npj | biofilms and microbiomes

Article

Published in partnership with Nanyang Technological University

https://doi.org/10.1038/s41522-025-00674-1

Prophages in the infant gut are pervasively induced and may modulate the functionality of their hosts

Tamsin A. Redgwell[1,8], Jonathan Thorsen[1,2,8], Marie-Agnès Petit[3], Ling Deng[4], Gisle Vestergaard[5], Jakob Russel[6], Bo Chawes[1], Klaus Bønnelykke[1], Hans Bisgaard[1,7], Dennis S. Nielsen[4], Søren Sørensen[6], Jakob Stokholm[1,4] & Shiraz A. Shah[1]

Gut microbiome (GM) composition and function is pivotal for human health and disease, of which the virome's importance is increasingly recognised. However, prophages and their induction patterns in the infant gut remain understudied. Here, we identified 10645 putative prophages in 662 metagenomes from 1-year-old children in the COPSAC2010 mother-child cohort and investigated their potential functions. No core provirome was found as the most prevalent vOTU was identified in only ~70% of the samples. The most dominant cluster of vOTUs in the cohort was related to Bacteroides phage Hanky p00', and it carried both diversity generating retroelements and genes involved in capsular polysaccharide synthesis. Paired analysis of viromes and metagenomes from the same samples revealed that most prophages within the infant gut were induced and that induction was unaffected by a range of environmental perturbers. In summary, prophages are major components of the infant gut that may have far reaching influences on the microbiome and its host.

The human gut microbiome is a complex and diverse ecosystem that is established at birth and becomes increasingly diverse over the following few years, until a stable 'adult-like' composition is reached during preschool years[1–5]. Within this period of maturation, several factors have been associated with differential development of the gut microbiome. The most well studied of these is birth mode[6,7], but factors such as medication use[8], diet[2,9,10], growing up in rural or urban environments[11–13], and the influence of siblings and pets have also been found to influence bacterial composition[14,15]. Early life bacterial gut microbiome dysbiosis has been associated with a range of disease outcomes in later life including asthma[16,17], allergy[18–20], and inflammatory bowel disease[21]. Whilst much work has been done on the bacterial component, recent studies have also shown the gut viral community to be altered in certain diseases[22–26], suggesting that they may also play a role in disease etiology. Viruses make up a large proportion (up to 5–6% of total DNA)[27] of the gut microbiome and are collectively referred to as the 'gut virome'. Their role in the gut microbiome is a research area of growing interest although much less is yet known about them compared to their bacterial counterparts.

Although a healthy human gut virome does contain some eukaryotic viruses, here we consider only bacteriophages (phages) that infect the bacteria in the gut and make up the vast majority of the human gut virome. Shifts in the phage composition of the gut have been associated with an increasing number of diseases such as Crohn's disease, Ulcerative Colitis[22,28], and arthritis[29]. These phages are thought to be virulent, but previous work has shown that the gut virome contains a significant proportion of temperate phages as well[30–32]. Temperate phages can integrate into the genomes of their bacterial hosts and be maintained through generations either as a part of the host genomes or as a plasmid-like genetic element; but can also return to the lytic cycle if the bacterial host is stressed by external factors[33]. While less than 20% of the phages found in faecal samples from adults were predicted to be temperate[31], they appear to be dominant in the infant gut virome[32] and carry a specific association with later risk of asthma[34]. Previous studies have also found that multiple strains of bacteria from the gut contain prophages, suggesting that lysogeny is a widespread phenomenon[35]. Whilst virulent phages infect and kill their host cell, temperate phages may be under a selective pressure to provide useful functions to their bacterial hosts.

[1]Copenhagen Prospective Studies on Asthma in Childhood, Copenhagen University Hospital, Herlev-Gentofte, Ledreborg Allé 34, DK-2820 Gentofte, Denmark. [2]Novo Nordisk Foundation Center for Basic Metabolic Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark. [3]Micalis institute, INRAE, Agroparistech, Université Paris-Saclay, Jouy en Josas, France. [4]Section of Food Microbiology and Fermentation, Department of Food Science, University of Copenhagen, Rolighedsvej 26, 1958 Frederiksberg C, Denmark. [5]Technical University of Denmark, Section of Bioinformatics, Department of Health Technology, 2800 Kgs Lyngby, Denmark. [6]Department of Biology, Section of Microbiology, University of Copenhagen, Copenhagen, Denmark. [7]Deceased: Hans Bisgaard. [8]These authors contributed equally: Tamsin A. Redgwell, Jonathan Thorsen. e-mail: shiraz.shah@dbac.dk

Prophages can influence the metabolism and function of their bacterial hosts by harbouring morons – genes that are not essential for the phage itself, but may benefit the bacterial host by providing additional functions that increase its fitness[36]. These can include antibiotic resistance genes[37–39], and toxins or virulence genes that increase the fitness of pathogenic bacteria during infection[40–42]. The presence of a prophage within a bacterium also has the added benefit of being protective from infections from other phages. Superinfection exclusion is a method by which prophages prevent infection of their host by related or more distant phages[43–45]. This is achieved by a variety of methods including alterations to the cell membrane[46,47], repressor-based immunity[48], and inhibition of DNA translocation into the cell cytoplasm[49,50], amongst others. Whilst there are several potential benefits to the bacterium from carrying a prophage, negative effects have also been identified in some cases. For example, in a monoxenic mouse model system the carriage of lambda prophage in *Escherichia coli* was detrimental to the host bacterium due to frequent reactivation of the prophage[51]. Additionally, in *Streptococcus pneumoniae* the carriage and expression of prophage element Spn1 has been shown to be detrimental to the fitness of the pathogen, by reducing its ability to colonise the nasopharynx[52]. Whether their effects are positive or negative, a growing body of evidence points to temperate phages being key components of the gut microbiome that can influence the bacterial host assemblage diversity and functional potential[31,35,53,54].

Since most phages within the human gut are still uncharacterised, cataloguing and quantifying them within sequenced virome samples has been challenging. Recently however, huge catalogues of human gut phages have been published, e.g. the Metagenomic Gut Virus (MGV) catalogue[55] and the Gut Phage Database (GPD)[56], and these aid researchers in cataloguing and tallying their samples by simply mapping virome reads, or even shotgun metagenomics reads against them. However, as these databases are based primarily on assembled bulk metagenomics data, it is still unclear if the genomes they contain originate from actively blooming viral populations or whether they comprise fragments of chromosomally integrated prophages that might even be inactive. Likewise, numerous major human virome studies have relied on bulk metagenomics for profiling the gut virome[57,58]. Again, here it is uncertain to what extent such profiling covers propagating viruses or integrated prophages.

One approach to resolving this is combining shotgun metagenomics data in addition to bonafide virome sequencing data for the same samples, making it possible to distinguish actively induced prophages from dormant ones. It is thought that temperate phages spend most of their lives as prophages, and this is supported by experimental models where induction is only triggered following strong chemical stress[59]. However, within natural environments, little is known about the overall extent of prophage induction, and analysing paired metagenomic and viromics for the same samples could answer that question systematically. In this work we utilised a set of 662 previously sequenced infant gut metagenomes[60] to identify prophages and explore their role by analysing their accessory genes. The samples originate from children of the Copenhagen Prospective Studies on Asthma in Childhood 2010 ($COPSAC_{2010}$) cohort, an ongoing mother-child cohort study followed since pregnancy and throughout early life with exhaustive phenotyping and sample collection. This allows for statistical testing of hypotheses about the biology of gut prophage composition during infancy and its potential link to later disease. Moreover, viral metagenomes (viromes) have previously been deeply sequenced for the same samples[32]. This configuration of deep metagenomics and viromics for the same samples allowed us to distinguish active prophages from dormant ones. By using the paired metagenome and virome data we were able to estimate if the prophages were actively induced and whether this was affected by external variables such as antibiotic usage. While our work is consistent with the current hypothesis that the high virome diversity in early life stems from induced prophages, it goes on to show that this induction is pervasive and constitutive, such that most prophages in the infant gut are induced most of the time.

Methods

Study design and workflow

An outline of the study design summarising the workflow described below is shown in Supplementary Fig 1.

Sample collection

The $COPSAC_{2010}$ cohort is an ongoing mother-child cohort study of 738 pregnant women and 700 children that have been followed from week 24 of pregnancy, in a protocol designed from the first COPSAC birth cohort ($COPSAC_{2000}$)[61]. The infant faecal samples studied here were collected at 1 year of age either at the research clinic or at home by the parents following detailed instructions. All samples were mixed with 1 mL of 10% vol/vol glycerol broth (Statens Serum Institut, Copenhagen, Denmark) and stored at −80 °C until use[16].

Metagenome and virome sequencing and assembly

The same samples were used for both metagenome and virome sequencing, and both datasets have been published previously[32,34,60]. Below, we have reproduced the relevant methodology descriptions from the original publications.

DNA extraction for metagenomic sequencing, taken from Stokholm et al.[16]

Genomic DNA was extracted from the infants' samples using the PowerMag® Soil DNA Isolation Kit optimized for epMotion® (MO-BIO Laboratories, Inc., Carlsberg, CA, USA) using the epMotion® robotic platform model 5075 (Eppendorf) according to the manufacturer's protocol with the following alterations to the workflow: 150–250 µL of the samples were added to the 96-well bead plate containing 750 µL bead/RNase A Solution and 60 µL lysis solution. Centrifugation steps were performed at 3220xRCF for 9 min. Removal of enzymatic inhibitors and DNA purification was performed as described by the manufacturer. Finally, the DNA was eluted with 100 µL Tris buffer (10 mM, pH 7.5). DNA concentrations were determined using the Quant-iT™ PicoGreen® quantification system (Life Technologies, CA, USA). Extracted DNA was stored at −20 °C.

Metagenomic sequencing for 1-year fecal samples and data processing, adapted from Li et al.[60]

Samples were prepared with the Kapa Hyper Prep kit (for Illumina) (KAPA Biosystems, Wilmington, MA, USA). Paired-end (150 bp) sequencing of the 663 samples in the DNA library (1-year samples) was performed with the Illumina NovaSeq apparatus by Admera Health (USA). Out of these 663 samples, only one failed to produce a library. In total, 662 gut samples at 1 year of age were sequenced for this study, generating between 32.6 and 215 million 150-bp paired-end reads per sample (mean ± SD: 48 ± 15.5 million reads). The samples were sequenced in a single batch to avoid any batch effect. Bioinformatic preprocessing was parallelized using GNU Parallel version 20180722[62]. Sequencing adapters were removed using BBDuk, from BBTools version 38.19 (https://sourceforge.net/projects/bbmap/), using the default options with the following exceptions: "ktrim = r k = 23 mink = 11 hdist = 1 hdist2 = 0 ptpe tbo". Low-quality sequences and reads shorter than 50 bases were filtered out using Sickle version 1.33[63]. Human contamination was filtered out using the BBMap feature of BBTools, with default values. The final dataset contained between 14 and 211 million clean reads per sample (mean ± SD: 46.7 ± 15.5 million reads). Clean reads were assembled with SPAdes version 3.12.0 using default metagenomics settings[64]. Metagenomics diversity was analysed using Nonpareil version 3.30, in kmer mode[65].

Virome extraction for 1-year fecal samples, adapted from Deng et al.[66]

Virome Isolation from Feces. After spiking with known phages, samples were poured into a stomacher filter bag (Interscience BagPage, 100 mL, Saint-Nom-la-Bretèche, France). The mixture was homogenized (Stomacher 80, Seward, UK) for 120 s at the high level setting. Homogenized samples, from the other side of the filter in the bag, were transferred to

50 mL tubes and centrifuged at 5000 × *g* for 30 min at 4 °C. After centrifugation, the supernatant was filtered through a 0.45 µm PES filter (Minisart® High Flow Syringe Filter, Sartorius, Göttingen, Germany) into the bottom of the outer tube of a Centriprep 50 K device (Millipore, Burlington, MA, USA). Afterwards, the filtrate was purified and concentrated using the Centriprep 50 K device by centrifuging at 1500× g three times in a row, first time for 30 min, second time for 10 min, and third time for 3 min. Extra centrifugation time was sometimes applied to allow the liquid level in the inner tube to be similar to the outer tube. The liquid filtered into the inner tube was poured off after each centrifugation step. A volume of 200 µL SM buffer was added to the inner tube at the end and centrifuged for 3 min. After the final centrifugation, 140 µL of the concentrated virome solution remaining in the outer tube was collected. The Centriprep filter membrane was cut out and added to the virome solution before storing at −80 °C until nucleic acids extraction. The remaining volume was stored at 4 °C for plaque assays.

Nucleic acid extraction of virome from feces. The concentrated virome solution and the cut filter membrane was first treated with 1 µL of 100 time diluted Pierce™ Universal Nuclease (Thermofisher Scientific, Waltham, MA, USA) for 5 min at room temperature, then the QIAmp viral RNA mini kit (Qiagen, Hilden, Germany) was used for viral DNA/RNA extraction following the procedures described by the manufacturer with modifications as described in ref. 67. Next, 10 µL of the extracted nucleic acids were amplified through Multiple Displacement Amplification (MDA) using the Genomephi V3 kit (GE Healthcare Life Sciences, Marlborough, MA, USA) following the instructions of the manufacturer, but the amplification time was shortened to 30 min (from 90 min). Finally, the amplified DNA was cleaned using a Genomic DNA Clean & Concentrator™ Kit (Zymo Research, Irvine, CA, USA) following the manufacturer's protocol.

Virome sequencing and QC, adapted from Shah et al.[32]

Virome libraries were sequenced on the Illumina HiSeq X platform to an average depth of 3 Gb per sample with paired-end 2× 150 bp reads. Satisfactory sequencing results were obtained for 647 out of 660 samples. Virome reads were quality filtered and trimmed using Fastq Quality Trimmer/Filter v0.0.14 (options -Q 33 -t 13 -l 32 -p 90 -q 13), and residual Illumina adaptors were removed using cutadapt (v2.0). Trimmed reads were de-replicated using the VSEARCH[68] (v2.4.3) derep_prefix.

Putative prophage contig identification and classification

Putative prophage sequences were identified using a combination of two methods: DeepVirFinder (v1.0)[69] and VIBRANT(v1.0.1)[70]. Assemblies from all 662 metagenomes were concatenated into a single pool and filtered to those above a minimum length of 4 kb. These assemblies were run through DeepVirFinder (v1.0) using default parameters after creating models from all known phage genomes, downloaded from the Millardlab database in September 2019[71]. Resulting contigs were filtered to include only those with a *p*-value of <0.05 after FDR correction. The assemblies were also run through VIBRANT (v1.0) setting the nucleotide input and parallelisation options only (VIBRANT_run.py -i assemlies.fna -t 16 …). Only those predicted phages of 'medium quality' and above were used further. Both sets of output were compared to the Refseq+plasmid database (available at https://mash.readthedocs.io/en/latest/tutorials.html) with a cut off of 95% identity using MASH (v2.2)[72] and any contigs with matches that might be contamination were removed. Resulting contigs from both methods were then combined and dereplicated at 95% ANI with dedupe2.sh[73]. This set was also run through CheckV(v0.7.0)[55] to give the details required for the minimum information about uncultivated viral genomes (MIUViG)[74].

Whilst measures have been taken to remove potential bacterial contamination in these sequences by applying appropriate cut-offs with the tools used, there is always the possibility that some bacterial traces still remain and this should be considered in any future analysis. Additionally, whilst the phages identified in this work are referred to as prophages, it is important to remember that this assignment is putative, as their lifestyle has not been experimentally validated and that DNA from phage particles could also contribute to the metagenomic assemblies in addition to the DNA from bacterial chromosomes.

Predicted prophages were clustered using a network analysis performed with vCONTACT2 (v0.9.8)[75], using the RefSeq release 88 database, with all other sequenced bacteriophages included using the Millardlab database as of January 2021[71]. The resulting network was visualised in Graphia (v2.1)[76].

Prophage annotation and functional analysis

Sequences were annotated using prokka (v1.14.5)[77] and a custom database made from all phage genes using the Millardlab database from September 2019. The –add-genes and –locus-tags options were also used. Resulting amino acid files were clustered at 90% identity using CD-Hit (v4.8.1)[78] and representative sequences from each cluster were analysed using EggNOG-mapper(v2.0)[79] and default parameters. Phage lifestyle prediction was calculated in the same way as Cook et al.[80] using HMM profiles for proteins that indicate a temperate lifestyle, and hmmscan (v3.3)[81].

Identification of previously isolated temperate phages

The genome sequences of a set of *E. coli* temperate phages that were isolated from the same infant faecal samples used here[82], were used to evaluate how well the temperate phage identification methods worked, and as reference genomes for comparison. To identify if any of the predicted sequences were those of the previously identified coliphages the reference genomes were dereplicated at 95% ANI to match that of the predicted contigs. The dereplicated contigs and the prophage contigs were then mapped against each other with minimap2 using the -asm20 option[83]. The sequences of any hits with >70% of the target contig covered was then extracted and clustered using Cluster_genomes.pl (v5.1)[84] to agglomerate contigs that were >95% similar over 90% of the genome. A cut-off previously accepted as the same genome[85], and the longest representative of the cluster was kept as the representative sequence.

Distribution of prophages in individuals

A set of reference contigs was constructed using the vOTUs identified from the metagenomes. The genomes of any remaining temperate coliphages isolated from the samples that had not been identified as vOTUs using the method described above were also added. Additionally, a set of 249 reference crAss phage genomes were downloaded from the dataset constructed by Guerin et al.[86], with the aim of capturing the diversity of the crAss-like family of viruses that are abundant in the gut. This set of predicted phages, coliphages and crAss phages were then dereplicated at 95% ANI using dedupe2.sh to remove any remaining redundancy[73].

Trimmed and QC'd reads from individual metagenomes were mapped against the set of reference contigs using bbsplit.sh with random mapping of ambiguous reads, and a minimum identity of 0.95 with the covstats option implemented[73]. A contig was considered present in a metagenome if there was coverage of >=1X across >=70% as used by Roux et al.[85]. Abundances were then calculated as counts per million (CPM). To determine how often the prophages are present in the metagenomes a binary presence/absence matrix was used so that extreme outliers in abundance would not skew results. The sum of the presence for each prophage was calculated and sorted to identify the prophages present in the most metagenomes. Alternatively, to determine how many prophages appeared in each child, the sum of presence in each metagenome was calculated.

When characterising the distribution of crAssphages in the samples the reference genomes along with any vOTUs that clustered together with them in vCONTACT were considered crAss-like prophages. The sum of presence of this subset of prophages was calculated from the presence/absence matrix.

Host prediction

Bacterial hosts of the prophages were predicted using CrisprOpenDB (v1.0)[87] with 1 mismatch allowed, and the host with the most prophages

predicted to infect them were identified in R and the host with more than 1% of the prophages predicted to infect them were visualised.

Functional analysis of the most abundant prophage cluster

Proteins from all members of cluster1819 (the most abundant prophage cluster) were extracted and clustered at 90% amino acid identity (AAI) using CD-Hit[78] to remove redundancy. These were then analysed with EggNOG-mapper(v2.0)[79] as previously described to assign Clusters of Orthologous Groups (COG) categories (Supplementary Table S2). Additionally, resulting Kegg Orthology (KO) codes were mapped onto metabolic pathways using KEGGmapper[88].

Phylogenetic analysis of the most abundant prophage cluster

An initial blast search showed that *Bacteroides* phage Hanky p00' (Hankyphage) was the only phage with significant sequence similarity to the members of cluster1819: the high quality vOTU_03578 was used as a representative of the cluster and had 99.90% identity and 74% query coverage with *Bacteroides* phage Hanky p00'; therefore, putative terminase genes from all cluster members were identified through analysis of the Hankyphage p00' genome. The terminase protein sequence of Hankyphage p00' was downloaded and used to identify the same protein in the cluster members using HMMsearch[66], as no protein had been annotated as such. The protein sequences were aligned with ClustalW in MEGA(v10.1.8) and a maximum likelihood tree was also produced in MEGA using the JTT model and 100 bootstraps[89]. The tree was visualised and manually coloured in iTOL[90].

Identification of diversity generating retroelements in the most abundant prophage cluster

Diversity generating retroelements were identified in members of cluster1819 by using both the myDGR web server[91] and MetaCSST(v1.0)[92] tools. To predict whether the target genes identified were putative tail fibre genes, as has previously been suggested[93], the proteins from all members were analysed with PhANNs(v1.0.0)[94] and the most significant hit to a tail fibre gene was carried forward. These were then compared with the results from the previous tools.

Determining active prophages

The paired nature of the metagenome and virome sequencing of the samples allowed for a novel exploration of whether the predicted prophages were induced at the time of sampling. Individual virome sample reads were mapped against the original metagenome assemblies containing the prophages that had been excised by the prediction tool VIBRANT[70] using bbsplit.sh with a minimum identity of 0.95 and random mapping of ambiguous reads[73]. A total of 4291 prophages were eligible for this analysis. Using the predicted coordinates of the prophages, the coverage of each background bacterial and predicted prophage region of an assembly was extracted from a bam file that had been sorted and indexed using samtools[95]. For this analysis it was assumed that there was only a single prophage region per contig; there were only a minority of cases where multiple prophage regions had been predicted for a contig and had passed the quality cut-offs used in this work. Of note, the assumption of a single prophage region may lead to a small number of false negatives in the induction analysis. A small number of virome samples were also mapped against three large chromosomal contigs that were not predicted to contain any prophages as a negative control. The number of reads mapped to sections of 40 kb (the mean size of prophages in this work) were extracted from different regions of the assembly, in the same way as described above to mimic the presence of a prophage and allow us to test for induction in these negative controls.

To determine statistically if prophages were induced, the number of reads mapped to the bacterial part of the assembly and the number of reads mapped against the predicted prophage part were tested for a binomial distribution using pbinom in R(version 3.6.1), and the resulting p-values were corrected for multiple testing using the Bonferroni method. Frequency of significant induction across samples was tested against predicted host and mean RPK per prophage using linear models. Induction patterns of vOTUs were derived from a binary matrix of vOTUs vs samples (Bonferroni-significant induction yes/no) using principal component analysis (PCA) with vOTUs as observations and samples as features to visualise similarities between vOTUs regarding which samples they were induced in. In parallel, the same matrix was transformed to a euclidean distance matrix and analysed with PERMANOVA (adonis2 from the R-package 'vegan' v. 2.6-4; with option by = "margin") to quantify similarities in induction patterns associated with predicted host (top hosts, genus level, vs others as shown in Fig. 5A) and viral cluster (each top cluster vs others as shown in Fig. 5C).

Environmental and clinical factors

To study factors potentially influencing prophage induction, we compared children according to key factors associated with microbiome composition: Antibiotics (yes/no), delivery mode(c-section/vaginal), furred pets(yes/no), gastrointestinal infection (yes/no), living environment(urban/rural), and siblings(any/none). Antibiotic exposure was defined as any prescription of ATC code starting with J01 (Antibacterials for systemic use) recorded at the 1-year visit in the Danish prescription register. Delivery mode, furred pets, birth address, and any siblings in the home was assessed by parental interview at the planned 1 week and 1 year visits to the research clinic. Living environment was defined by converting addresses to coordinates and mapping to 100x100m raster maps from the CORINE database of European land cover (https://land.copernicus.eu/) in a 3 km radius and performing PAM clustering on the composition of 5 major land cover types, as previously described in detail[13]. Gastrointestinal infections were assessed from a prospective symptom diary as any diarrhoea or vomiting within 7 days of the sample collection. Associations between these factors and induction rates were assessed using Wilcoxon tests of sample-wise induction percentages (within-sample matched vs non-matched virome-contig pairs analysed separately) and multiple testing was controlled using false discovery rate (FDR) adjustment and expressed as q-values.

Results

Identification and classification of novel prophages in the infant gut

From the 662 infant metagenomes obtained at age 1 year, we identified 10645 vOTUs after dereplication (Supplementary Table 1).

We grouped these viral contigs into approximate genus or subfamily-level classifications using vCONTACT2 and included the genomes of all sequenced bacteriophages as references, resulting in 2934 clusters. Of these, 364 were singletons and 2221 were outliers. Of all the clusters identified 953 were comprised solely of vOTUs identified in this work and of these 953 clusters, 177 were made up of a single member (Fig. 1). The hosts of 65% of the vOTUs could be predicted; the most common assignment at the genus level was *Bacteroides* (12.2%), followed by *Salmonella/E. coli* (6.3%) and *Bifidobacterium* (5.8%) (Fig. 2A)

Previous work using these samples resulted in isolation of 35 *Escherichia* temperate phages and sequencing of their genomes[82]. There were five vOTUs that shared significant similarity with these isolated phages (Table 1). The longest sequence from each cluster was kept as representative, resulting in four predicted sequences being replaced with the isolated phage genomes and one isolated phage genome replaced by a predicted prophage genome. The ability to identify five vOTUs with similarity to previously isolated phages may reflect a level of microdiversity within this group of closely related phages. Both microdiversity and close similarity of sequences within a sample are known to cause problems with assembly and may explain why more were not found[96,97]. It may also reflect the fact that the isolated phages may not be abundant enough in the metagenomic sequence data to assemble fully.

Abundance analysis suggests no core provirome is established in infants

The distribution of the number of prophage vOTUs in each sample shows a mean of ~100, with values as low as four and as high as 400. No single vOTU

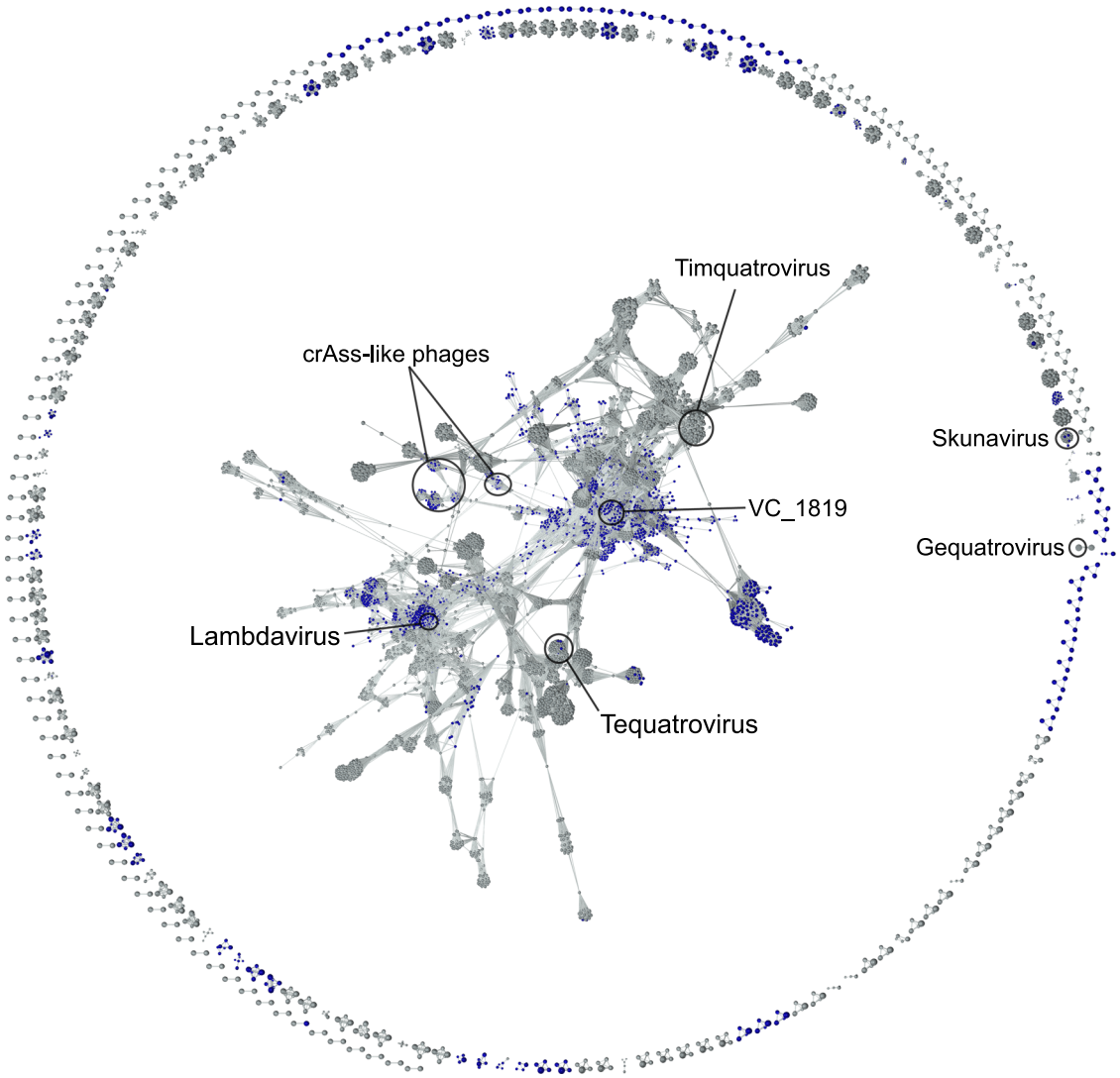

Fig. 1 | vCONTACT2 network analysis of vOTUs from this study and a database of phage genomes extracted from millardlab.org in January 2021. Each node represents a viral genome: vOTUs identified in this work are coloured in blue and reference genomes are grey. The largest and key viral families have been annotated, and viral clusters characterised in this work (VC_1819 and additional crAss-like prophages) have also been annotated. The number of clusters highlighted in blue, and their distribution throughout the network reflects the diversity of vOTUs identified.

was found in all children. The most widespread vOTU was found in ~70% of the children (Fig. 2B), which is below the 95% cut off used in this work to designate a prophage as core. Using a 50% cut off that has been used in previous work for the same designation[34,98] results in one additional vOTU. Whilst no individual prophage could be found in all samples, the top 50 most prevalent prophages were spread between only eight viral clusters (excluding those without an assigned cluster) (Fig. 2B), showing more conservation of the genus/subfamily level than of individual viral contigs. To further examine this, we compared the prevalences of each viral cluster analogously to Fig. 2B and found no evidence of a core provirome at the cluster level either (supplementary Fig 2), with only one viral cluster present in more than 50% of the samples, excluding singletons and outliers.

CrAss-like prophages were present in 195 (29%) of the samples sequenced, and if a sample had crAss-like prophages identified, it was likely to possess only one type, as only 38% of the crAss-positive subjects contained two or more types. The identification of 109 vOTUs that clustered together with the reference crAssphage genomes has also expanded our knowledge of crAssphage and crAss-like phages, particularly those of the infant gut: an environment where they are thought to be present much less frequently than in adults[86].

The functional potential of viral OTUs showed no significant patterns on the individual phage level

The percentage of all proteins involved in the different COG categories showed that the majority (64.9%) of proteins were assigned to category S – those of Unknown Function. Followed by categories reflecting viral replication – Replication, Recombination, and Repair (12.9%); Transcription (6.6%), and Cell Wall/Membrane Biogenesis (3%). Other categories were present in very small percentages of the total protein amount.

Cluster1819 is the most abundant phage cluster and contains an abundance of DGRs and morons

Cluster1819, containing 82 members, was found to be the most abundant in the samples (Fig. 3A) and is the second most prevalent cluster in the children (Supplementary Fig 2) so was characterized in more detail. Phylogenetic analysis of the large terminase gene revealed a single relative: *Bacteroides* Hankyphage p00' (Accession BK010646) (Fig. 3B). The bacterial host for Hankyphage was previously identified as a *Bacteroides* which is the same as the predicted host for many members of this cluster[81]. However, there were variations on this with some vOTUs predicted to infect *Prevotella* and *Butyricimonas*.

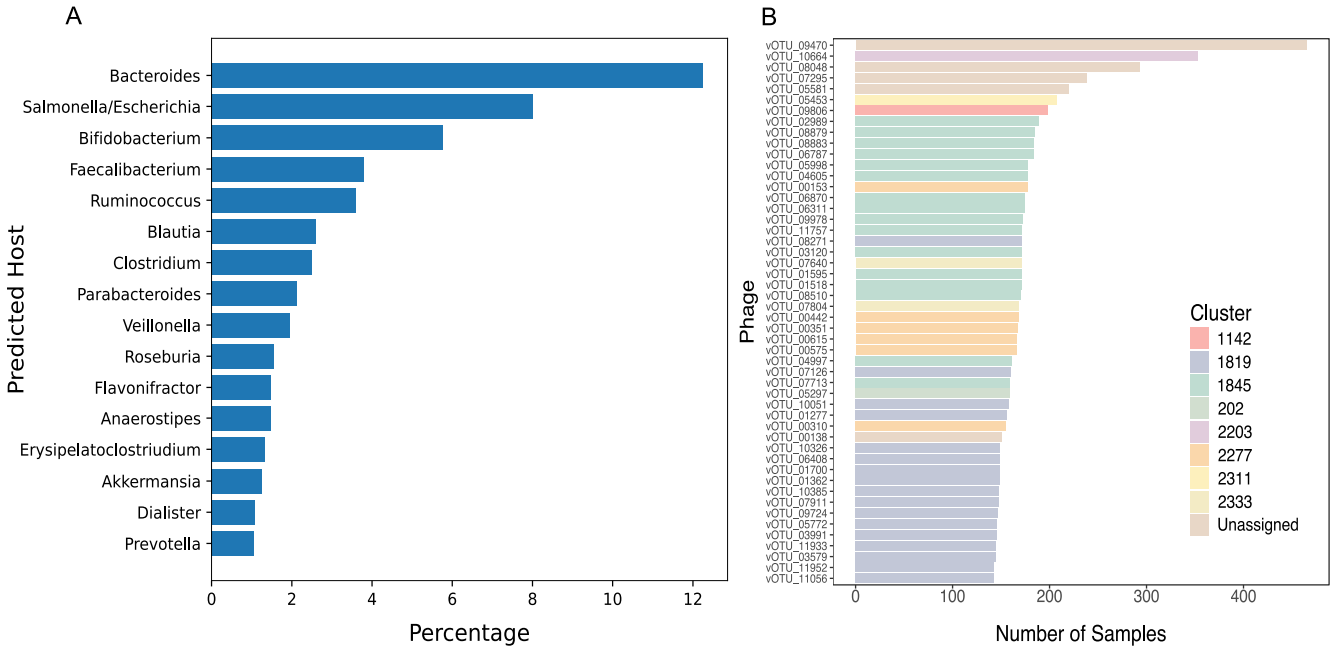

Fig. 2 | Characterising putative prophages. A Host prediction for the putative prophages. Only those that make up greater than 1% of the predicted hosts are shown. Salmonella and Escherichia have been grouped together, as CrisprOpenDB is known to not be able to differentiate between the two, due to the high number of Salmonella spacers available in the database. This is likely to be an overprediction of Salmonella. **B** Top 50 most prevalent vOTUs coloured by their viral cluster.

Table 1 | Percent identity of vOTUs identified bioinformatically in this work, and Escherichia coli temperate phages isolated in previous work[82]

Sequence kept	Length (bp)	Sequence replaced	Length (bp)	% Identity
Escherichia coli phage Lambda ev017	50126	vOTU_10421	43620	99.2
Escherichia phage mEp460 ev081	45865	vOTU_01494	44384	99.6
Escherichia coli phage P2 2H1	32662	vOTU_09503	32101	99.9
vOTU_05922	32386	Escherichia coli phage P2 4C9	32150	99.9
Escherichia coli phage ESSI2 ev015	30584	vOTU_03208	26975	88.8

The E.coli temperate phages had been isolated from the same samples used for metagenomic and virome sequencing, but sequenced independently. Five vOTUs were found to share sufficient percent identity to be classified as the same phage.

The combination of MetaCSST and MyDGR identified 53 diversity-generating retroelements (DGRs) in the cluster, an element that is present in Hankyphage. Of the cluster members 49/82 were found to contain a DGR, as some contained multiple, and the target sequences were used to predict the gene it would generate diversity in. PhaNNs was used to predict the structural genes for the cluster members including the tail fibre genes, commonly a target for DGRs; when the target gene sequences were compared to the structural gene predictions all target genes were predicted to be tail fibres.

The majority of identified proteins in cluster 1819 belong to category S – those of unknown function (Fig. 3C). This is followed by category L and represents proteins involved in replication, recombination, and repair; categories O and V are equally abundant and represent those proteins involved in post-translational modification, and defence mechanisms respectively. Cell cycle control and nucleotide transport and metabolism are also categories of note. Combining these COG codes with the KEGG pathway database gave a clearer overview of what host pathways could be affected by these phages. These pathways included dTDP-L-rhamnose biosynthesis and menaquinone biosynthesis amongst others (Supplementary Table S2).

Estimation of the proportion of putative active prophages

Previously, we have shown that the mean relative abundances of both virulent and temperate phages in the virome are highly positively correlated with the corresponding abundances of their cognate host bacteria in the metagenome, at least at the host genus level and study-wide[32]. However, differences could still exist between different prophage clusters or even between samples, holding important clues about their biology.

Here we used the results of the virome reads mapped against the large metagenome contigs containing prophages, to discover prophages that were potentially induced and present in the samples as actively propagating viral particles. A subset comprising 4291 of the predicted prophages were able to be tested for induction via a read mapping approach, as they had been excised from a larger assembly in the prediction process, thus allowing for a comparison against the bacterial background on the same contig. We quantified and tested induction as a degree of preferential mapping of virome reads inside the predicted bounds of the prophage compared with the rest of the contig; for examples see Fig. 4A, B. This showed that induction is a widespread phenomenon; 4041 (94.2%) of the prophages were induced in at least one sample and remained significantly so ($p = <0.05$) after Bonferroni correction, resulting in 4.59% significant virome-prophage contig pairs, see Fig. 4C. Only 250 prophages were never found to be significantly induced in any sample, see Fig. 4D. When only considering contigs with 100 reads or more mapping in a sample, 83,418 out of 321,232 pairs (26.0%) were found to be significant.

Comparing prophages across different predicted hosts, we found differences in the fraction of induced samples (Fig. 5A, log linear model $p < 2e$

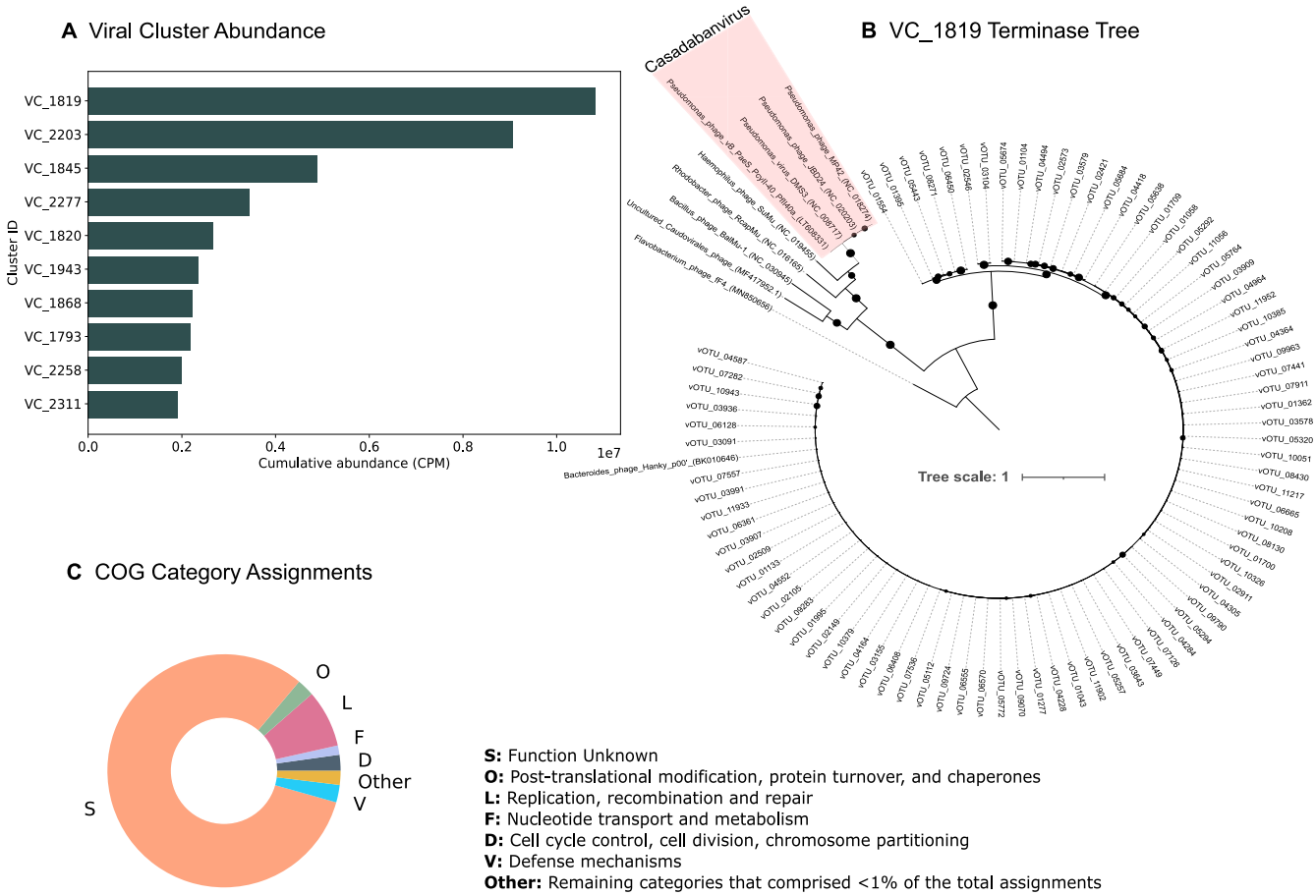

Fig. 3 | Characterisation of cluster1819. A Cumulative abundance of the top 10 most abundant viral clusters. **B** Phylogenetic tree based on a terminase protein alignment for members of cluster1819 and including the Bacteroides phage Hankyp00'. Only bootstrap values > 70% are shown as circles. The next most closely related phages were used as an outgroup due to their limited genome similarity. Any known taxonomy has been highlighted in red. **C** COG Category assignments for proteins belonging to members of cluster1819. The 'other' group is made up of those categories that comprised less than 1% of the total assignments.

−16). Among the genera with highest rates of induction were *Blautia*, *Bifidobacterium*, and *Erysipelatoclostridium*. The most abundant group of prophages in the metagenomes, cluster1819, was among the most commonly induced (Fig. 5B, C). Comparing the cluster against the rest of the predicted *Bacteroides* phages showed that the phages belonging to cluster1819 were more often significantly induced than the rest of the group, and to all prophages with different predicted hosts.

We found similar patterns of induction across prophages belonging to the same viral cluster (Fig. 5C), ie. which samples they were significantly induced in, suggesting that specific clusters of prophages are induced together within a sample (PERMANOVA, $F = 85.1$, $R^2 = 13.7\%$, $p < 0.001$). A similar phenomenon was seen when comparing induction patterns between vOTUs according to predicted host ($F = 20.2$, $R^2 = 7.0\%$, $p < 0.001$). This only attenuated slightly after including both viral cluster and predicted host in the model; both had significant contributions to the variation in induction patterns (Viral cluster, $F = 75.7$, $R^2 = 11.5\%$, $p < 0.001$; Predicted host, $F = 16.0$, $R^2 = 4.9\%$, $p < 0.001$).

The frequency of significant induction was also associated with the mean RPK of the prophages (log linear model, Supplementary Fig 3), however adjusting for this did not change the conclusions above.

Next, we examined induction among virome-prophage contig pairs originating from the same sample, compared with those from different samples. For matched virome-prophage contig pairs, a much higher rate of induction was seen (Fig. 6A) than for non-matched pairs. However, the mapping rate was also much higher among these pairs (Fig. 6B, C), which also influenced induction rates. Therefore, we considered both factors simultaneously and saw that in matched pairs, higher mapping rates were associated with higher rates of induction, up to around 75% in contigs attracting many reads. This was only partially seen for non-matched pairs, which increased until 100–200 reads and thereafter declined slightly (Fig. 6D).

Finally, we examined whether induction rates differed between samples according to environmental and clinical factors known to be associated with microbiome composition. We analysed matched vs non-matched virome-contig pairs separately, summarising per child to allow a fair comparison. We found no association to induction rate according to recent antibiotic treatment, delivery mode, having furred pets, or an urban vs rural living environment. Children who had siblings had a slightly lower induction rate than those without siblings, but this did not remain significant after FDR correction (Wilcoxon test, $p = 0.00495$, $q = 0.059$).

Discussion

This work sought to deeply characterise prophages in the infant gut, and to highlight novel aspects of the associated phage biology. By combining machine learning based identification methods we maximised our ability to identify integrated prophages from metagenomic assemblies. Of those identified, no single prophage could be found in more than 70% of the samples suggesting that there is no core set of prophages in the infant gut. Much debate has taken place previously over the existence of a core virome with early work on the topic identifying 23 phage contigs shared by at least 50% of samples from 62 individuals[98]; however, more recent studies with a larger sample size found that the most ubiquitous viral population was only present in 39% of the metagenomes used, and most of the populations were only sporadically detected at all.[99]. These results suggest that infant viral

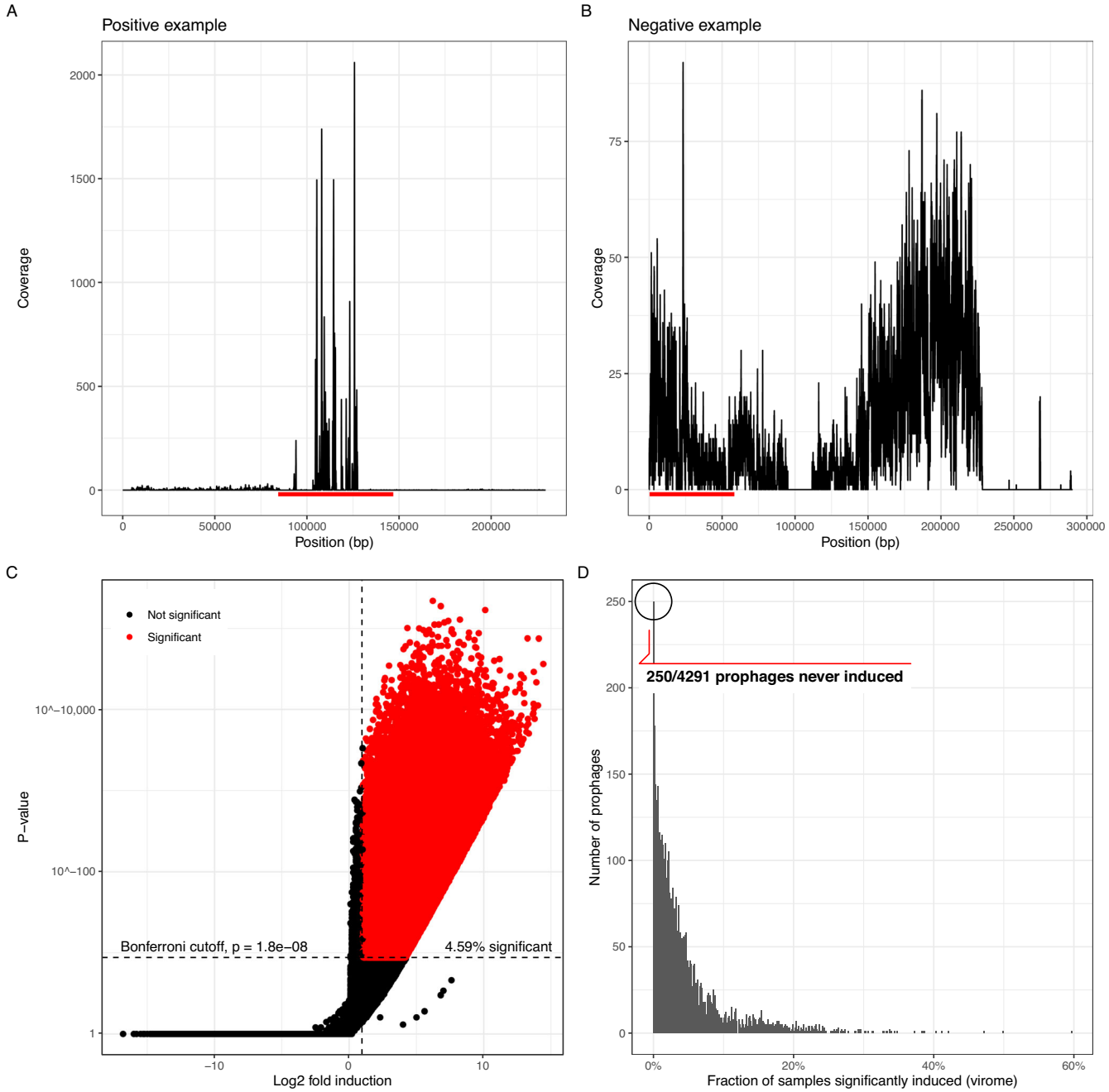

Fig. 4 | Evaluating induction in prophages using virome read mapping. **A**, **B** Looking across the entire prophage carrying contig, we assessed induction as differential virome read coverage inside the predicted prophage region (red lines) versus the rest of the contig, which was considered as background. These two examples were chosen to illustrate this phenomenon as a positive (**A**) and negative (**B**) example. **C** Volcano plot showing log2-fold induction vs. *p*-value (double log scale) distribution of all prophage/sample pairs. All prophage/sample pairs with an induction value > 1 and passing the Bonferroni cutoff were considered significant (red dots), which comprised 4.59% of the entire set. The red area looks larger due to massive overplotting in the lower part of the panel. **D** Histogram of all prophages by how many children they were significantly induced in, ranging from 0% (no children, 250 prophages) to ~60% of all the children.

communities are more unique to the individual than they are commonly shared. Considering the dynamic nature of the infant gut microbiome, it may not be surprising to find prophages distributed sporadically throughout the samples as they adapt to changing bacterial host abundances[100,101].

We were able to assign predicted hosts to ~65% of the prophages identified here with *Bacteroides*, *Salmonella*/*Escherichia*, and *Bifidobacterium* the most common hosts; all key members of the infant gut microbiome[1,16]. *Salmonella* and *Escherichia* predictions were grouped together as the tool used has difficulty distinguishing between them[87]. Thus, these are likely to be *Escherichia* infecting phages rather than *Salmonella*. It is important to remember that these are predictions and putative hosts have not been experimentally validated. We included some experimentally validated coliphages from the same samples with very high sequence similarity. While the detection of phages with complementing methods is a strength of the study, there is also a potential for bias from increased sensitivity to those coliphages.

We also specifically looked at the prevalence of crAssphages, a well-known and large family of phages that are widespread in gut viromes[86]. In addition to the reference crAss-like phages that were used we also identified 109 additional vOTUs that clustered together and so were considered part of

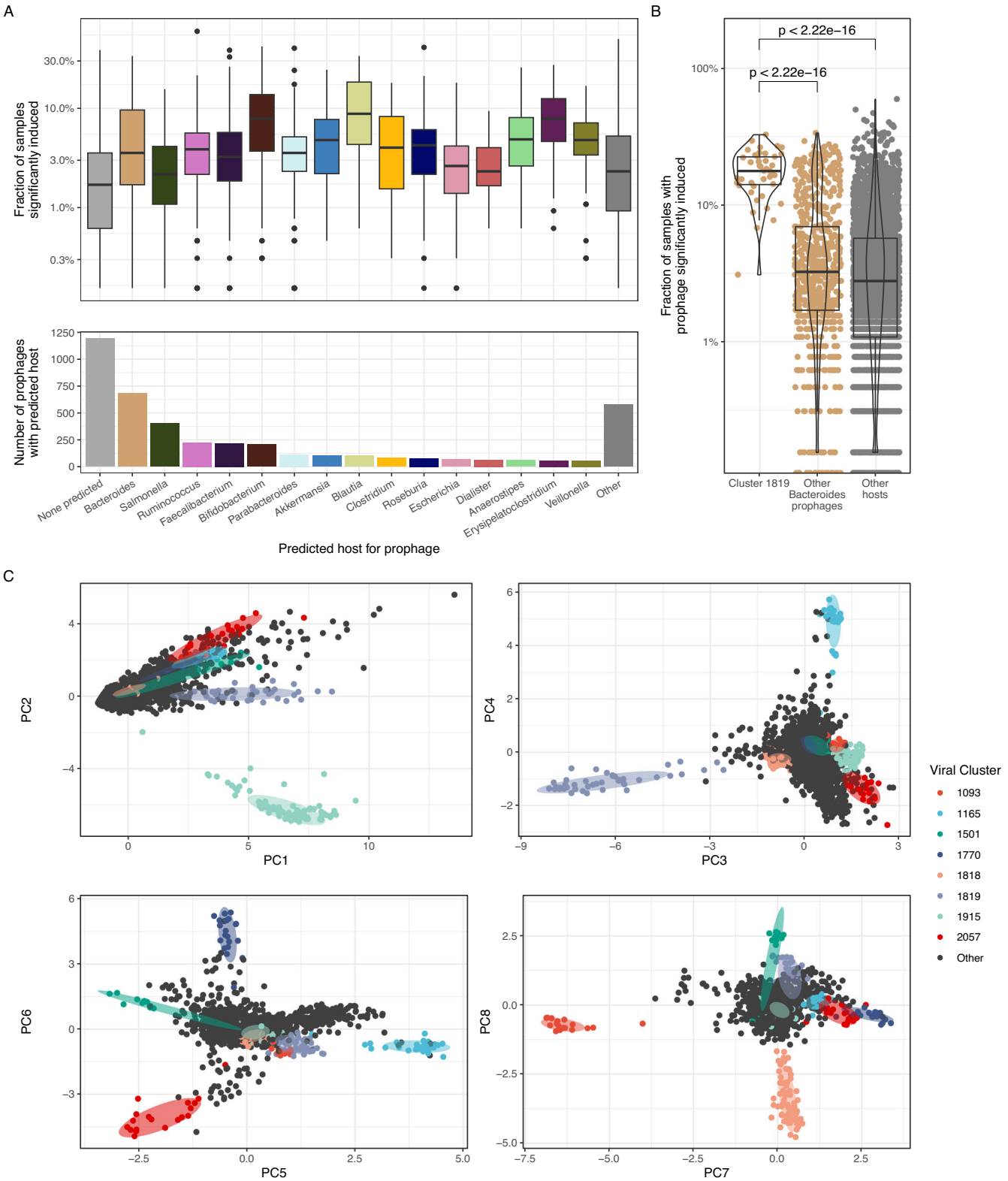

Fig. 5 | Comparing induction frequencies across vOTU and sample characteristics. A Induction fraction by predicted host of the prophages, showing sizable differences between hosts (overall $p < 2e{-}16$), even when adjusted for mean RPK of the prophages. **B** Highlighting the induction frequencies of cluster 1819 prophages are much higher than other prophages with Bacteroides as their predicted host, as well as all other prophages with different predicted hosts. **C** Principal Component Analysis (PCA), each dot is a prophage ordinated according to their induction pattern, ie. which children these prophages were induced in. Points which are close together signify prophages that tend to be induced in the same children. Points were coloured according to some of the major Viral Clusters that show homogenous inductions patterns, notably including cluster 1819.

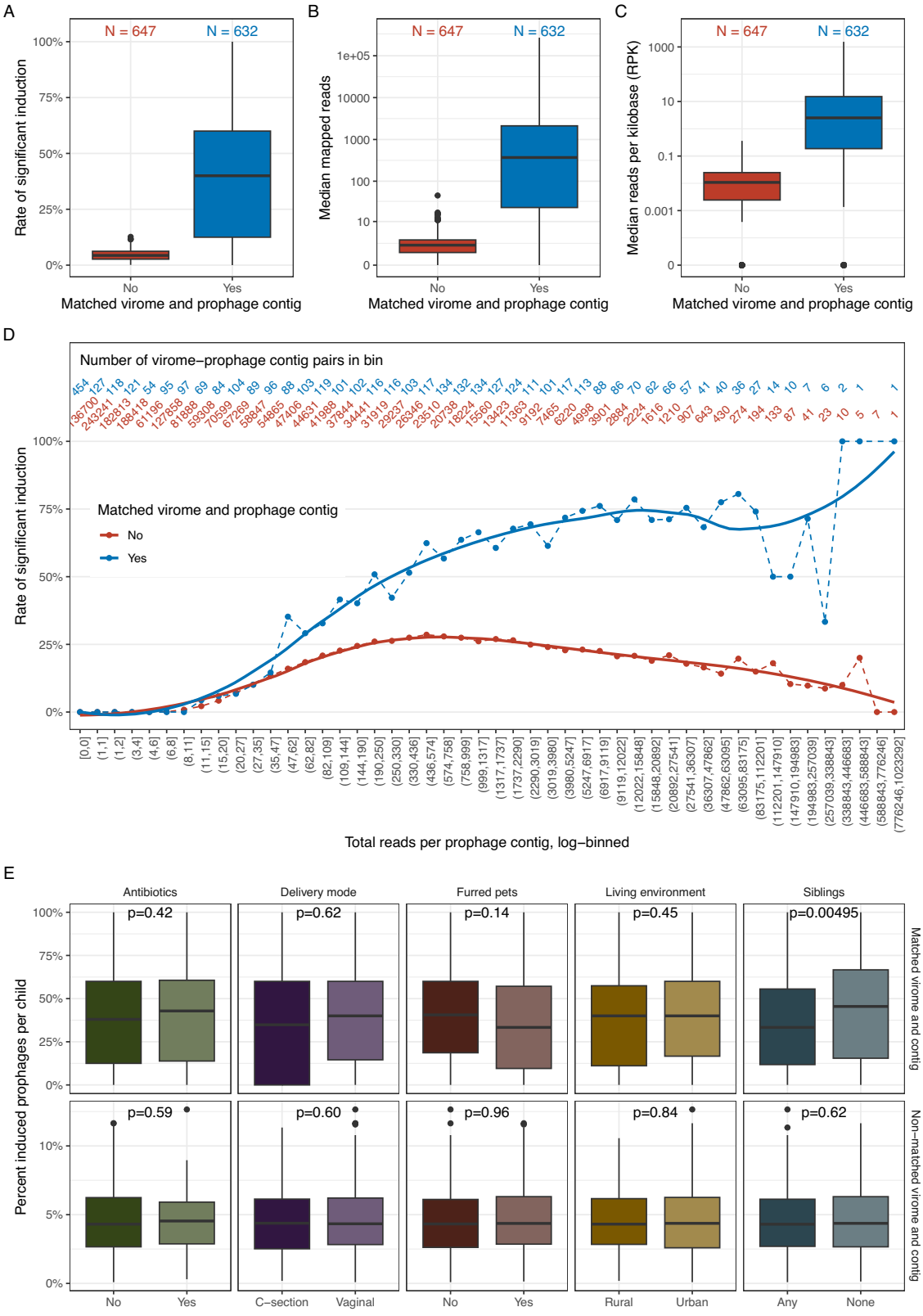

Fig. 6 | Investigating within-sample induction and contributions of environmental factors. Comparing all samples' viromes mapped against prophage contigs assembled in same vs other samples, expressed as sample-wise **A** mean induction rate (ie per-sample mean of prophages that were significantly induced in same vs other samples), **B** Median number of reads mapped to contig (per-sample mean of same vs other), and **C** Median reads per kilobase (RPK); i.e. length-adjusted mapping rate. **D** Analysis of within-sample vs between-sample induction estimates adjusted for number of reads mapped to contig, showing that especially contigs with high mapping rates exhibit high rates of induction. Rates of all virome-contig matches are summarized within bins of the number of total reads per contig. **E** Comparison of sample-wise induction rates split in sample-matched and non-matched virome-contig pairs, comparing samples according to key environmental factors. Children with siblings exhibit lower rates of induction compared to those without siblings.

the crAss-like group. Our results strongly support previous work that has suggested that crAssphages are not abundant or prevalent in infants. Initial reports on crAss-phages demonstrated an indeterminate infection strategy, showing no clear signs of lysogeny and unusual lytic infection behaviour[102]. Our results support the suggestion that at least some of the crAss-like phages are temperate[103].

Classification of the putative phages combined with the sequences of all sequenced bacteriophage genomes resulted in the creation of 2934 clusters with 364 singletons. Of all the clusters identified, 953 were comprised solely of vOTUs identified in this work and of these 953 clusters, 177 were made up of a single member.

A more in-depth analysis of the most abundant cluster of prophages revealed an interesting perspective into their potential role in the infant gut. The most abundant group, Cluster1819, falls within the proposed *Hannahviridae* family[32] and is comprised of 82 prophages closely related (genus or sub-family level) to *Bacteroide*s phage Hanky p00'. This phage was originally identified as an integrated prophage of *Bacteroides dorei* from metagenomic data[93]. Hankyphage has been predicted to be present in half of the human population from geographically distant regions, found to lysogenize at least 13 different species of *Bacteroides*[93], and now abundant in this young age group.

The broad host range of these phages is due to the possession of diversity generating retroelements (DGRs) that target tail fibres[93], which were also found in the present study. Over half of the members of the viral cluster identified here were found to contain at least one DGR, all of which were predicted to target tail fibres. This adaptation to infect different *Bacteroides* may be vital for their success in this dynamic environment. The cluster was predicted to infect a few different hosts; which could be an artefact of the host prediction method used or due to changing tail fibres. Without experimental evidence this cannot be validated but remains intriguing. Importantly, CRISPR-Cas systems are subject to horizontal gene transfer between related host genera, making CRISPR-based host predictions inherently uncertain, and especially insensitive for bacteria that do not normally carry them.

In addition to harbouring DGRs, members of the cluster also possessed several morons, or auxiliary metabolic genes, that may prove beneficial to their host or influence bacterial metabolism. We found genes involved in the dTDP-L-rhamnose biosynthesis pathway, which is responsible for biosynthesis of the O-antigen of lipopolysaccharides in Gram negative bacteria[104–106]. *Bacteroides* in particular are known to produce a number of phase-variable capsular polysaccharides (CPS)[107,108], which are involved in host-tropism of *Bacteroides*-targeting phages[107]. The phase-variable expression of these CPS creates diversity that may help to ensure host infection resilience by maintaining differentially susceptible subpopulations[107] and help the phage by superinfection exclusion of other phages[46,47]. This echoes the piggy-back-the-winner model proposed for the crAss-like phage crAss001 and its *Bacteroides* host[109]. Finally, the O-antigen is of importance for recognition of the human immune system and the pathogenicity of the bacterium[110,111], as well as the bacteria's ability to bind to and infect epithelial cells[112].

Another gene of interest found in the cluster was *menA*: a component of the menaquinone (vitamin K2) biosynthesis pathway, an important part of the electron transfer pathway in prokaryotes and vital to humans in the blood clotting process, and bone and nervous system health[113–117]. The importance of microbially synthesised vitamin K is debated due to the low amount of total vitamin K it would be contributing[118–120], which may be more significant in infants[121].

Further work is needed to characterise other families of prophages in the infant gut.

Our induction analysis shows that out of the subset of phages we were able to test, most prophages that were identified in a sample were also induced. This suggests that these prophages are an active part of the community and may play a prominent role in the shaping of the bacterial community in the gut. Previous work has suggested that the infant gut in particular may be dominated by temperate phages that may be induced due to the high turnover rate/constant maturation of the bacterial community during the first few years of life[30,31,35]. However, whether pervasive prophage induction is also characteristic of more mature gut microbial compositions is still not known. One study of mice colonised with the Oligo Mouse Microbiota community (OMM12) also found widespread prophage induction, but this could as well be attributed to their gnotobiosis, known for the stress it causes hosts[122].

Environmental conditions can lead to the induction of prophages from their bacterial hosts, leading to lytic replication and the production of progeny phages. A number of factors have been shown to induce prophages such as certain chemicals (mitomycin c) and antibiotics such as fluoroquinolones[123–125]. More recently, the use of common oral medications such as nonsteroidal anti-inflammatory drug diclofenac, and other antibiotics including ampicillin, norfloxacin, and ciprofloxacin were shown to induce prophages from bacterial isolates of the human gut[126]. Our work showed no major effects of the clinical and environmental factors tested on the proportion of induced prophages. A priori, we would have assumed that especially antibiotics could have potential for influencing the overall induction level, but no differences were seen. This may be due to the specific antibiotics used as most children received regular penicillins which may not have the same induction potential as the specific aforementioned drugs. It could also be due to time limitations of the method; evidence of induction may have already been turned over in the 4 weeks preceding sampling, and this may be why we cannot see it here. Furthermore, we only had one time point per child; sampling before and after treatment may better uncover changes in induction. We found a significant reduction in the overall induction level in children with siblings. However, this did not remain significant after FDR adjustment. Future clinical studies should examine the potential effect of siblings further.

Sequence composition and genomic structure may influence sequencing efficiency which can lead to uncertainty or noise in the mapping. Together, this highlights the need to be careful of interpreting the 'snapshot in time' of a population from a single time point and indicates the need for more longitudinal data to characterise the biology of prophages in the infant gut.

In summary, our results show that prophages of the infant gut form a diverse community that is different in each individual; no conserved core provirome of temperate phages was apparent at the vOTU level. Our work utilises a large infant cohort to support the previous observation that crAss-like phages are present in small numbers early in life. We also identified a novel cluster of prophages that are the most abundant in the metagenomes, which fall within the newly proposed *Hannahviridae* and are related to *Bacteroides* phage Hanky p00'. The possession of DGRs targeting tail fibres in members of this cluster suggest they may be able to infect a range of bacterial hosts. We also found evidence that they may modify host LPS through possession of components of the dTDP-L-Rhamnose pathway. Therefore, this group of phages possess elements that may allow them to maintain differentially susceptible subpopulations of their host bacterium, whilst also containing DGRs that could expand their host range. By utilising the paired metagenome and virome sequencing we were able to show that out of the identified prophages we were able to test, the majority of them were induced. However, testing induction against antibiotic usage and other factors revealed no significant associations, although this may be a reflection of the speed at which the evidence of induction is turned over in the gut, highlighting the need for more longitudinal data in the field.

Data availability

Original raw sequence data and metadata for all metagenomic samples has been previously reported under project PRJNA715601. Assembled and filtered prophages were submitted under accession ERZ2947210. Virome reads are available under project PRJEB46943. Participant-level personally identifiable data are protected under the Danish Data Protection Act and European Regulation 2016/679 of the European Parliament and of the Council (GDPR) that prohibit distribution even in pseudo-anonymized form, but can be made available under a data transfer agreement as a collaboration effort.

Code availability

The code used can be obtained from the corresponding author upon request, however, most tools have been run with standard parameters.

Received: 19 April 2024; Accepted: 21 February 2025;
Published online: 19 March 2025

References

1. Roswall, J. et al. Developmental trajectory of the healthy human gut microbiota during the first 5 years of life. *Cell Host Microbe* **0**, https://doi.org/10.1016/j.chom.2021.02.021 (2021).
2. Koenig, J. E. et al. Succession of microbial consortia in the developing infant gut microbiome. *Proc. Natl. Acad. Sci*., https://doi.org/10.1073/pnas.1000081107 (2010).
3. Yatsunenko, T. et al. Human gut microbiome viewed across age and geography. *Nature* **486**, 222–227 (2012).
4. Palmer, C., Bik, E. M., DiGiulio, D. B., Relman, D. A. & Brown, P. O. Development of the human infant intestinal microbiota. *PLoS Biol.* **5**, e177 (2007).
5. Bergström, A. et al. Establishment of Intestinal Microbiota during Early Life: a Longitudinal, Explorative Study of a Large Cohort of Danish Infants. *Appl. Environ. Microbiol.* **80**, 2889–2900 (2014).
6. Dominguez-Bello, M. G. et al. Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. *Proc. Natl Acad. Sci.* **107**, 11971–11975 (2010).
7. Stokholm, J. et al. Delivery mode and gut microbial changes correlate with an increased risk of childhood asthma. *Sci. Transl. Med*. **12**, https://doi.org/10.1126/scitranslmed.aax9929 (2020).
8. Tanaka, S. et al. Influence of antibiotic exposure in the early postnatal period on the development of intestinal microbiota. *FEMS Immunol. Med. Microbiol.* **56**, 80–87 (2009).
9. Fallani, M. et al. the INFABIO team, Determinants of the human infant intestinal microbiota after the introduction of first complementary foods in infant samples from five European centres. *Microbiology* **157**, 1385–1392 (2011).
10. Fan, W., Huo, G., Li, X., Yang, L. & Duan, C. Impact of diet in shaping gut microbiota revealed by a comparative study in infants during the six months of life. *J. Microbiol. Biotechnol.* **24**, 133–143 (2014).
11. Ayeni, F. A. et al. Infant and Adult Gut Microbiome and Metabolome in Rural Bassa and Urban Settlers from Nigeria. *Cell Rep.* **23**, 3056–3067 (2018).
12. Dhakal, S. et al. Amish (Rural) vs. non-Amish (Urban) Infant Fecal Microbiotas Are Highly Diverse and Their Transplantation Lead to Differences in Mucosal Immune Maturation in a Humanized Germfree Piglet Model. *Front. Immunol*. **10**, https://doi.org/10.3389/fimmu.2019.01509 (2019).
13. Lehtimäki, J. et al. Urbanized microbiota in infants, immune constitution, and later risk of atopic diseases. *J. Allergy Clin. Immunol.* **148**, 234–243 (2021).
14. Laursen, M. F. et al. Having older siblings is associated with gut microbiota development during early childhood. *BMC Microbiol* **15**, 154 (2015).
15. Christensen, E. D. et al. The developing airway and gut microbiota in early life is influenced by age of older siblings. *Microbiome* **10**, 106 (2022).
16. Stokholm, J. et al. Maturation of the gut microbiome and risk of asthma in childhood. *Nat. Commun.* **9**, 141 (2018).
17. Depner, M. et al. Maturation of the gut microbiome during the first year of life contributes to the protective farm effect on childhood asthma. *Nat. Med.* **1**, 10 (2020).
18. Björkstén, B., Sepp, E., Julge, K., Voor, T. & Mikelsaar, M. Allergy development and the intestinal microflora during the first year of life. *J. Allergy Clin. Immunol.* **108**, 516–520 (2001).
19. Bisgaard, H. et al. Reduced diversity of the intestinal microbiota during infancy is associated with increased risk of allergic disease at school age. *J. Allergy Clin. Immunol.* **128**, 646-652.e1–5 (2011).
20. Hoskinson, C. et al. Delayed gut microbiota maturation in the first year of life is a hallmark of pediatric allergic disease. *Nat. Commun.* **14**, 4785 (2023).
21. Quince, C. et al. Extensive Modulation of the Fecal Metagenome in Children With Crohn's Disease During Exclusive Enteral Nutrition. *J. Am. Coll. Gastroenterol. ACG* **110**, 1718 (2015).
22. Norman, J. M. et al. Disease-Specific Alterations in the Enteric Virome in Inflammatory Bowel Disease. *Cell* **160**, 447–460 (2015).
23. Gogokhia, L. et al. Expansion of Bacteriophages Is Linked to Aggravated Intestinal Inflammation and Colitis. *Cell Host Microbe* **25**, 285–299.e8 (2019).
24. Zuo, T. et al. Temporal landscape of human gut RNA and DNA virome in SARS-CoV-2 infection and severity. *Microbiome* **9**, 91 (2021).
25. Yang, K. et al. Alterations in the Gut Virome in Obesity and Type 2 Diabetes Mellitus. *Gastroenterology* **161**, 1257–1269.e13 (2021).
26. Zhao, G. et al. Intestinal virome changes precede autoimmunity in type I diabetes-susceptible children. *Proc. Natl. Acad. Sci*. **114**, https://doi.org/10.1073/pnas.1706359114 (2017).
27. Shkoporov, A. N. & Hill, C. Bacteriophages of the Human Gut: The "Known Unknown" of the Microbiome. *Cell Host Microbe* **25**, 195–209 (2019).
28. Clooney, A. G. et al, Whole-Virome Analysis Sheds Light on Viral Dark Matter in Inflammatory Bowel Disease. *Cell Host Microbe*, https://doi.org/10.1016/j.chom.2019.10.009 (2019).
29. Mangalea, M. R. et al. Individuals at risk for rheumatoid arthritis harbor differential intestinal bacteriophage communities with distinct metabolic potential. *Cell Host Microbe* **29**, 726–739.e5 (2021).
30. Sausset, R., Petit, M. A., Gaboriau-Routhiau, V. & Paepe, M. D. New insights into intestinal phages. *Mucosal Immunol.* **13**, 205–215 (2020).
31. Reyes, A. et al. Viruses in the faecal microbiota of monozygotic twins and their mothers. *Nature* **466**, 334–338 (2010).
32. Shah, S. A. et al. Expanding known viral diversity in the healthy infant gut. *Nat. Microbiol.* **8**, 986–998 (2023).
33. Nanda, A. M. et al. Analysis of SOS-Induced Spontaneous Prophage Induction in Corynebacterium glutamicum at the Single-Cell Level. *J. Bacteriol.* **196**, 180–188 (2014).
34. Leal Rodríguez, C. et al. The infant gut virome is associated with preschool asthma risk independently of bacteria. *Nat. Med*. **30**, 138–148 (2023).
35. Kim, M.-S. & Bae, J.-W. Lysogeny is prevalent and widely distributed in the murine gut microbiota. *ISME J.* **12**, 1127–1141 (2018).
36. Juhala, R. J. et al. Genomic sequences of bacteriophages HK97 and HK022: pervasive genetic mosaicism in the lambdoid bacteriophages1. *J. Mol. Biol.* **299**, 27–51 (2000).
37. Wendling, C. C., Refardt, D. & Hall, A. R. Fitness benefits to bacteria of carrying prophages and prophage-encoded antibiotic-resistance genes peak in different environments. *Evol. Int. J. Org. Evol.* **75**, 515–528 (2021).
38. Kondo, K., Kawano, M. & Sugai, M. Distribution of Antimicrobial Resistance and Virulence Genes within the Prophage-Associated Regions in Nosocomial Pathogens. *mSphere* **6**, https://doi.org/10.1128/msphere.00452-21 (2021).
39. López-Leal, G., Santamaria, R. I., Cevallos, M. Á., Gonzalez, V. & Castillo-Ramírez, S. Letter to the Editor: Prophages Encode Antibiotic Resistance Genes in Acinetobacter baumannii. *Microb. Drug Resist* **26**, 1275–1277 (2020).
40. O'Brien, A. D. et al. Shiga-Like Toxin-Converting Phages from Escherichia coli Strains That Cause Hemorrhagic Colitis or Infantile Diarrhea. *Science* **226**, 694–696 (1984).

41. Waldor, M. K. & Mekalanos, J. J. Lysogenic Conversion by a Filamentous Phage Encoding Cholera Toxin. *Science* **272**, 1910–1914 (1996).
42. Bae, T., Baba, T., Hiramatsu, K. & Schneewind, O. Prophages of Staphylococcus aureus Newman and their contribution to virulence. *Mol. Microbiol.* **62**, 1035–1047 (2006).
43. Ptashne, M. et al. Autoregulation and function of a repressor in bacteriophage lambda. *Science* **194**, 156–161 (1976).
44. Cumby, N., Edwards, A. M., Davidson, A. R. & Maxwell, K. L. The bacteriophage HK97 gp15 moron element encodes a novel superinfection exclusion protein. *J. Bacteriol.* **194**, 5012–5019 (2012).
45. Susskind, M. M., Botstein, D. & Wright, A. Superinfection exclusion by P22 prophage in lysogens of Salmonella typhimurium. III. Failure of superinfecting phage DNA to enter sieA+ lysogens. *Virology* **62**, 350–366 (1974).
46. Newton, G. J. et al. Three-component-mediated serotype conversion in Pseudomonas aeruginosa by bacteriophage D3. *Mol. Microbiol.* **39**, 1237–1247 (2001).
47. McGrath, S., Fitzgerald, G. F. & van Sinderen, D. Identification and characterization of phage-resistance genes in temperate lactococcal bacteriophages. *Mol. Microbiol.* **43**, 509–520 (2002).
48. Shinedling, S., Parma, D. & Gold, L. Wild-type bacteriophage T4 is restricted by the lambda rex genes. *J. Virol.* **61**, 3790–3794 (1987).
49. Mahony, J., McGrath, S., Fitzgerald, G. F. & van Sinderen, D. Identification and Characterization of Lactococcal-Prophage-Carried Superinfection Exclusion Genes. *Appl. Environ. Microbiol.* **74**, 6206–6215 (2008).
50. Sun, X., Göhler, A., Heller, K. J. & Neve, H. The ltp gene of temperate Streptococcus thermophilus phage TP-J34 confers superinfection exclusion to Streptococcus thermophilus and Lactococcus lactis. *Virology* **350**, 146–157 (2006).
51. Paepe, M. D. et al. Carriage of λ Latent Virus Is Costly for Its Bacterial Host due to Frequent Reactivation in Monoxenic Mouse Intestine. *PLOS Genet* **12**, e1005861 (2016).
52. DeBardeleben, H. K., Lysenko, E. S., Dalia, A. B. & Weiser, J. N. Tolerance of a Phage Element by Streptococcus pneumoniae Leads to a Fitness Defect during Colonization. *J. Bacteriol.* **196**, 2670–2680 (2014).
53. Waller, A. S. et al. Classification and quantification of bacteriophage taxa in human gut metagenomes. *ISME J.* **8**, 1391–1402 (2014).
54. Lugli, G. A. et al. Prophages of the genus Bifidobacterium as modulating agents of the infant gut microbiota. *Environ. Microbiol.* **18**, 2196–2213 (2016).
55. Nayfach, S. et al. CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nat. Biotechnol.* **39**, 578–585 (2021).
56. 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.e9 (2021).
57. Lou, Y. C. et al. Infant gut DNA bacteriophage strain persistence during the first 3 years of life. *Cell Host Microbe* **32**, 35–47.e6 (2024).
58. Gulyaeva, A. et al. Discovery, diversity, and functional associations of crAss-like phages in human gut metagenomes from four Dutch cohorts. *Cell Rep.* **38**, 110204 (2022).
59. Silpe, J. E., Duddy, O. P. & Bassler, B. L. Induction mechanisms and strategies underlying interprophage competition during polylysogeny. *PLOS Pathog.* **19**, e1011363 (2023).
60. Li, X. et al. The infant gut resistome associates with E. coli, environmental exposures, gut microbiome maturity, and asthma-associated bacterial composition. *Cell Host Microbe* **29**, 975–987.e4 (2021).
61. Bisgaard, H. The Copenhagen Prospective Study on Asthma in Childhood (COPSAC): design, rationale, and baseline data from a longitudinal birth cohort study. *Ann. Allergy Asthma Immunol.* **93**, 381–389 (2004).
62. Tange, O. GNU Parallel 2018, https://doi.org/10.5281/zenodo.1146014 (2018).
63. Joshi, N. A. & Fass, J. Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33)[Software] (available at https://scholar.google.com/scholar?cluster=6706699853004034677&hl=en&oi=scholarr) (2011).
64. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. metaSPAdes: a new versatile metagenomic assembler. *Genome Res.* **27**, 824–834 (2017).
65. Rodriguez-R, L. M., Gunturu, S., Tiedje, J. M., Cole, J. R. & Konstantinidis, K. T. Nonpareil 3: Fast Estimation of Metagenomic Coverage and Sequence Diversity. *mSystems* **3**, https://doi.org/10.1128/msystems.00039-18 (2018).
66. Deng, L. et al. A Protocol for Extraction of Infective Viromes Suitable for Metagenomics Sequencing from Low Volume Fecal Samples. *Viruses* **11**, 667 (2019).
67. Conceição-Neto, N. et al. Modular approach to customise sample preparation procedures for viral metagenomics: a reproducible protocol for virome analysis. *Sci. Rep.* **5**, 16532 (2015).
68. Rognes, T., Flouri, T., Nichols, B., Quince, C. & Mahé, F. VSEARCH: a versatile open source tool for metagenomics. *PeerJ* **4**, e2584 (2016).
69. Ren, J. et al. Identifying viruses from metagenomic data using deep learning. *Quant. Biol. Beijing China* **8**, 64–77 (2020).
70. 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).
71. Cook, R. et al. INfrastructure for a PHAge REference Database: Identification of Large-Scale Biases in the Current Collection of Cultured Phage Genomes. *PHAGE N. Rochelle N.* **2**, 214–223 (2021).
72. Ondov, B. D. et al. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.* **17**, 132 (2016).
73. BBMapSourceForge. Available at https://sourceforge.net/projects/bbmap/ (2023).
74. Roux, S. et al. Minimum Information about an Uncultivated Virus Genome (MIUViG). *Nat. Biotechnol.* **37**, 29–37 (2019).
75. Jang, H. Bin et al. Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. *Nat. Biotechnol.* **37**, 632–639 (2019).
76. Freeman, T. C. et al. Graphia: A platform for the graph-based visualisation and analysis of high dimensional data. *PLOS Comput. Biol.* **18**, e1010310 (2022).
77. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).
78. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinforma. Oxf. Engl.* **22**, 1658–1659 (2006).
79. Huerta-Cepas, J. et al. Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. *Mol. Biol. Evol.* **34**, 2115–2122 (2017).
80. Cook, R. et al. Hybrid assembly of an agricultural slurry virome reveals a diverse and stable community with the potential to alter the metabolism and virulence of veterinary pathogens. *Microbiome* **9**, 65 (2021).
81. Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res* **39**, W29–W37 (2011).
82. Mathieu, A. et al. Virulent coliphages in 1-year-old children fecal samples are fewer, but more infectious than temperate coliphages. *Nat. Commun.* **11**, 378 (2020).
83. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
84. ClusterGenomes/Cluster_genomes_5.1.pl at master · simroux/ClusterGenomes*GitHub* (available at https://github.com/simroux/ClusterGenomes/blob/master/Cluster_genomes_5.1.pl).

85. Roux, S., Emerson, J. B., Eloe-Fadrosh, E. A. & Sullivan, M. B. Benchmarking viromics: an in silico evaluation of metagenome-enabled estimates of viral community composition and diversity. *PeerJ* **5**, e3817 (2017).
86. Guerin, E. et al. Biology and Taxonomy of crAss-like Bacteriophages, the Most Abundant Virus in the Human Gut. *Cell Host Microbe* **24**, 653–664.e6 (2018).
87. Dion, M. B. et al. Streamlining CRISPR spacer-based bacterial host predictions to decipher the viral dark matter. *Nucleic Acids Res* **49**, 3127–3138 (2021).
88. Kanehisa, M. & Sato, Y. KEGG Mapper for inferring cellular functions from protein sequences. *Protein Sci. Publ. Protein Soc.* **29**, 28–35 (2020).
89. Kumar, S., Stecher, G., Li, M., Knyaz, C. & Tamura, K. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Mol. Biol. Evol.* **35**, 1547–1549 (2018).
90. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* **49**, W293–W296 (2021).
91. Sharifi, F. & Ye, Y. MyDGR: a server for identification and characterization of diversity-generating retroelements. *Nucleic Acids Res.* **47**, W289–W294 (2019).
92. Yan, F. et al. Discovery and characterization of the evolution, variation and functions of diversity-generating retroelements using thousands of genomes and metagenomes. *BMC Genomics* **20**, 595 (2019).
93. Benler, S. et al. A diversity-generating retroelement encoded by a globally ubiquitous Bacteroides phage. *Microbiome* **6**, 191 (2018).
94. Cantu, V. A. et al. PhANNs, a fast and accurate tool and web server to classify phage structural proteins. *PLOS Comput. Biol.* **16**, e1007845 (2020).
95. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
96. Bellas, C. M., Schroeder, D. C., Edwards, A., Barker, G. & Anesio, A. M. Flexible genes establish widespread bacteriophage pan-genomes in cryoconite hole ecosystems. *Nat. Commun.* **11**, 4403 (2020).
97. Martinez-Hernandez, F. et al. Single-virus genomics reveals hidden cosmopolitan and abundant viruses. *Nat. Commun.* **8**, 15892 (2017).
98. Manrique, P. et al. Healthy human gut phageome. *Proc. Natl Acad. Sci.* **113**, 10400–10405 (2016).
99. 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.e8 (2020).
100. Thingstad, T. F. Elements of a theory for the mechanisms controlling abundance, diversity, and biogeochemical role of lytic bacterial viruses in aquatic systems. *Limnol. Oceanogr.* **45**, 1320–1328 (2000).
101. Knowles, B. et al. Lytic to temperate switching of viral communities. *Nature* **531**, 466–470 (2016).
102. Shkoporov, A. N. et al. ΦCrAss001 represents the most abundant bacteriophage family in the human gut and infects Bacteroides intestinalis. *Nat. Commun.* **9**, 4781 (2018).
103. Yutin, N. et al. Discovery of an expansive bacteriophage family that includes the most abundant viruses from the human gut. *Nat. Microbiol.* **3**, 38 (2018).
104. Reeves, P. Evolution of Salmonella O antigen variation by interspecific gene transfer on a large scale. *Trends Genet. TIG* **9**, 17–22 (1993).
105. Rajakumar, K. et al. Nucleotide sequence of the rhamnose biosynthetic operon of Shigella flexneri 2a and role of lipopolysaccharide in virulence. *J. Bacteriol.* **176**, 2362–2373 (1994).
106. Tsukioka, Y., Yamashita, Y., Oho, T., Nakano, Y. & Koga, T. Biological function of the dTDP-rhamnose synthesis pathway in Streptococcus mutans. *J. Bacteriol.* **179**, 1126–1134 (1997).
107. Porter, N. T. et al. Phase-variable capsular polysaccharides and lipoproteins modify bacteriophage susceptibility in Bacteroides thetaiotaomicron. *Nat. Microbiol.* **5**, 1170–1181 (2020).
108. Krinos, C. M. et al. Extensive surface diversity of a commensal microorganism by multiple DNA inversions. *Nature* **414**, 555–558 (2001).
109. 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).
110. Reeves, P. Role of O-antigen variation in the immune response. *Trends Microbiol.* **3**, 381–386 (1995).
111. Rahman, M. M., Guard-Petter, J. & Carlson, R. W. A virulent isolate of Salmonella enteritidis produces a Salmonella typhi-like lipopolysaccharide. *J. Bacteriol.* **179**, 2126–2131 (1997).
112. Vaca-Pacheco, S., Paniagua-Contreras, G. L., García-González, O. & de la Garza, M. The clinically isolated FIZ15 bacteriophage causes lysogenic conversion in Pseudomonas aeruginosa PAO1. *Curr. Microbiol.* **38**, 239–243 (1999).
113. Ramotar, K., Conly, J. M., Chubb, H. & Louie, T. J. Production of menaquinones by intestinal anaerobes. *J. Infect. Dis.* **150**, 213–218 (1984).
114. Takahashi, M., Naitou, K., Ohishi, T., Kushida, K. & Miura, M. Effect of vitamin K and/or D on undercarboxylated and intact osteocalcin in osteoporotic patients with vertebral or hip fractures. *Clin. Endocrinol.* **54**, 219–224 (2001).
115. Manfioletti, G., Brancolini, C., Avanzi, G. & Schneider, C. The protein encoded by a growth arrest-specific gene (gas6) is a new member of the vitamin K-dependent proteins related to protein S, a negative coregulator in the blood coagulation cascade. *Mol. Cell. Biol.* **13**, 4976–4985 (1993).
116. Crivello, N. A., Casseus, S. L., Peterson, J. W., Smith, D. E. & Booth, S. L. Age- and brain-region-specific effects of dietary vitamin K on myelin sulfatides. *J. Nutr. Biochem.* **21**, 1083–1088 (2010).
117. Ferland, G. Vitamin K and the nervous system: an overview of its actions. *Adv. Nutr. Bethesda Md.* **3**, 204–212 (2012).
118. Lippi, G. & Franchini, M. Vitamin K in neonates: facts and myths. *Blood Transfus.* **9**, 4–9 (2011).
119. Suttie, J. W. The importance of menaquinones in human nutrition. *Annu. Rev. Nutr.* **15**, 399–417 (1995).
120. Conly, J. M., Stein, K., Worobetz, L. & Rutledge-Harding, S. The contribution of vitamin K2 (menaquinones) produced by the intestinal microflora to human nutritional requirements for vitamin K. *Am. J. Gastroenterol.* **89**, 915–923 (1994).
121. Shearer, M. J. Vitamin K and Vitamin K-Dependent Proteins. *Br. J. Haematol.* **75**, 156–162 (1990).
122. Lamy-Besnier, Q. et al. Chromosome folding and prophage activation reveal specific genomic architecture for intestinal bacteria. *Microbiome* **11**, 111 (2023).
123. Otsuji, N., Sekiguchi, M., Iijima, T. & Takagi, Y. Induction of Phage Formation in the Lysogenic Escherichia coli K-12 by Mitomycin C. *Nature* **184**, 1079–1080 (1959).
124. George, J., Castellazzi, M. & Buttin, G. Prophage induction and cell division in E. coli. III. Mutations sfiA and sfiB restore division in tif and lon strains and permit the expression of mutator properties of tif. *Mol. Gen. Genet. MGG* **140**, 309–332 (1975).
125. López, E. et al. Induction of Prophages by Fluoroquinolones in Streptococcus pneumoniae: Implications for Emergence of Resistance in Genetically-Related Clones. *PLOS ONE* **9**, e94358 (2014).
126. Sutcliffe, S. G., Shamash, M., Hynes, A. P. & Maurice, C. F. Common Oral Medications Lead to Prophage Induction in Bacterial Isolates from the Human Gut. *Viruses* **13**, 455 (2021).

Acknowledgements

We would like to acknowledge and thank the families of the $COPSAC_{2010}$ cohort, without whose continued support this work would not be possible. S.A.S. is a recipient of a Novo Nordisk Foundation project grant in basic bioscience, NNF18OC0052965, which also supported the salary of T.R. J.T

is supported by the BRIDGE – Translational Excellence Programme (bridge.ku.dk) at the Faculty of Health and Medical Sciences, University of Copenhagen, funded by the Novo Nordisk Foundation (Grant agreement no. NNF18SA0034956).

Author contributions

S.A.S. obtained funding for the work. H.B., J.S., B.C., and K.B. contributed to the design of the cohort and collection of samples. L.D., D.S.N., G.V., J.R., S.A.S., J.S., M.A.P. and S.J.S. generated the data for analysis. T.R. and J.T. conceived the idea, analysed the data, and drafted the manuscript. All authors contributed to interpretation of the data. All authors approved and have contributed to the final manuscript.

Competing interests

The authors declare no competing interests.

Ethics approval

The study was approved by the Capital Region of Denmark's Local Ethics Committee (H-B-2008-093) and the Danish Data Protection Agency (2015-41-3696). The study was conducted following the guiding principles of the Declaration of Helsinki. Before enrolment, parents gave their oral and written informed consent.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41522-025-00674-1.

Correspondence and requests for materials should be addressed to Shiraz A. Shah.