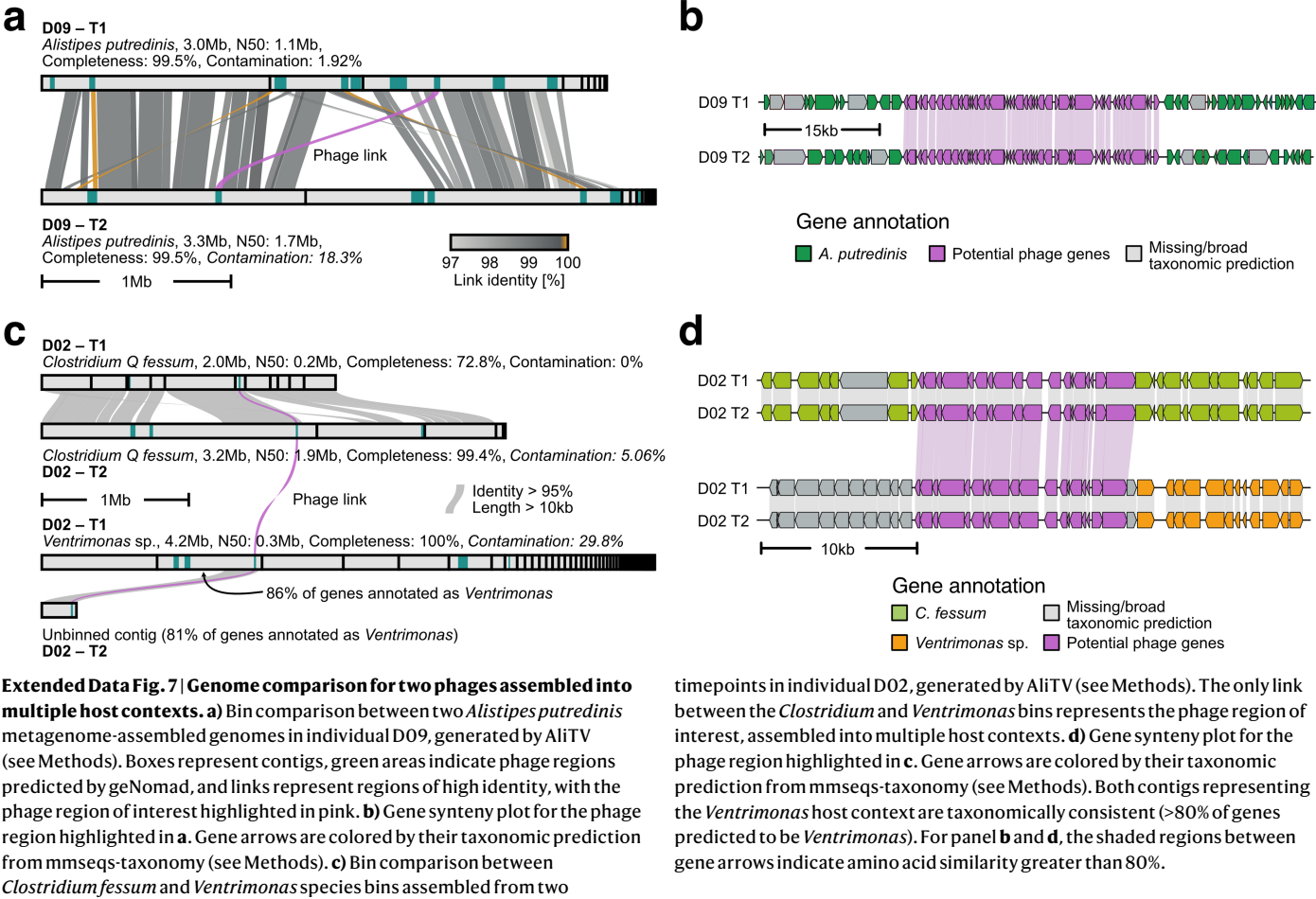
a) Schematic showing the clustering and classification of integrated phages within individuals. In short, all integrated phages and their adjacent host regions were clustered separately using the CheckV companion scripts for genome identity and coverage calculation based on blast mappings. Clustering was done with high identity (99%) and genome coverage (90%) cutoffs. Then, clusters were refined by comparing between phage and host clusters. **b)** Histogram showing the number of clusters with the specified number of phage regions. Cluster annotation is indicated according to the figure legend. **c)** Scatter plot showing the median read coverage in timepoint 1 against coverage in timepoint 2 for each phage cluster, split by their classification. For each cluster, the representative phage genome is shown. Dot shapes indicate the timepoint of assembly. **d)** Dotplot comparing two contigs within individual D08, showing the gain of a 137 kb integrated phage into an otherwise stable host. The phage region (representing a dynamic phage) is indicated by a shaded pink area. **e)** Dotplot comparing two contigs within individual D05, showing a cluster classified as 'same host, different phage' due to variations in phage

annotation. The phage regions are indicated by shaded pink areas. While both contigs align perfectly, the region annotated as phage in both timepoints do not overlap over more than 90%, resulting in disparate clustering outcomes. The reason for this inconsistency is a single nucleotide difference changing the predicted genes in T2 (annotated above). **f)** Dotplot comparing two contigs within individual D04, showing a cluster classified as 'same host, different phage' due to a large structural variation (deletion in T2). The phage regions are indicated by shaded pink areas. The gap in the alignment represents a 6.5 kb deletion in T2 compared to T1, resulting in disparate clustering outcomes because of respective genome coverage below 90%. **g)** Line graph showing the mean proportion of strain replacement across individuals depending on the ANI cutoff used to call strain replacement. The analysis is based on high-quality and high coverage (>30x median coverage) MAGs that had the same species annotation across timepoints ($n = 90$). Colored bars indicate the number of dynamic phages that were lost (blue) or gained (orange) from paired MAGs with a shared ANI.

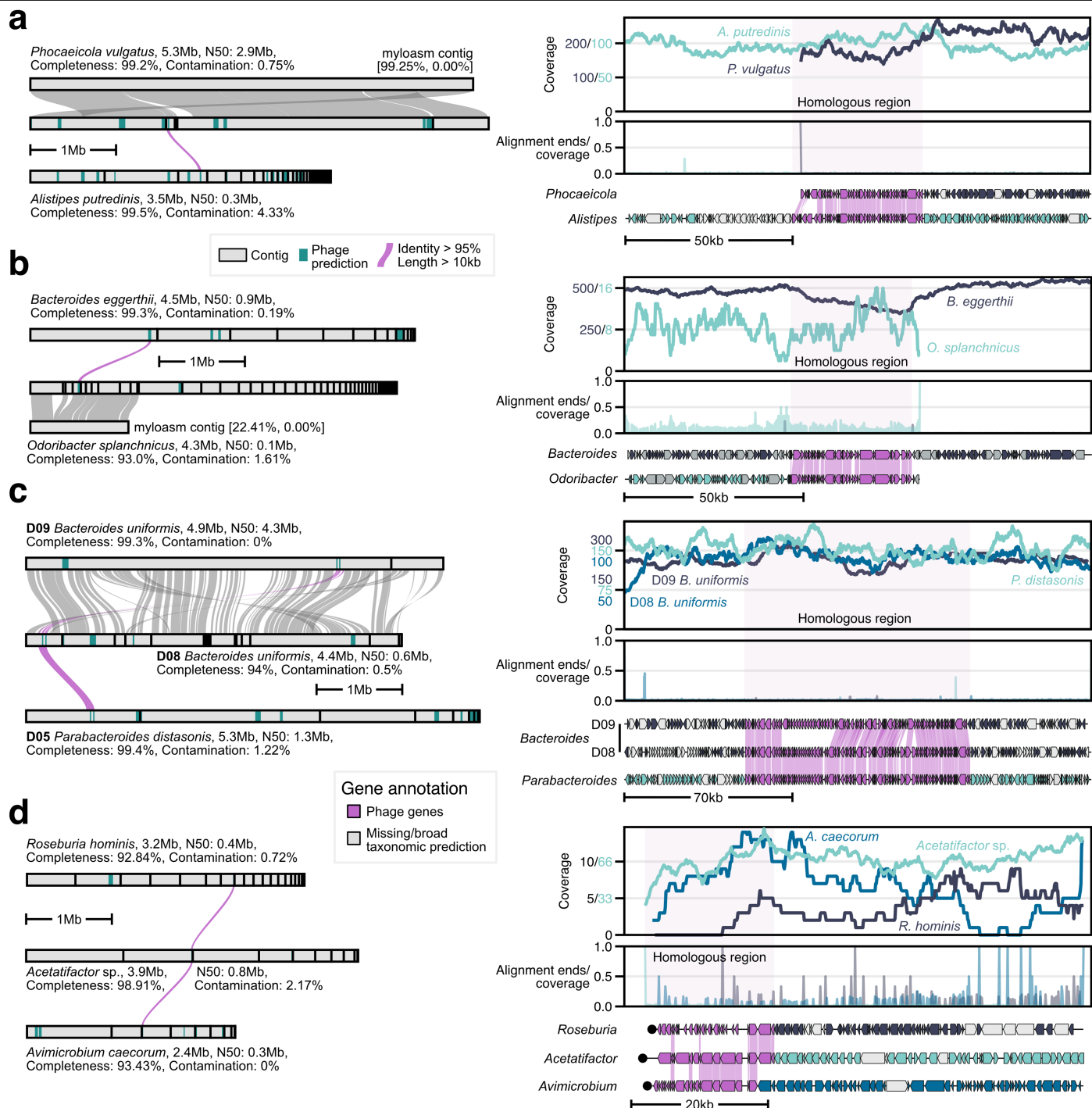


Extended Data Fig. 6 | Detection of exact phage boundaries through long-read mapping and structural variation calling. **a)** Schematic representation of the identification of exact phage boundaries by structural variant (SV) calling. In short, SVs were identified with Sniffles2 on the basis of read mapping within and across timepoints for the same individual. In these cases, linked read alignments (either supporting a deletion or a duplication, represented by blue lines in the schematic) can be used to identify SVs. Duplication SVs can be interpreted as the presence of circular phage genomes. We considered all phage-SV overlaps that covered at least 50% of both phage and SV to be high-quality phages identified with base-pair accuracy. **b)** Read alignment plot illustrating a phage identified by a deletion SV in individual D02. Read alignments are separated into linked and single alignments. Each line represents the alignment of a single read with dots showing the start and end of each alignment. Linked alignments are colored by strand and connected by a thin grey line to indicate that they originate from the same read. On the top, the coordinates from phage predictions are shown: boundaries of the original phage prediction are shown with 50% shading, while boundary correction with CheckV is shown in full opacity. **c)** Same plot as **b**, but for a duplication SV (circular phage genome), identified in individual D05. **d)** Same plot as **b**, showing for a phage region present in individual D05 that there is read evidence for the simultaneous presence of circular phage genomes (linked reads supporting a duplication SV), hosts without the integrated phage (linked reads

supporting a deletion SV), and hosts with the integrated phage (single alignments) in T2. **e)** Number of phages predicted by different tools with and without CheckV boundary correction that overlapped SVs called by Sniffles2, colored by their overlap being in range (more than 50% of the phage and the SV region) or not within range. Virsorter2 tends to predict very long phages, resulting in many comparisons where less than 50% of the phage was covered by the SV region. CheckV boundary correction increases the number of Virsorter2 predictions that fall in range. **f)** Evaluation of the total boundary error calculated as the sum of the absolute difference on either side of the prediction with and without CheckV boundary correction. This is done for all phage-SV overlaps within range (see **e**). Box plots show the interquartile ranges (IQRs) as boxes, with the median as a black horizontal line, whiskers extending up to the most extreme points within 1.5-fold IQR. A one-sided Wilcoxon rank sum test was used to determine if the phage boundary error with CheckV correction was shorter than the original predicted boundary of each tool. No significant difference was found for the geNomad predictions (Original $n = 2161$, CheckV-corrected $n = 2119$, $p = 0.2$). CheckV-boundary correction significantly decreased the boundary error for VIBRANT (Original $n = 1746$, CheckV-corrected $n = 1688$, $p = 0.009$) and Cenote-taker3 (Original $n = 1649$, CheckV-corrected $n = 1550$, $p = 0.001$), but had the most significant effect on the virsorter2 predictions (Original $n = 2027$, CheckV-corrected $n = 1580$, $p < 2.2e-16$).

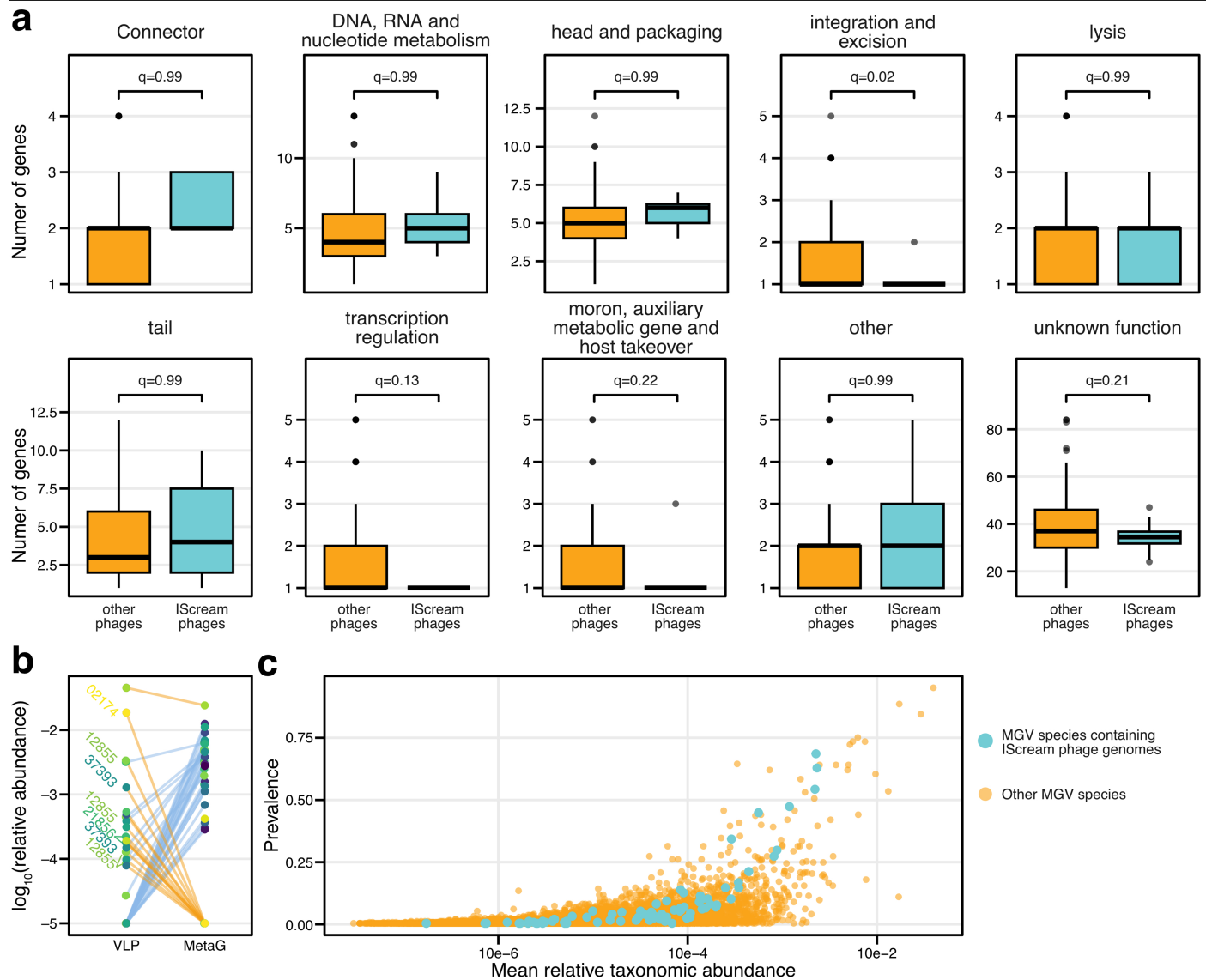


timepoints in individual D02, generated by AliTV (see Methods). The only link between the *Clostridium* and *Ventrimonas* bins represents the phage region of interest, assembled into multiple host contexts. **d)** Gene synteny plot for the phage region highlighted in **c**. Gene arrows are colored by their taxonomic prediction from mmseqs-taxonomy (see Methods). Both contigs representing the *Ventrimonas* host context are taxonomically consistent (>80% of genes predicted to be *Ventrimonas*). For panel **b** and **d**, the shaded regions between gene arrows indicate amino acid similarity greater than 80%.



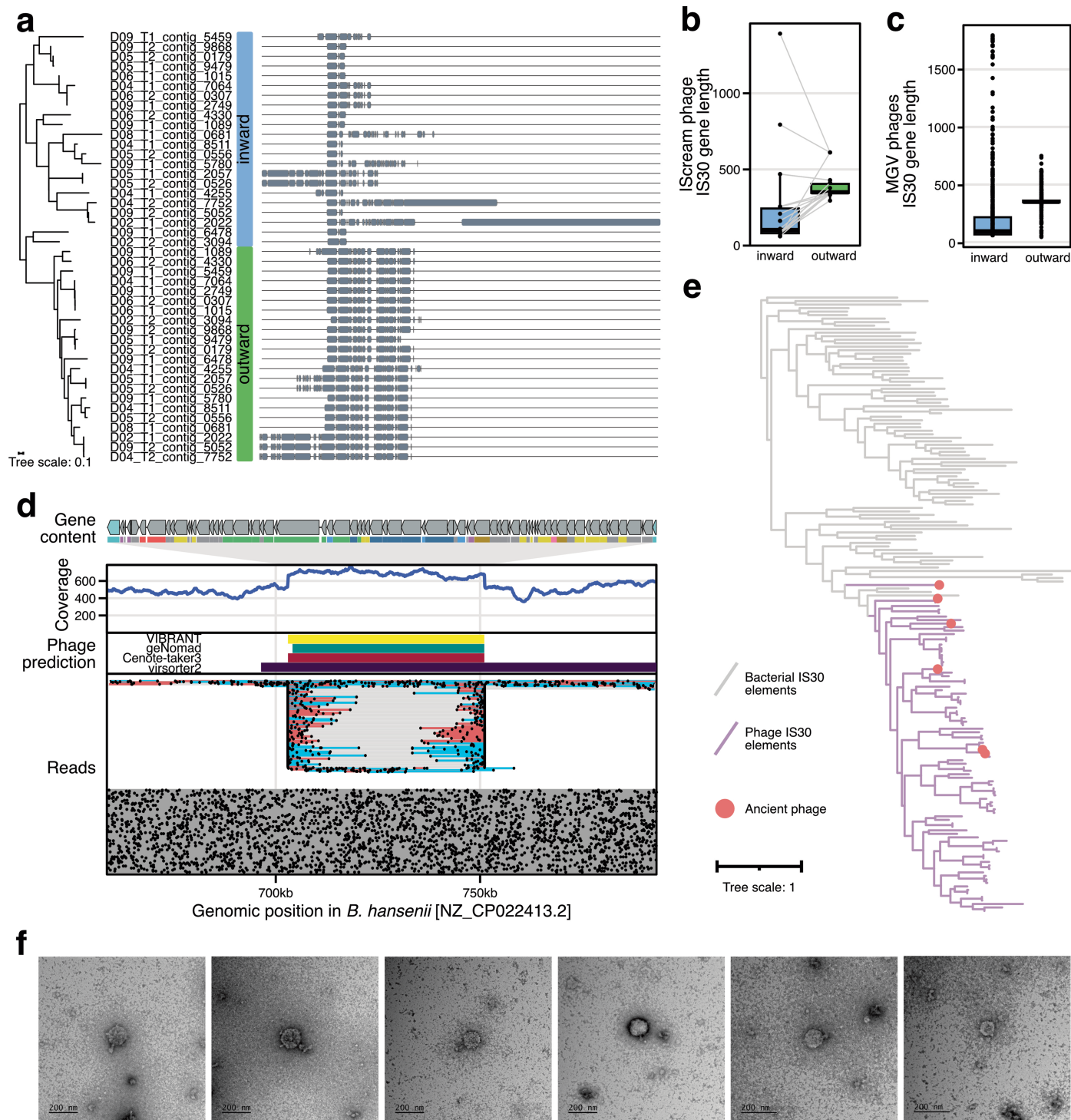
Extended Data Fig. 8 | Evidence for phages assembled in taxonomically distinct host contexts. a) Chart representing high-quality metagenome-assembled genomes for two different species, annotated with phage predictions from geNomad. A single genomic region (>10 kb) is shared between both genomes with >95% nucleotide identity, annotated by a pink link. As the phage region was assembled at the edge of a contig, we sought orthogonal validation by identifying and comparing the corresponding contig generated from an alternative assembly method (myloasm⁴¹). The myloasm contig is annotated with its individual CheckM completeness and contamination result. The right panel shows the coverage and synteny plot for the annotated phage region. The top panel shows read coverage, the middle panel shows the number of alignment ends divided by coverage, and the bottom panel shows the gene

arrows across the three genomes. Genes are colored according to their taxonomic predictions. **b)** Same plot organization as in **a**, for a different example phage region. **c)** Same plot organization as in **a**. In this example, two phage regions were identified as part of a larger shared region between two genomes found in different bacterial families. This region is more complex, as the comparison of two closely related *Bacteroides uniformis* genomes reveals complex rearrangement between these closely related genomes, involving the region in question. **d)** Same plot organization as in **a**, for a phage found in three bacterial hosts of different classes. All phage regions are at the ends of contigs and could not be verified by myloasm. The shaded regions in the synteny plots indicate amino acid similarity greater than 80%.



Extended Data Fig. 9 | IScream phages have phage-like gene content and are present in MGVS. a) Number of genes classified as different functional groups, identified from pharokka, shown for both IScream phages and all other integrated phages. Both groups of phages were identified through structural variation analysis. Each panel is annotated by a Benjamini-Hochberg-corrected P-value, resulting from testing differences in gene numbers with a two-sided Wilcoxon test (Other phages $n = 569$, IScream phage $n = 22$). Box plots show the interquartile ranges (IQRs) as boxes, with the median as a black horizontal line, whiskers extending up to the most extreme points within 1.5-fold IQR, and outliers are indicated as dots. **b)** Log10-transformed relative taxonomic

abundance for all phage species containing IScream phages identified in MGVS genomes for data from Liang et al.⁴⁷ (see Methods). Each dot represents a phage species within one individual, sequenced either in bulk metagenomic (MetaG) or virus-like particle (VLP)-enriched samples, and are connected by lines to indicate the same individual. Different phage species are indicated by different colors. **c)** Mean relative taxonomic abundance is plotted against prevalence for all phages identified by phanta in the metagenomes of healthy individuals from Yachida et al.⁷⁷ (see Methods). Phage species containing IScream phages identified in MGVS genomes are highlighted in cyan.



Extended Data Fig. 10 | See next page for caption.

Extended Data Fig. 10 | IS30 clustering and *Blautia hansenii* IScream phage.

a) Multiple sequence alignment of all IS30 open reading frames in IScream phages with structural variation evidence, assembled in this study, visualized through the ete3 toolkit⁷⁸ (boxes indicate alignments and empty areas indicate gaps in the alignment). The outward- and inward-directed IS30 proteins form two separate clusters, indicated by the tree reconstructed from the multiple sequence alignment. **b)** Box plot showing the IS30 gene length for outward- and inward-directed IS30 genes from IScream phages assembled here ($n = 22$). Box plots show the interquartile ranges (IQRs) as boxes, with the median as a black horizontal line, whiskers extending up to the most extreme points within 1.5-fold IQR, and all data points are indicated as dots. **c)** Boxplot showing the IS30 gene length for outward- and inward-directed IS30 genes from all potential IScream phages identified in MGVI ($n = 1705$). All boxplots show the interquartile ranges (IQRs) as boxes, with the median as a black horizontal line, whiskers extending up to the most extreme points within 1.5-fold IQR, and outliers indicated as dots. **d)** Evidence for the presence of

circular phage genomes from read alignments against the *B. hansenii* reference genome (NZ_CP022413.2). Read alignments are separated into linked alignments (top part) and single alignments (bottom part). Each line represents the alignment of a single read with dots showing the start and end of each alignment. Linked alignments are colored by strand and connected by a thin grey line to indicate that they originate from the same read. The remainder of this figure panel shows coordinates of phage predictions, read coverage, and the gene content of the phage genome (as determined by structural variation calling). **e)** Tree constructed from a multiple sequence alignment of IS30 proteins clustered at 70% amino acid identity (outward-directed IS30 proteins from MGVI IScream phages, IScream phages assembled here, and bona fide bacterial IS30 elements) together with IS30 proteins found overlapping phages predicted in assemblies of paleofeces. **f)** Raw, uncropped TEM images of virus-like particles from the lysate of an overnight culture of *B. hansenii*. Scale bar for 200 nm included on each image.

Reporting Summary

Nature Portfolio 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 Portfolio policies, see our [Editorial Policies](#) 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 (<i>n</i>) 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 <i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i>
☒	☐ 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. <i>F</i> , <i>t</i> , <i>r</i>) with confidence intervals, effect sizes, degrees of freedom and <i>P</i> value noted <i>Give P values as exact values whenever suitable.</i>
☒	☐ 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 <i>d</i> , Pearson's <i>r</i>), 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	No software used for data collection.
Data analysis	Source code for analysis and figure generation is publicly available at Zenodo under https://doi.org/10.5281/zenodo.15192469 and on GitHub under https://github.com/bhattlab/long_read_benchmark . Analysis tools: NextFlow v22.10.5, HTStream SuperDeduper v1.3.3, TrimGalore v0.6.7, bwa v0.7.17, metahit v1.2.9, bakta v1.8.2, MetaBAT v2.5, CheckM v1.2.2, GTDB-tk v2.3.0 using the GTDB r214 database, dorado v0.5.3 using the 'super-high accuracy' model v3.4.0, NanoPlot v1.41.6, minimap2 v2.26-r1175, meta-flye v2.9.2-b1786, Seqtk subseq v1.4-r130, geNomad v1.7.6, VIBRANT v1.2.1, virstorter2 v2.2.4, Cenote-taker3 v3.4.0, CheckV v1.0.1, snakemake (v5.26.0), blast v2.2.31+, bowtie2 v2.5.4, samtools v1.21, FastANI v1.34, NGMLR v0.2.7, samtools v1.9, Sniffles2 v2.2, iPHoP v1.3.3, mmseqs2 v14.7e284, myloasm v0.1.0, PharoKka v1.7.3, LoVis4u v0.1.4.1, HMMER v3.4, phanta v1.0, ete3 toolkit v3.1.3, Clustal Omega v1.2.4, FastTree v2.1.8, trimAl v1.4.rev6, DamageProfiler v1.1, R v4.2.2, ggplot2 v3.5.1, tidyverse v2.0.0

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 and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [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 description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

The raw sequencing data for all samples sequenced in this study are available from the European Nucleotide Archive under the study identifier PRJEB88320. The short read sequencing data from T1 had been included in the earlier publication by Maghini, Dvorak et al. 2024 and are available under the identifier PRJNA940499. Data for the Metagenome Gut Virus catalogue (MGV) from the publication by Almeida et al. 2020 is available under <https://portal.nersc.gov/MGV/>. The raw sequencing data of ancient metagenomic samples from the publication by Wibowo et al. 2021 are available under PRJNA561510. The raw data for the phanta-profiled datasets from Liang et al. 2020 is available PRJNA524703 and under PRJDB4176 for the study from Yachida et al. 2019.

Research involving human participants, their data, or biological material

Policy information about studies with [human participants or human data](#). See also policy information about [sex, gender \(identity/presentation\), and sexual orientation](#) and [race, ethnicity and racism](#).

Reporting on sex and gender	Sex and gender data were not collected, as sex and gender are not expected to have an influence on prophage dynamics.
Reporting on race, ethnicity, or other socially relevant groupings	Race, ethnicity, or other socially relevant grouping data were not collected as they are not expected to have an influence on prophage dynamics.
Population characteristics	All participants are adults residing in California, USA. No additional metadata was collected for each donor due to IRB approval restrictions. No covariates are expected to have an influence on sample preservative efficacy.
Recruitment	The study was advertised verbally and participants volunteered for the study on a first-come, first-served basis until the predetermined sample size was met. Local recruitment may have resulted in a self-selection bias in which participants have shared environmental, lifestyle, and socio-economic impacts. This may impact how generalizable the findings are to other populations.
Ethics oversight	Ethical oversight for human subject research by the Stanford University IRB.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

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	No sample size calculation was performed. Sample size was determined opportunistically to acquire longitudinal sampling. This sample size is sufficient to investigate prophage dynamics as it expands upon previous studies that have either had smaller sample size ($n = 1$) across similar timescales or explored shorter timescales (1 year).
Data exclusions	No data was excluded.
Replication	DNase treatment and PCR analyses were repeated on three separate occasions with complete DNase digestion and successful detection of circular phage DNA. One additional attempt had incomplete DNase digestion and we were therefore unable to evaluate phage particle presence accurately using this method.
Randomization	Not applicable. There were no experimental groups for participants in this study.
Blinding	Not applicable. There were no experimental groups for participants in this study. Sequencing data from all sample collection groups were processed through the same computational pipelines.

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 and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

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

Plants

Seed stocks

Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.

Novel plant genotypes

Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.

Authentication

Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosaicism, off-target gene editing) were examined.