HHS Public Access

Author manuscript

Nature. Author manuscript; available in PMC 2020 October 15.

Published in final edited form as:

Nature. 2020 May ; 581(7809): 470–474. doi:10.1038/s41586-020-2192-1.

Step-wise assembly of the neonatal virome modulated by breastfeeding

Guanxiang Liang[1,2], **Chunyu Zhao**[2], **Huanjia Zhang**[2], **Lisa Mattei**[2], **Scott Sherrill-Mix**[1], **Kyle Bittinger**[2], **Lyanna R. Kessler**[1], **Gary D. Wu**[3], **Robert N. Baldassano**[2], **Patricia DeRusso**[2], **Eileen Ford**[2], **Michal A. Elovitz**[4], **Matthew S. Kelly**[5], **Mohamed Z. Patel**[6], **Tiny Mazhani**[6], **Jeffrey S. Gerber**[7], **Andrea Kelly**[8], **Babette S. Zemel**[2], **Frederic D. Bushman**[1,*]

[1]Department of Microbiology, Perelman School of Medicine, University of Pennsylvania, 3610 Hamilton Walk, Philadelphia, PA 19104-6076 USA

[2]Division of Gastroenterology, Hepatology, and Nutrition, Children's Hospital of Philadelphia, Philadelphia, PA 19104-4319 USA

[3]Division of Gastroenterology, Perelman School of Medicine, University of Pennsylvania, 421 Curie Boulevard, Philadelphia, PA 19104-6076 USA

[4]Maternal and Child Health Research Center, Department of Obstetrics and Gynecology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, PA 19104-6076 USA

[5]Division of Pediatric Infectious Diseases, Duke University, Durham, NC 27710 USA

[6]Department of Paediatric and Adolescent Health, Faculty of Medicine of the University of Botswana, Gaborone, Botswana

[7]Division of Infectious Diseases, Children's Hospital of Philadelphia, Philadelphia, PA 19104-4319 USA

[8]Division of Endocrinology and Diabetes, Children's Hospital of Philadelphia, Philadelphia, PA 19104-4319 USA

Abstract

*Corresponderce and requests for materials should be addressed to FDB. Bushman@pennmedicine.upenn.edu.

Author contributions: GL carried out biochemical analysis, sequencing, and bioinformatic analysis; LRK, HZ and LM assisted with biochemical manipulations; CZ, SS and KB assisted with bioinformatic analysis; GCW, RNB, PD, EF, ME, JG, AK, BSZ, MSK, MZP, TM carried out sample collection; GL and FDB conceived the project and wrote the paper.

Competing interests: The authors declare no competing interests.

Extended data is available for the paper at www.nature.com/nature.

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

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

Data and software availability

Sample information and raw sequences are available in the National Center for Biotechnology Information Sequence Read Archive under BioProject ID PRJNA524703 (Supplementary Table 8). The isolated bacterial genome sequences have been deposited at DDBJ/ENA/GenBank under the accession WVTF00000000-WVUC00000000 (Supplementary Table 3). All bioinformatic scripts are available on Github (https://github.com/guanxiangliang/liang2019).

The gut of healthy human neonates is usually devoid of viruses at birth, but quickly becomes colonized, in some cases leading to gastrointestinal disorders[1–4]. Here we report that viral community assembly in neonates takes place in distinct steps. Fluorescent staining of virus-like particles purified from infant meconium/early stool samples show few or no particles, but by one month of life particle numbers achieve 10^9 per gram, and these numbers appear to persist through life[5–7]. We investigated the origin of these viral populations using shotgun metagenomic sequencing of viral-enriched preparations and whole microbial communities, and followed up with targeted microbiological analyses. Results indicate that, early after birth, pioneer bacteria colonize the infant gut, and by one month prophage induced from these bacteria provide the predominant population of virus-like particles. By four months of life, identifiable viruses that replicate in human cells become more prominent. Multiple human viruses were more abundant in stool samples from babies exclusively fed formula versus those fed partially or fully on breast milk, paralleling reports that breast milk can be protective against viral infections[8–10]. Phage populations also differed associated with breastfeeding. Evidently colonization of the infant gut is stepwise, first mainly by temperate bacteriophages induced from pioneer bacteria, and later by viruses that replicate in human cells, with the second phase modulated by breastfeeding.

To investigate early life viral colonization, we first analyzed stool samples from 20 healthy infants longitudinally (Supplementary Table 1). Samples included meconium/early stool samples collected at day 0 to 4 days after birth, (median of 17+/− hrs after birth, range 11 to 152 hrs; henceforth “month 0”) and also stool samples collected at one and four months of life. Infants were from an urban United States cohort of African American infants. As an initial step, virus-like particles (VLPs) were purified from meconium or stool and stained with SYBR gold, which binds nucleic acids, and subsequently visualized by epifluorescence microscopy (Fig. 1a). VLPs were undetectable in the majority of meconium samples, with only three of 20 samples showing detectable counts (Fig. 1b). By month one, most samples were positive, and VLP counts averaged 1.6×10^9 per gram of stool. Values at month 4 were similar. We also tested the VLP counts from twelve 2–5-year-old children, which averged 9.4×10^8 per gram of stool, which were not distinguishable from month 1 and 4 samples (P = 0.48, Wilcoxon rank-sum test). This number is also close to that reported for adults[5–7]. Thus we conclude that the high VLP counts seen in one month old infants typically persist into adulthood.

To characterize bacterial content, DNA was purified from whole meconium/stool and analyzed by qPCR to quantify bacterial 16S rRNA gene copy number (Fig. 1c). Some published studies have suggested microbial colonization of the infant begins *in utero* (e. g. [11]), but recent studies indicate that colonization more likely begins with rupture of membranes and delivery[12–14]. Quantification showed low or undetectable levels of bacterial 16S rRNA genes for 14 of the 20 meconium/early stool samples from month 0, and relatively low levels for the other six (median 3.3×10^6). In contrast, for months one and four, most samples were positive for 16S rRNA gene sequences, with a median value of 3.1×10^8. To query the early life microbiome further, total DNA from each sample was subjected to metagenomic sequencing. For the month 0 samples, many were dominated by human DNA, which is characteristic of the neonatal gut prior to colonization (Extended Data Fig. 1a). Some month 0 samples also contained bacterial DNA, indicating early colonization and/or

reagent contamination. Levels of human DNA decreased with time after delivery ($P = 0.04$, Spearman's rank-order correlation $\rho = -0.45$; Extended Data Fig. 1b), consistent with bacterial colonization. For month one and month four samples, bacterial DNA predominated (Extended Data Fig. 1a). The bacterial richness and diversity at month 0 was lower compared to month 1 and month 4 samples. Early bacterial colonizers included Proteobacteria, Actinobacteria, Bacteroidetes and Firmicutes (Extended Data Fig. 1c–e), consistent with previous studies[2,15].

To investigate the origin of the early life virome, DNA and RNA were purified from preparations of VLPs from stool from each of the three time points and characterized by metagenomic sequencing. After filtering out human DNA, we assembled sequence reads into contigs and annotated open reading frames. Previous literature indicates that many gut viruses are uncharacterized bacteriophage (henceforth "phage")[1,16–19], which are challenging to identify in metagenomic sequence data because the proportion of phage genomes in databases is small compared to the number of global phage types. Viruses that infect human cells are more fully characterized and thus more readily recognizable. To address this challenge, we required half of all reading frames within a contig to be annotated as viral to assign that contig as viral. Quantification of viral species richness showed low values at month 0, but higher richness after one and four months (Fig. 1d). After taxonomic assignment of viral contigs, we found that, despite the difficulty of identifying phages, the great majority of VLP classifications were in fact phage families (Fig. 1e). Most were from DNA phage, consistent the reported rarity of RNA phage[20]. For DNA phage, an average of 31% of reads could be assigned as viral at month 1, and 38% at month 4. The nature of the remainder is unknown but some likely represent unstudied phages. Values were lower at month 0 (11%), likely reflecting a relative increase of background due to the low numbers of particles recovered (Extended Data Fig. 2a–h).

To assess community interactions, we compared bacterial abundance from 16S rRNA gene qPCR data, bacterial richness and bacterial diversity from sequence data, against VLP counts, viral richness and viral diversity, revealing strong positive correlations (Extended Data Fig. 3a, b; Supplementary Table 2).

For the minority of viruses detected that are known to replicate in human cells (Fig. 1e), at month 0 a single sample was positive for Herpesviridae, and another for Picornaviridae. By month four, human cell viruses were more prominent, including Adenoviridae, Anelloviridae, Caliciviridae, and Picornaviridae.

Either of two modes of phage production could generate the observed VLP populations[21,22]. Lytic phages only grow by infection, replication and lysis (Extended Data Fig. 4a). Previous reports, focusing on older infants and adults, suggested that lytic growth and predator-prey interactions between phage and bacteria were prominent in early life communities[2]. Temperate phages have a second strategy available, involving integration of the phage DNA into the host bacterial DNA, followed by quiescent growth as an integrated prophage (Extended Data Fig. 4a). Exposure to an inducing signal causes integrated prophage to excise and resume lytic growth. Induction can also take place at low levels spontaneously[23].

The prophage state is commonly maintained by repressor proteins, which also serve to exclude infection by similar or identical (homoimmune) phage strains[21,22].

To test the idea that the early life virome is composed of strictly lytic phage, we purified 24 bacterial strains from infant gut, including three *Escherichia coli*, three *Klebsiella*, and ten *Enterococcus* strains (Supplementary Table 3), and plated virome fractions from the cognate infant back on these bacteria. In no case did virome VLPs form plaques on lawns of these bacteria, thus providing no evidence for lytic replication. Note that temperate phages are not expected to form plaques on host cells already harboring that phage as an integrated prophage due to repressor-mediated homoimmune exclusion[21,22]. Analysis of our DNA VLP contigs using PHACTS, a Random Forest-based approach to classifying phage lifestyles[24], indicated that the great majority of genomes more closely resembled temperate phage than lytic phage (Extended Data Fig. 4b).

We next investigated the idea that virome populations resulted from prophage induction by quantifying VLP production from the 24 infant bacterial strains described above. Bacterial strains were analyzed for spontaneous VLP production during growth in liquid culture, and for production after induction with the DNA damaging agent mitomycin C, using the fluorescent staining assay for VLP particles. Experiments were carried out under aerobic and anaerobic conditions. We detected spontaneous VLP production from 32% of strains. After induction with mitomycin C, 80% of strains produced VLPs (Fig. 2a). Sixteen strains showed VLP production of at least 10^7 particles per ml under at least one condition. Thus we conclude that infant gut bacteria are commonly capable of high-level VLP production following prophage induction.

The hypothesis that prophage induction yields the bulk of VLPs in infant gut samples predicts that VLP sequences found in stool should be detectable as integrated prophages in bacterial genome sequences. We thus sequenced genomes of the 24 infant bacterial strains described above, and also VLPs produced from those strains in the presence or absence of mitomycin C (Supplementary Table 3). Prophage sequences could be readily detected in the bacterial genomes, and many of these were detectable both in sequences from the induced VLP samples and also in the VLP samples from infant stool. Examples are shown in Fig. 2 b and c, where VLP sequence reads are shown aligned to bacterial contigs, so that spikes indicate VLP detection after mitomycin C induction (red), in the absence of induction (blue) and in purified stool VLPs (green). To test the infant specificity of each community, we mapped the stool VLP reads back to the viral contigs assembled from VLPs induced from the 24 bacterial strains. Although the steps of VLP nucleic acid amplification prior to sequencing can distort abundances, we nevertheless found that the induced VLPs from each purified bacterial strain were more similar to stool VLP sequences from the infant from which the bacterial strain was isolated than to VLPs from unmatched infants ($P < 0.0001$; Extended Data Fig. 5a), consistent with production of infant gut phage by prophage induction.

In addition, there was a significant positive correlation between the proportion of each bacteria in the infant gut community and the proportion of that bacteria's prophages in that infant's gut virome ($P = 0.0008$; Spearman's rank-order correlation $\rho = 0.53$;; Fig. 2d;

Extended Data Fig. 5b). The abundance of the bacterial strains in the gut communities ranged from 0.03% to 99.1%, indicating that in at least some cases a large proportion of the gut community was interrogated.

The abundant crAssphages, which infect Bacteriodetes and do not integrate during replication[25,26], were scarce in samples from month one, but more common by month four and in samples from 2–5-year-old children (Extended Data Fig. 6). Evidently this group of lytic phages colonize children predominantly later in life, potentially reflecting sequential acquisition of Bacteriodes strains and later crAssphages from the environment.

These findings support the idea that prophage induction from pioneer bacteria is the main source of the observed virome community by month one. This is supported by the findings that: 1) replication of lytic phage was undetectable; 2) many purified bacterial strains from infants produced VLPs at high levels; 3) sequences of genomes from these induced VLPs could be identified as integrated prophage in bacteria isolated from these infants; 4) stool VLP genome sequences could be identified as integrated prophage in the bacterial genomes; 5) VLP abundance in stool was proportional to the abundance of the host bacteria in the same sample; and 6) VLP contigs annotated primarily as lysogenic phage and not lytic phage.

We then compared features of the VLP data from infant stool samples to metadata on feeding history, mode of delivery, sex and other variables (Supplementary Table 1). Unexpectedly, this revealed a strong influence of breastfeeding, which was associated with lower accumulation of viruses that replicate in human cells. Taking a conservative threshold for detection, requiring coverage of 33% of the viral genome, viruses infecting human cells were only found in those infants fed formula exclusively (Fig. 3a and Extended Data Fig. 7a). Statistically this only achieved a p value of 0.11 (Fisher's exact test) due to the small sample size and unbalanced distribution (Fig. 3a, Extended Data Fig. 7b). Delivery type (spontaneous vaginal delivery versus cesarean section) did achieve statistical significance (P = 0.01, Fisher's exact test; Extended Data Fig. 7f, g).

To challenge these conclusions, we analyzed a validation cohort of an additional 125 infants, focusing on stool samples taken at 3–4 months of life. These samples were obtained from mixed race cohorts of urban U. S. infants (Supplementary Table 1). In these samples, delivery mode did not show a significant influence (Extended Data Fig. 7h–j), but a protective effect of breastfeeding was evident— 30% of formula fed babies were positive for viruses infecting human cells, while 9% of babies fed breast milk or breast milk plus formula were positive (P = 0.003, Fisher's exact test; Fig. 3a). Results based on requiring from 0.1% coverage up to 60% coverage of viral genomes for scoring detection yielded similarly significant results (Extended Data Fig. 7c, d). A comparison after normalizing for sequencing depth also yielded a significant difference (P < 0.0001, Wilcoxon rank-sum test; Extended Data Fig. 7e). As a control, preparations of formula were subjected to VLP purification and sequence analysis, which yielded no detections of animal cell viruses (data not shown), indicating that these viruses were unlikely to originate as contamination in the formula products themselves.

To validate the metagenomic detections, VLP DNA and RNA samples were also tested by qPCR for their content of Adenovirus, Torque teno virus, Enterovirus, Astrovirus, Sappovirus and Norovirus sequences (detailed in Supplementary Table 4). The qPCR analysis also showed enrichment of viruses infecting human cells in the exclusively formula fed cohort. ($P = 0.0002$, Fisher's exact test; Fig. 3b).

Both populations studied above were from urban cohorts in the United States. To begin to investigate whether our results hold more widely, we analyzed samples from a cohort of African newborns from Botswana. Infants 4 months of age were sampled using rectal swabs, so only qPCR assays and not sequence-based assays were attempted. All infants were delivered vaginally. We again found an association between exclusive formula feeding and colonization of viruses that grow on human cells ($P = 0.011$, Fisher's exact test; Fig. 3b).

Feeding type and other variables were then tested for effects on phage populations. Phage genes were annotated on assembled contigs, and their abundances used to calculate Bray-Curtis distances between communities. Significant differences in phage population structure could be detected based on feeding mode ($P = 0.001$, PERMANOVA; Extended Data Fig. 8a, c) but not for another twelve metadata variables (Extended Data Fig. 8a–e). To probe the taxa involved, shotgun sequence analysis of whole stool from pooled discovery and validation cohorts was queried and found to show higher abundance of *Bifidobacterium* (FDR = 0.02; Fig. 3c) and *Lactobacillus* (FDR = 0.03; Fig. 3c) in breastfed infants. Paralleling host abundances, VLP sequences aligning to temperate phages of *Bifidobacterium* and *Lactobacillus* were also enriched in breastfed infants ($P = 0.03$ and $P < 0.0001$, Fisher's exact test; Fig. 3d), in part explaining the effects of feeding mode on phage populations.

Our data thus indicate that viral colonization in early life is stepwise, with the first phase characterized by induction of prophage from pioneer bacteria, and a subsequent phase involving colonization with viruses infecting human cells, the latter of which is modulated by breastfeeding (Fig. 3e). Previous epidemiological studies have emphasized the protective effects of breastfeeding in reducing viral gastroenteritis and infant death[9,10]. Mixed feeding of formula and breast milk is also reported to be protective compared to formula only[8], as was seen in the metagenomic analysis reported here. Activities in breast milk that are known to inhibit viral colonization include maternal antibodies, human milk oligosaccharides, lactoferrin, and additional breast milk proteins (reviewed in[27–29]). The work reported here further develops our understanding of protection by breastfeeding in several respects. The metagenomic data 1) documents the extent of subclinical infections with potentially pathogenic viruses, 2) highlights the potency of viral inhibition by breastfeeding, and 3) specifies the diversity of viruses affected, including viral families that cannot be grown in the laboratory and for which inhibition by breastfeeding is unstudied. For the African cohort, we found viruses that grow on human cells more commonly in exclusively formula fed babies, but we also found more colonization in both feeding groups compared to US babies, emphasizing potential opportunitities to intervene to reduce viral transmission to infants. Going forward, the metagenomic methods described here should be useful in assessing the effects of different feeding strategies in diverse global settings to optimize protection of infants from gut viral infection.

Methods

Experimental model and human subjects

Three cohorts of newborn infants were studied. Detailed subject information is in Supplementary Table 1. All experimentation complied with ethical regulations, and written informed consent was obtained from all human subjects.

The Infant Growth and Microbiome Study (IGram) was approved by the Committee for the Protection of Human Subjects at The Children's Hospital of Philadelphia (IRB14–010833). African-American women planning to deliver at the Hospital of the University of Pennsylvania and their infants were enrolled. Inclusion and exclusion criteria are listed in Supplementary Table 5. Study visits were conducted at The Children's Hospital of Philadelphia. A total of 20 healthy, term infants were recruited for the discovery cohort. Stool samples were collected longitudinally at day 0 to 4 days after birth (meconium samples, Month 0), month 1 (Month 1), and month 4 (Month 4). The participants in an independent validation cohort had the same inclusion and exclusion criteria as the discovery cohort (only at month 4, n = 86). Fresh stool specimens from the healthy infants were collected from diapers and aliquoted into feces collection tubes (Sarstedt, Nümbrecht, Germany). All samples were stored at −80°C. Metadata regarding delivery mode, infant feeding and health outcomes was collected by medical chart review and in-person interview by trained research personnel.

The Microbiome, Antibiotic, and Growth Infant Cohort (MAGIC) Study was approved by the Committee for the Protection of Human Subjects at Children's Hospital of Philadelphia (IRB 15–012623). The study enrolled children born at Pennsylvania Hospital, Philadelphia, PA, receiving preventive health care in the CHOP Primary Care Network or participating in private practices, together with their biological mothers. The distribution of race, ethnicity, and sex of the newborns reflected the general distribution in the participating sites. All subjects enrolled were less than 120 hours of age, greater than 36 weeks gestation, greater than 2000 grams, and spent less than 120 hours in the neonatal care unit. Mothers were over the age of 18 and spoke English. A total of 39 heathy, term babies were used for this cohort. Study visits were conducted at Children's Hospital of Philadelphia. Stool samples were collected and questionnaires administered at birth and every 3 months until the subject reached 24 months of age. Stool samples obtained at 3 months of life were used for this cohort. Fresh stool specimens were collected at home by parents using a sterile fecal collection tube to scoop a pea sized amount from a used diaper. Samples were then transported by courier on dry ice, aliquoted, and stored at −80°C. Mother and baby clinical and metadata were collected via medical chart review and parent questionnaires.

The Botswana Infant Microbiome Study was approved by the Botswana Ministry of Health (IRB HPDME 13/8/1) and Institutional Review Boards at the University of Pennsylvania (IRB 822692) and Duke University (IRB 319561). Mother-infant pairs (n = 300) were enrolled within 48 hours of delivery at Princess Marina Hospital and two public clinics in or near Gaborone, Botswana. Exclusion criteria included maternal age less than 18 years, infant birth weight less than 2000 grams, multiple gestation pregnancy, and Caesarian delivery. Participants were seen for monthly study visits until the infant was 6 months of age and

every other month thereafter until the infant was 12 months of age. At all visits, a questionnaire was administered and clinical samples were obtained from the infant and the mother. Rectal swab samples obtained at 4 months of age from 100 infants were used for this cohort. These samples were collected into eNAT® medium (Copan Italia, Brescia, Italy), stored on ice, and transported within 4 hours to the National Health Laboratory in Gaborone for processing and storage at −80°C. Metadata, including data regarding infant feeding practices, were collected by medical chart review and in-person interview by trained research personnel.

In no cases were infants from any cohort reported to be suffering from gastroenteritis at the time of sampling.

VLP purification from stool samples

VLPs were purified as described[30]. Approximately 200 mg of stool was homogenized in 30 mL of SM buffer (50 mM Tris-HCl pH 7.5, 100 mM NaCl, 8 mM $MgSO_4$), spun down and filtered through a 0.2-μm-pore-size filter (Thermo Fisher Scientific, Waltham, MA, USA). The filtrate was concentrated using a 100-kDa-molecular-mass Amicon Ultra-15 Centrifugal filter (Millipore, Burlington, MA, United States), resuspended in 30 ml SM buffer, and concentrated for the second time to a final volume of ~500 μl. The concentrate was treated with DNase I and RNase (Roche, Basel, Switzerland) at 37°C for 30 min to degrade nonencapsulated nucleic acids. A total of 200 μl VLP preparation was used for viral nucleic acid extraction immediately after DNase I and RNase treatment; the remainder was stored at 4°C up to 3 months. In order to detect enveloped viruses, no chloroform was used to treat the VLPs. To test the VLPs purification efficiency, 16S qPCR was used to quantify the 16S copy number before (total microbial DNA) and after VLPs purification (VLP viral DNA). Samples showed an average reduction of 99.9% after purification (Extended Data Fig. 9).

Control spiking experiments with bacteriophage lambda showed that after addition to stool, ~90% of plaque forming units could be recovered using the above methods.

VLP enumeration

Thirty-five μl of purified VLPs were diluted in 10 ml SM buffer and filtered onto a 0.02-μm Anodisc polycarbonate filter (Whatman, Maidstone, UK). Filters were stained with 2×SYBR Gold (Thermo Fisher Scientific) for 15 min, then washed with H_2O once. After drying, the filter was mounted on a glass slide with 15 μl of mountant buffer (100ul 10% ascorbic acid + 4.9ml 7.4 PBS + 5ml 100% glycerol; filtered through 0.02um). For each filter, viruses were counted in 5 to 10 fields of view selected randomly. The filter was visualized using a motorized inverted system microscope IX81 (Shinjuku, Tokyo, Japan) for fluorescence. VLPs were counted using imageJ. Stained particles <0.5 μm in diameter were regarded as VLPs (larger particles were not counted). Purified lambda phage with known plaque-forming unit counts (PFU) per ml were used as a positive control to adjust image color, saturation, level and contrast. VLPs mock purified from SM buffer was used as negative controls. At least one count per microscope field was set as a threshold for a positive detection, which was equal to ~6.6×10^6 counts per gram feces. Lower than this threshold,

the VLP counts were considered to be below the level of detection. The results were listed in Supplementary Table 2.

Viral nucleic acid extraction and amplification

Viral DNA and RNA were extracted from VLPs using the AllPrep DNA/RNA Mini kit (Qiagen, Hilden, Germany) according to the manufacturer's instructions. DNA was stored at −20°C and RNA at −80°C. Viral DNA was subjected to DNA whole genome amplification using the GenomiPhi V2 Amplification kit (GE Healthcare, Little Chalfont, UK). Viral RNA was treated with DNase, reverse transcribed and PCR-amplified as described[31]. Specifically, 20 μl of RNA was treated with 10 units of RNAse-free, recombinant DNAse (Roche) for 20 min at 37°C. A total of 5 μL of each sample was then reverse transcribed. First strand cDNA synthesis was completed using SuperScript III First Strand Synthesis kit (Thermo Fisher Scientific) and Primer A (5'-GTTTCCCAGTCACGATCNNNNNNNNN-3'), to allow for random priming[31]. The second strand synthesis was performed using DNA Polymerase I, Large (Klenow) Fragment (New England BioLab, Ipswich, MA, USA). The dsDNA product was then amplified by adding Primer B (5'-GTTTCCCAGTCACGATC-3') with AccuPrime Taq High Fidelity DNA polymerase (Thermo Fisher Scientific) with the following reaction conditions: 75.5 μl of molecular grade H_2O, 10 μl of 10xPCR Buffer I, 4 μl of 50 mM $MgCl_2$, 2.5 μl 10 mM dNTPs, 1 μl 100 μM Primer B, 1μl Taq and 6μl dsDNA product. Products were amplified at 94°C for 2 min, 94°C 30 s, 40°C 30 s, 50°C 30 s, 72°C 1 min for 40 cycles. Amplified DNA and cDNA were stored at −20°C.

For the African cohort, samples were only available as rectal swabs stored in eNAT® medium (Copan Italia, Brescia, Italy), which contains guanidine thyocianate. Thus it was not possible to purify VLPs prior to analysis. Based on experience, such samples have such high human DNA content that shotgun metagenomic analysis yields overwhelmingly human sequences—we thus carried out only qPCR analysis of these samples. Nucleic acids were purified using the AllPrep DNA/RNA Mini kit (Qiagen, Hilden, Germany) as mentioned above. No pre-amplification was performed for either DNA or RNA.

Total microbial DNA extraction

Approximately 200 mg of stool were used for total microbial DNA extraction. Total microbial DNA was purified from each sample using the Mo Bio PowerSoil kit (Mo Bio, Carlsbad, CA, USA) following the manufacturer's instructions. A total of 50 μl total microbial DNA was obtained for each sample, and stored at −20°C.

Stool virome library and total microbial shotgun library construction and sequencing

Amplified viral DNA, cDNA and total microbial DNA were used for shotgun library construction. DNA concentration was measured by Quant-iT PicoGreen dsDNA Assay kit (Thermo Fisher Scientific), and the fluorescence was detected by EnVision Multilabel Plate Reader (Waltham, MA, USA). Libraries were made using an Illumina Nextera XT Samples Prep kit (Illumina, San Diego, CA, USA), quantified using both Quant-iT PicoGreen dsDNA Assay kit and KAPA Library Quantification kit (Kapa Biosystems, Basel, Switzerland). The size distribution of the libraries was checked by 5300 Fragment Analyzer (Agilent, Santa Clara, CA, USA). Libraries were pooled for sequencing. The concentration of the pooled

libraries was measured using Qubit (Invitrogen, Carlsbad, CA, USA), and the size distribution of the pooled libraries was checked by Agilent Technology 2100 Bioanalyzer using a High Sensitivity DNA chip (Agilent). Sequence was acquired using the Illumina Miseq (250 bp paired-end reads, Illumina) and HiSeq (150 bp paired-end reads, Illumina).

Isolation of bacterial strains

A total of 24 bacterial strains were isolated from the stool samples (19 samples from 12 subjects) using three media: Lysogeny broth (LB) medium in aerobic conditions, and Bifidus selective (BSM) medium (Sigma-Aldrich, St. Louis, MO, USA) and Eosin methylene blue (EMB) medium (Sigma-Aldrich) in anaerobic conditions, which were incubated at 37°C for up to 72 hours. Single colonies were picked and re-streaked in media plates at least three times to isolate pure bacterial strains. The bacterial taxonomy was determined by matrix-assisted laser desorption ionization time-of-flight mass spectrometry (MALDI-TOF MS) using the MALDI-TOF BD instrument (BD, Franklin Lakes, NJ, USA) and default software. The taxonomy was further validated by mapping scaffolds to a 16S rRNA gene database.

Whole genome sequencing of isolated bacterial strains

All 24 bacterial strains were cultured in LB broth overnight or until OD600 > 1. DNA was extracted using the phenol/chloroform method. DNA quality control was as above. TruSeq DNA PCR-Free kit (Illumina) was used to make genomic DNA sequencing libraries; quality control and pooling was as above. Sequencing was performed using the Illumina Miseq (250 bp paired-end reads, Illumina).

In vitro prophage induction and induced VLP sequencing

Overnight cultures of isolated bacterial strains were diluted 1/100 into 10 ml medium and grown until log phase (OD = 0.6). Mitomycin C (Sigma-Aldrich) was then added to a final concentration of 5 μg/ml. The OD values were measured and VLPs were purified after 6 hours of culture. VLPs were purified from the bacterial culture using the same method as stool VLP purification without the homogenization step. The purified VLP DNA was extracted and amplified for virome sequencing as described above, and also enumerated by SYBR Gold staining. RNA phage are generally not thought to form prophages; RNA was interrogated for three samples, and no phage were identified.

In vitro phage infection

Overnight cultures of isolated bacterial strains were diluted 1/100 into 10 ml medium, grown to log phase (OD = 0.6), and then 100 μl bacteria were mixed with 100 μl of serial dilutions of isolated infant stool VLPs with the addition of $MgSO_4$ (with a final concentration of 10 nM). The mixture was incubated at 37°C for 30 min, diluted in 3 ml of warm soft agar, and then plated on a pre-warmed LB plate. Lambda phage was used as a positive control.

16S rRNA gene qPCR

Bacterial abundance was quantified using qPCR of the V1-V2 region of the 16S rRNA gene using a TaqMan-based assay (Applied Biosystems, Foster City, CA, USA). Primer, probe sequences and the PCR program were described in[32] and are presented in Supplementary

Table 6. The reaction was conducted on a 7500 Fast Real Time qPCR system (Thermo Fisher Scientific). Triplicate reactions were performed. Results show the mean values (Supplementary Table 2). The limit of detection in the 16S qPCR assay was determined to be 20 copies per reaction, which was equal to ~2000 copies per gram faces.

Stool VLP sequence read quality control and taxonomic classification

Quality control for the stool VLP reads was performed using the Sunbeam pipeline[33] with a custom Sunbeam extension (https://github.com/guanxiangliang/sbx_dedup). In brief, low quality reads and adapter sequences were removed by Trimmomatic[34], low-complexity reads were identified and discarded by Komplexity (https://github.com/eclarke/komplexity), and then duplicate identical sequences (inferred PCR replicates) were filtered out by BBmap (https://jgi.doe.gov/data-and-tools/bbtools/). Dereplicated reads were aligned using BWA to the host (GRCh38 for human genome and GRCm38 for mouse genome) or phix174 and removed. The quality-controlled reads were classified by Kraken using a custom database which included all complete human, bacterial, archeal, and viral genomes in Refseq release 89 (released on July 9, 2018), with low-complexity regions masked prior to building the database.

To investigate environmental contamination or experimental reagent contamination, negative control samples were analyzed, including empty diaper samples, empty stool container samples, and reagent-only samples. The Decontam package in R was used on the Kraken classification data to remove contaminating species with “prevalence” method at a threshold of 0.5[35]. Taxa including Klebsiella phage 0507-KN2–1, Choristoneura occidentalis granulovirus, Vibrio phage pYD38-A, Pseudomonas phage PPpW-4, Burkholderia phage ST79, Burkholderia phage KS9, Bacillus virus phi29, Simbu orthobunyavirus, and Shamonda orthobunyavirus were removed from downstream analysis.

Stool VLP sequence read assembly and annotation, and phage lifestyle prediction

The quality-controlled reads were assembled into contigs using megahit[36] within each subject. To quantify contigs in each sample, quality-controlled reads were mapped back to the contigs using Bowtie2[37], and the number of mapped reads were calculated by processing Sam files using custom code. To remove differences in sequencing depth, reads per million total reads (RPM) were calculated for each contig. Assembled contigs from virome libraries with length larger than 3000 bp were selected to predict open reading frames (ORFs) using Prodigal in “meta” mode[38]. The predicted ORFs were mapped to the viral protein database in UniProt Knowledgebase (TrEML and Swiss-Prot)[39] using BLASTP with E value < 10e-5.

In order to exclude contigs resulting from contamination, we mapped negative control sample reads to the built VLP contigs. If the maximal RPM of negative control samples for the sample contig was greater than in the stool samples, then that contig was marked as contamination and not used for downstream analysis.

We defined an assembled contig as a viral contig if it had 1) at least one viral protein per 10kbp of VLP contig, and 2) 50% of the predicted ORFs were viral ORFs. The taxonomy of each contig was classified as described previously[40] modified to compile attributions over multiple reading frames to generate a single taxonomic assignment. The ORFs were

assigned with taxa based on the best-hit viral protein in UniProt Knowledgebase. The majority taxonomic assignment over all ORFs within a contig was given to the contig. Contigs that cannot be assigned to any taxa were classified as “Others”. Contigs that were not assigned as “Bacteriophage” were mapped to NCBI nt database with a threshold 80% coverage and 80% identity to futher remove contigs from non-viral genomes. In total, we identified 2552 viral contigs among all 20 subjects (Extended Data Fig. 2a–b). Contigs sharing the same taxonomic assignments were collapsed to yield pooled RPM values for each taxon. Viral richness was calculated by observed species number with RPM > 10. DNA virome reads that could be assigned to our set of viral contigs accounted for 11.3% ± 4.7% (mean ± s.e.m.) at month 0, 31.2% ± 5.6% at month 1, and 37.7% ± 5.4% at month 4 for all non-human reads (Extended Data Fig. 2c–e). Other reads come from contamination, other microorganism genomes and unassigned categories. We conjecture that some of the unassiged reads represent unstudied bacteriophage, where there were not sufficient ORFs matching database viral ORFs to label the contig as viral. For the RNA virome data, assembly from 12 out of 20 subjects yielded contigs larger than 3000 bp. The RNA virome reads that could be assigned to viral contigs accounted for a mean of 4.5% ± 4.7% (mean ± s.e.m.) at month 0, 15.9% ± 13.0% at month 1, and 10.0% ± 5.1% at month 4 of total non-human reads (Extended Data Fig. 2f–h).

Viral contigs were scored as temperate or lytic bacteriophage using PHACTS[24]. In order to obtain strong predictions, only viral contigs with at least 10 ORFs were analyzed. Of 2552 viral contigs, 1029 were classified as “Bacteriophages” and contained more than 10 ORFs and used for the PHACTS analysis. Ten replicate PHACTS predictions were performed. Probability values obtained from PHACTS were standardized between −1 and 1, which was presented as probability of “Lytic” or “Temperate” (Extended Data Fig. 4b).

To test the abundance of crAssphage in the infant gut, we mapped the stool VLP reads to 37 genomes which belong to the crAssphage family[25,26,41,42]. At least 33% genome coverage was considered to be a positive detection (Extended Data Fig. 6). In this analysis, we included stool VLP sequencing data from a group of older healthy children (2–5 years old, n = 19, Supplementary Table 1).

Profiling human-cell viruses

Seven viral families that replicate on human cells were detected by Kraken. To further investigate the accumulation of these viruses, the viral genomes in RefSeq and Viral Neighbor databases that represent these families were retrieved from NCBI. The stool VLP sequences were mapped to these genomes to estimate genome coverage using Bowtie2 with global alignment option[37]. The output sam files were process by Samtools[43], Bedtools[44] and custom code (https://github.com/guanxiangliang/liang2019) to quantify the fraction of the genome covered. We favor use of percent coverage as a metric for genome detection[45]; amplification during sequence library preparation can yield many copies of single genome regions, yielding many sequence reads but with low genome coverage. Comparisons in several studies thus indicate coverage is a more reliable measure. We found that the negative control samples could contain coverage of up to ~10% of a viral genome (Extended Data Fig. 7k).

Human-cell virus qPCR

The numbers of selected human-cell viral genome copies in stool samples were determined by qPCR using TaqMan-based assays (Applied Biosystems). Primers and probes that target Adenoviruses[46], Human Torque teno viruses[47], Enteroviruses[48], Astroviruses[49], Sappovirus GI strains[50] and Norovirus GII strains[51] were used in this study. All primer and probe sequences are listed in Supplementary Table 6. The qPCR reactions were conducted on a 7500 Fast Real Time qPCR system using TaqMan Fast Advanced Master Mix (Thermo Fisher Scientific) in a final volume of 20 μl with 900 nM primers and 250 nM probe. All qPCR reactions were performed without preamplification for both VLP DNA or RNA. Triplicate reactions were performed and the results showing the mean values and standed deviations are listed in Supplementary Table 7.

The availability of metagenomic virome sequence data and qPCR data allowed assessment of qPCR efficiency given sporadic mismatches of viral sequences to qPCR primers. A comparision between virome sequencing and qPCR data is presented in Supplementary Table 4.

Total microbial shotgun metagenome sequencing read quality control and taxonomic classification

The quality control for the shotgun metagenome sequencing reads were performed using the default pipeline in Sunbeam. The quality-controlled reads were classified by Kraken using the same database as was used for stool VLP sequence analysis. To calculate the bacterial richness and diversity, 15,000 paired reads were randomly selected from each sample, and MetaPhlAn2 was used to align reads to different levels of bacterial taxonomy[52]. Bacterial richness was calculated as observed species number, and Shannon diversity was calculated using the Vegan package in R. The Decontam package in R was used to remove contaminating sequences with “prevalence” method at a threshold of 0.5[35].

Bacterial whole genome sequence assembly and quality control

The quality control for the bacterial whole genome sequence reads was performed using Sunbeam without removing low-complexity reads. The quality-controlled reads were assembled by SPAdes[53], followed by SSPACE to make scaffolds[54]. The quality of scaffolds (completeness and contamination) was evaluated by CheckM[55]. The assembled scaffolds revealed good quality for each bacterial strain (Supplementary Table 3).

Integrated analysis of stool VLP, induced VLP, shotgun metagenome and whole genome sequence data

To analyze whether stool viruses were from the stool VLP sequences matched sequences of induced VLP from purified infant gut bacterial strains, reads from induced VLP and stool VLP were mapped to corresponding bacterial scaffolds using Bowtie2[37]. The mapped reads number and bedgraph files for coverage plots were generated using Samtools[43], Deeptools[44]and custom code. Induced VLP sequences from isolated bacterial strains were mapped to the stool virome contigs using the same method to assess whether the induced phages from isolated bacteria were more similar to stool VLP sequences from subjects donating the bacterial strain than in VLPs from unmatched subjects. The prophage genome

annotation was performed by PHASTER[56] targeting bacterial genomic scaffolds longer than 100k.

VLP contigs for analysis were identified as follows 1) We asked whether assembled contigs from the induced VLPs from the 24 purified bacterial strains could be identified in the 24 sequenced bacterial genomes. We required that more than 50% of the induced VLP contigs length was matched to a bacterial genome scaffold or vice versa (Blastn with E value < 10e-10). 2) We asked whether induced VLP contigs recognized as candidate prophage encoded proteins that were present in the UniProt viral protein database. At least 50% of the ORFs were required to be virus-like proteins (Blastp with E value < 10e-5). 3) In the induced VLPs, contigs were required to comprise at least 5% of all reads for inclusion in the analysis.

To evaluate whether the induced prophages from purified bacteria were more similar to stool VLP sequences from infants donating the bacterial strain than in VLPs from unmatched infants, we mapped the stool VLP reads to the corresponding induced VLP contigs from the same infant (within infants) as well as unmatched infants (between infants) using Bowtie2 (Extended Data Fig. 5a).

Several further analyses were performed to investigate the correlation between the proportion of each bacteria in the infant gut community and the proportion of that bacteria's prophages in the infant's gut virome.

Stool VLP sequences were mapped to induced VLP contigs identified above using Bowtie2. The proportion of mapped reads from stool VLP were divided by total stool VLP read number to obtain the proportion, which represents the abundance of each bacterial prophage in the infant gut virome. The proportion of isolated bacteria in the infant gut community was represented by the proportion of shotgun reads that can be mapped to the isolated bacterial genome divide by all nonhuman reads. The bacterial prophages abundance was plotted against the isolated bacterial abundance (Fig. 2d). This analysis was conducted using data based on both mitomycin C induction (Fig. 2d) and spontaneous induction (Extended Data Fig. 5b).

Phage population structure analysis

To interrogate phage populations, 185 samples pooled from both discovery and validation US cohorts were analyzed. Assembled contigs from individual DNA virus libraries with length larger than 3000 bp were selected to predict ORFs as described above. Accurately assigning taxonomic ranking of viral contig is still a challenge, therefore, we preformed a taxonomy-independent population analysis. ORFs were mapped to Pfam database using HmmScan (HMMER 3.1; http://hmmer.org/) with E value < 10e-5. Pfam entries that belong to phages, and those that were shared by phages and bacteria, were selected for further analysis. The coordinates of each Pfam entry on the contigs were identified by custom code, and VLP reads were aligned to these coordinates by featureCounts[57] to evaluate the abundance of each Pfam entry. The Pfam annotations for each sample were cataloged, and a matrix generated for annotation over all samples. Clustering was evaluated using Bray-Curtis dissimilarities. Bray-Curtis dissimilarities were plotted using PCoA, and differences

among groups (infant age, infant feeding type, infant delivery type, infant gender, mother body type, formula type, mother pregnancy induce HTN or diabetes and mother Chorioamnionitis) were tested using PERMANOVA analysis. Continuous variables (gestational age, infant birth weight, household underage number, household number, and mother pregnancy weight gain) were fit to the PCoA ordination by regression using the Envfit function. *P* values were determined using 999 permutations. The analysis were carried out using the Vegan R package.

Bifidobacterium and *Lactobacillus* phage analysis

Forty-two *Lactobacillus* phage genomes were downloaded from RefSeq and used for comparison. RefSeq did not contain any Bifidobacterium phage genome sequences, but two *Bifidobacterium* phage genomes were available in NCBI (accession number GQ141189.1 and MH444512.1) and were used for analysis here. Genome coverage was estimated using the same method as was used for animal virus coverage analysis. The *Bifidobacterium* and *Lactobacillus* phages that were highly covered (> 33%) by sequencing all contain annotated "Integrase" proteins, suggesting temperate replication cycles.

Quantification and statistical analysis

Statistical tests were conducted using R. Non-parametric tests were used for comparing two independent groups (Wilcoxon rank-sum test), two related groups (Wilcoxon signed-rank test), and multiple groups (Kruskal-Wallis test with Bonferroni correction). Non-parametric correlation was performed using Spearman's rank-order correlation (*R* represents Spearman's ρ). Fisher's Exact test was used to test the difference between two categorical variables. *P* values for multiple comparisons were corrected using the Benjamini-Hochberg false discovery rate (FDR) method. $P < 0.05$ or FDR < 0.05 was considered significant. All reported *P* values are from two-sided comparisons. All acquired data were included for analyses.

Gnotobiotic mouse control

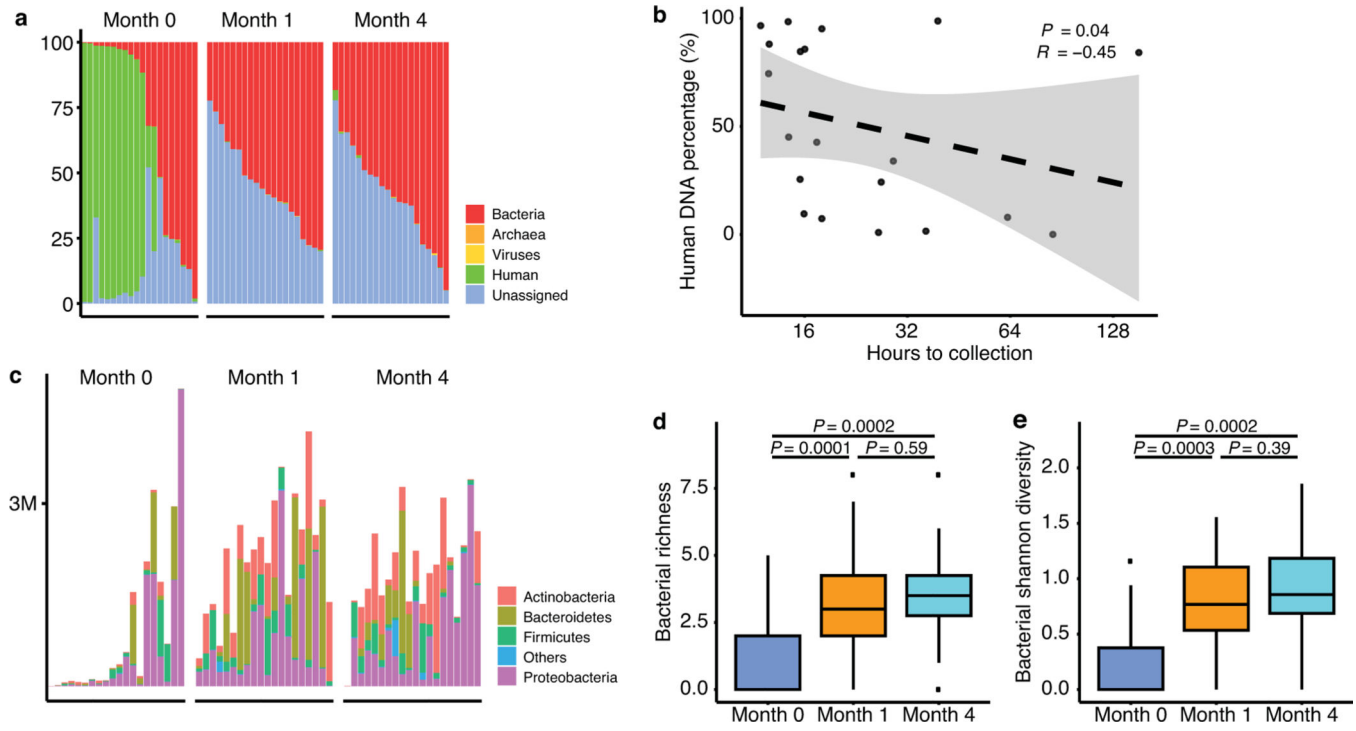
As a control for this study, we prepared and analyzed VLPs from stool samples from gnotobiotic mice, and to our surprise did find viral sequences that were not present in contamination controls. In this case, the particles found were in fact derived from murine endogeneous retroviruses, specifically murine leukemia virus, which is known to be present in the germ line of the mouse strain C57BL/6 used here[58,59]. Evidently endogenous retroviral particles can be shed into the murine gut and detected by our methods, providing a positive control for our analysis of human samples. No additional VLP contigs passing quality filtering were detected.

Possible contribution of Human Endogeneous Retroviruses (HERVs).

Human endogenous retroviruses are another candidate source of viral particles in neonates, and low levels of these sequences could be detected in VLP DNA fractions. However, HERV particles contain RNA, and HERVs were not detected significantly in RNA VLP fractions (Extended Data Fig. 10). Quality control studies showed HERV DNAs were likely

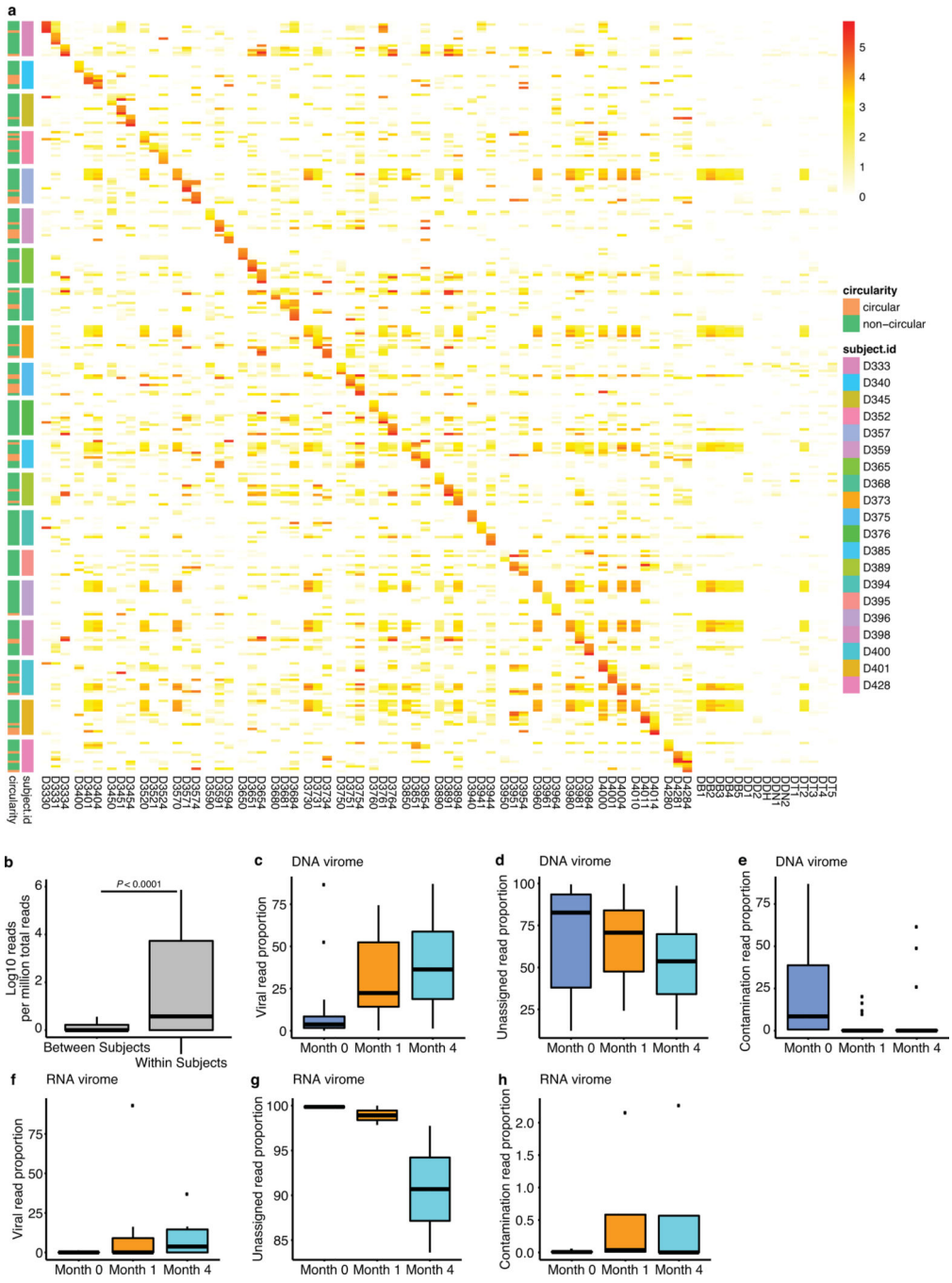
contributed by contaminating human DNA, and were present in proportions predicted given the frequencies of other human genomic repeated sequences.

Extended Data



Extended Data Fig. 1 |. Overview of total stool microbial shotgun metagenomic sequencing.

a, Percentage of reads mapped to human, microbial genomes or unassigned. Times of sampling are indicated above the graphs. Types of DNA detected are summarized at the right. **b**, Correlation between human DNA percentage with sampling time after delivery using month 0 samples (n = 20). The percentage of human DNA is shown on the y-axis, and the sampling time after delivery is shown on the x-axis. The black dashed line shows the linear regression line and the gray-shaded region shows the 95% confidence interval for the slope. Two-sided Spearman's rank-order correlation method was used to test significance (*R* represents Spearman's ρ). **c**, Taxonomic composition of bacteria at the phylum level. The total read number was shown on the y-axis, and x-axis represents different samples. **d**, Bacterial richness. The Y-axis indicates the richness calculated by observed species number. **e**, Bacterial diversity. Y-axis indicates the Shannon index. In **d** and **e**, two-sided Wilcoxon rank-sum test was used to test the difference between different ages (n = 20 subjects at three time points). The horizontal lines in boxplots represent the third quartile, median and first quartile. The dots represent the outliers.

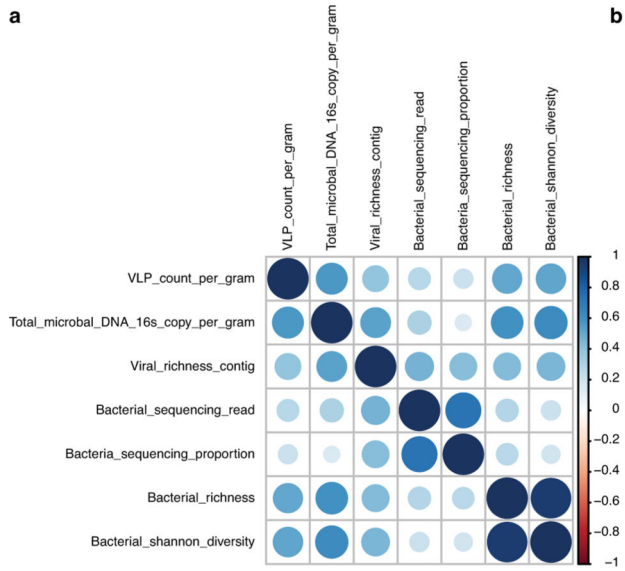


Extended Data Fig. 2 |. Summary of infant stool virome sequencing.

a, Heat map summarizing representation of the top five most abundant DNA viral contigs in each sample. Samples are grouped sequentially by subject on both the x-axis and y-axis. The last group of subjects on x-axis are negative control samples. Circularity indicates whether a contig is circular (orange color) or not (light green color). The heatmap map color represents the abundance (log transformed reads per million total reads value) of each contig in each sample. **b**, Contig read abundance compared between different subjects versus within the same subjects. Time points were pooled for each individual. **c-e**, Percentage of DNA virome

reads assigned as Viruses (**c**), unassigned (**d**), and contamination (**e**). **f**-**h**, Percentage of RNA virome reads assigned as Viruses (**f**), unassigned (**g**), and contamination (**h**). In **b**-**h**, n = 20 subjects at three time points were tested. The horizontal lines in boxplots represent the third quartile, median and first quartile. The dots represent the outliers.

a

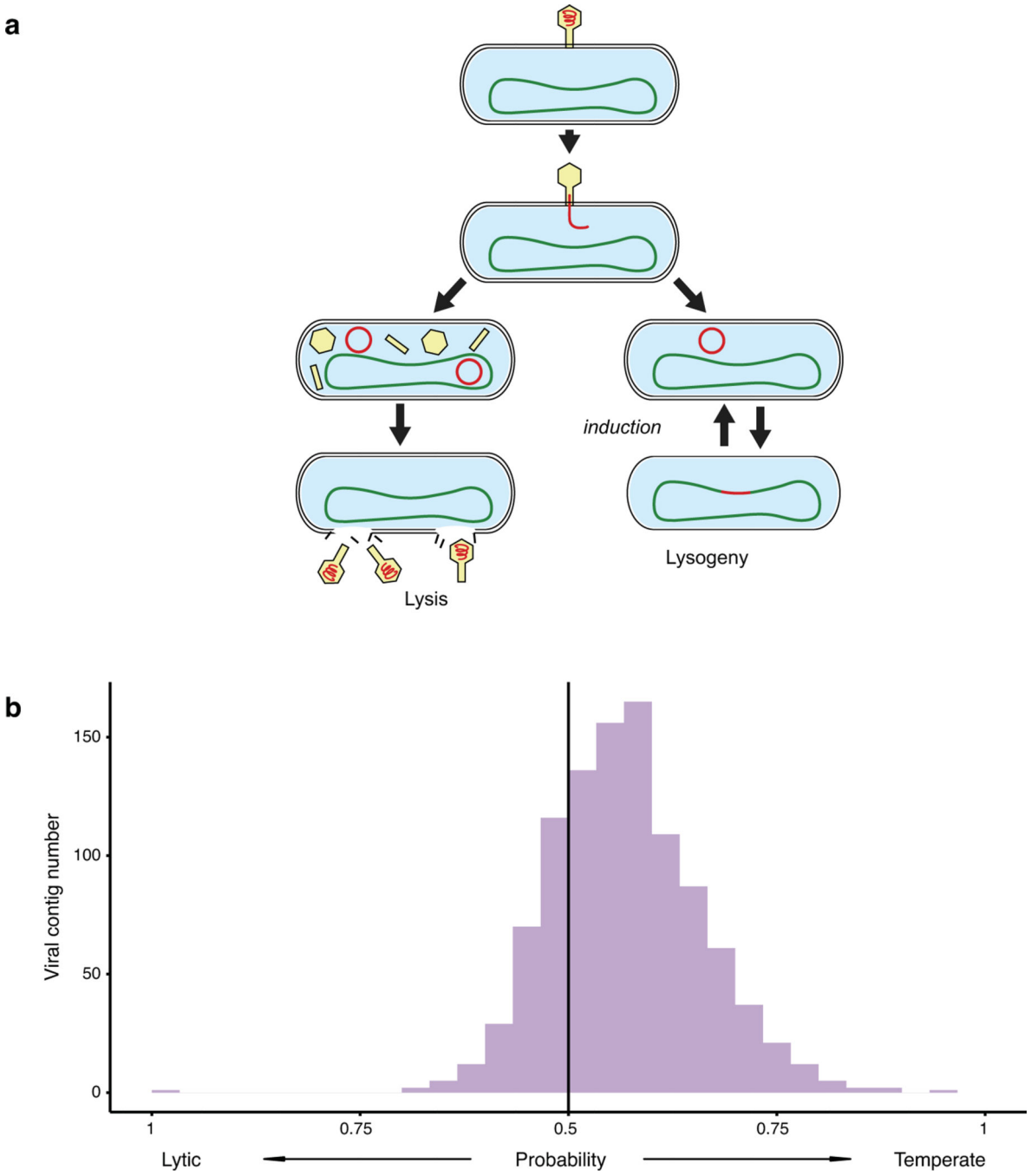


b

data1	data2	R.value	P.value	FDR
VLP_count_per_gram	VLP_count_per_gram	1.00	0.00e+00	0.00e+00
VLP_count_per_gram	Viral_richness_contig	0.40	1.68e−03	2.77e−03
VLP_count_per_gram	Total_microbal_DNA_16s_copy_per_gram	0.57	1.61e−06	5.68e−06
VLP_count_per_gram	Bacterial_shannon_diversity	0.52	1.92e−05	5.76e−05
VLP_count_per_gram	Bacterial_sequencing_read	0.28	3.00e−02	3.86e−02
VLP_count_per_gram	Bacterial_richness	0.51	2.98e−05	8.32e−05
VLP_count_per_gram	Bacteria_sequencing_proportion	0.22	9.81e−02	1.09e−01
Viral_richness_contig	VLP_count_per_gram	0.40	1.68e−03	2.77e−03
Viral_richness_contig	Viral_richness_contig	1.00	0.00e+00	0.00e+00
Viral_richness_contig	Total_microbal_DNA_16s_copy_per_gram	0.53	1.20e−05	3.88e−05
Viral_richness_contig	Bacterial_shannon_diversity	0.46	2.20e−04	4.81e−04
Viral_richness_contig	Bacterial_sequencing_read	0.47	1.40e−04	3.65e−04
Viral_richness_contig	Bacterial_richness	0.43	5.69e−04	1.07e−03
Viral_richness_contig	Bacteria_sequencing_proportion	0.43	6.06e−04	1.09e−03
Total_microbal_DNA_16s_copy_per_gram	VLP_count_per_gram	0.57	1.61e−06	5.68e−06
Total_microbal_DNA_16s_copy_per_gram	Viral_richness_contig	0.53	1.20e−05	3.88e−05
Total_microbal_DNA_16s_copy_per_gram	Total_microbal_DNA_16s_copy_per_gram	1.00	0.00e+00	0.00e+00
Total_microbal_DNA_16s_copy_per_gram	Bacterial_shannon_diversity	0.63	8.00e−08	3.41e−07
Total_microbal_DNA_16s_copy_per_gram	Bacterial_sequencing_read	0.32	1.28e−02	1.76e−02
Total_microbal_DNA_16s_copy_per_gram	Bacterial_richness	0.60	3.85e−07	1.48e−06
Total_microbal_DNA_16s_copy_per_gram	Bacteria_sequencing_proportion	0.15	2.51e−01	2.58e−01
Bacterial_shannon_diversity	VLP_count_per_gram	0.52	1.92e−05	5.76e−05
Bacterial_shannon_diversity	Viral_richness_contig	0.46	2.20e−04	4.81e−04
Bacterial_shannon_diversity	Total_microbal_DNA_16s_copy_per_gram	0.63	8.00e−08	3.41e−07
Bacterial_shannon_diversity	Bacterial_shannon_diversity	1.00	0.00e+00	0.00e+00
Bacterial_shannon_diversity	Bacterial_sequencing_read	0.22	9.42e−02	1.07e−01
Bacterial_shannon_diversity	Bacterial_richness	0.95	4.53e−32	2.82e−31
Bacterial_shannon_diversity	Bacteria_sequencing_proportion	0.19	1.40e−01	1.52e−01
Bacterial_sequencing_read	VLP_count_per_gram	0.28	3.00e−02	3.86e−02
Bacterial_sequencing_read	Viral_richness_contig	0.47	1.40e−04	3.65e−04
Bacterial_sequencing_read	Total_microbal_DNA_16s_copy_per_gram	0.32	1.28e−02	1.76e−02
Bacterial_sequencing_read	Bacterial_shannon_diversity	0.22	9.42e−02	1.07e−01
Bacterial_sequencing_read	Bacterial_sequencing_read	1.00	0.00e+00	0.00e+00
Bacterial_sequencing_read	Bacterial_richness	0.29	2.24e−02	2.98e−02
Bacterial_sequencing_read	Bacteria_sequencing_proportion	0.74	0.00e+00	0.00e+00
Bacterial_richness	VLP_count_per_gram	0.51	2.98e−05	8.32e−05
Bacterial_richness	Viral_richness_contig	0.43	5.69e−04	1.07e−03
Bacterial_richness	Total_microbal_DNA_16s_copy_per_gram	0.60	3.85e−07	1.48e−06
Bacterial_richness	Bacterial_shannon_diversity	0.95	4.53e−32	2.82e−31
Bacterial_richness	Bacterial_sequencing_read	0.29	2.24e−02	2.98e−02
Bacterial_richness	Bacterial_richness	1.00	0.00e+00	0.00e+00
Bacterial_richness	Bacteria_sequencing_proportion	0.27	3.42e−02	4.26e−02
Bacteria_sequencing_proportion	VLP_count_per_gram	0.22	9.81e−02	1.09e−01
Bacteria_sequencing_proportion	Viral_richness_contig	0.43	6.06e−04	1.09e−03
Bacteria_sequencing_proportion	Total_microbal_DNA_16s_copy_per_gram	0.15	2.51e−01	2.58e−01
Bacteria_sequencing_proportion	Bacterial_shannon_diversity	0.19	1.40e−01	1.52e−01
Bacteria_sequencing_proportion	Bacterial_sequencing_read	0.74	0.00e+00	0.00e+00
Bacteria_sequencing_proportion	Bacterial_richness	0.27	3.42e−02	4.26e−02
Bacteria_sequencing_proportion	Bacteria_sequencing_proportion	1.00	0.00e+00	0.00e+00

Extended Data Fig. 3 |. Correlation between viral and bacterial community.

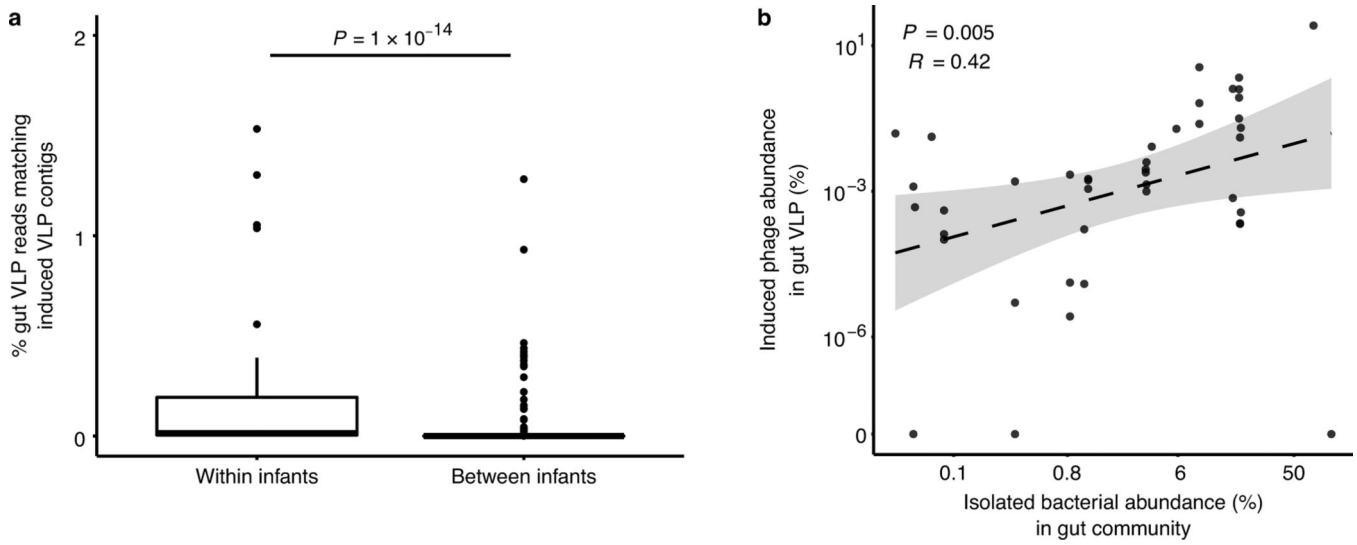
a, Pairwise correlations among sample measures including: VLP count number, bacterial 16S qPCR copy number, viral richness, bacterial sequence read proportion, bacteria richness and diversity. The size of circles indicates the R value of the correlation. Blue color indicates positive correlation, and red color indicates negative correlation. Samples from different time points were pooled (n = 60). Two-sided Spearman's rank-order correlation method was used in this analysis. **b**, As in **a**, but showing the raw data of the statistical analysis. *P* values, FDR corrected p values and *R* (Spearman's ρ) are presented.



Extended Data Fig. 4 |. Life cycles of bacteriophages.

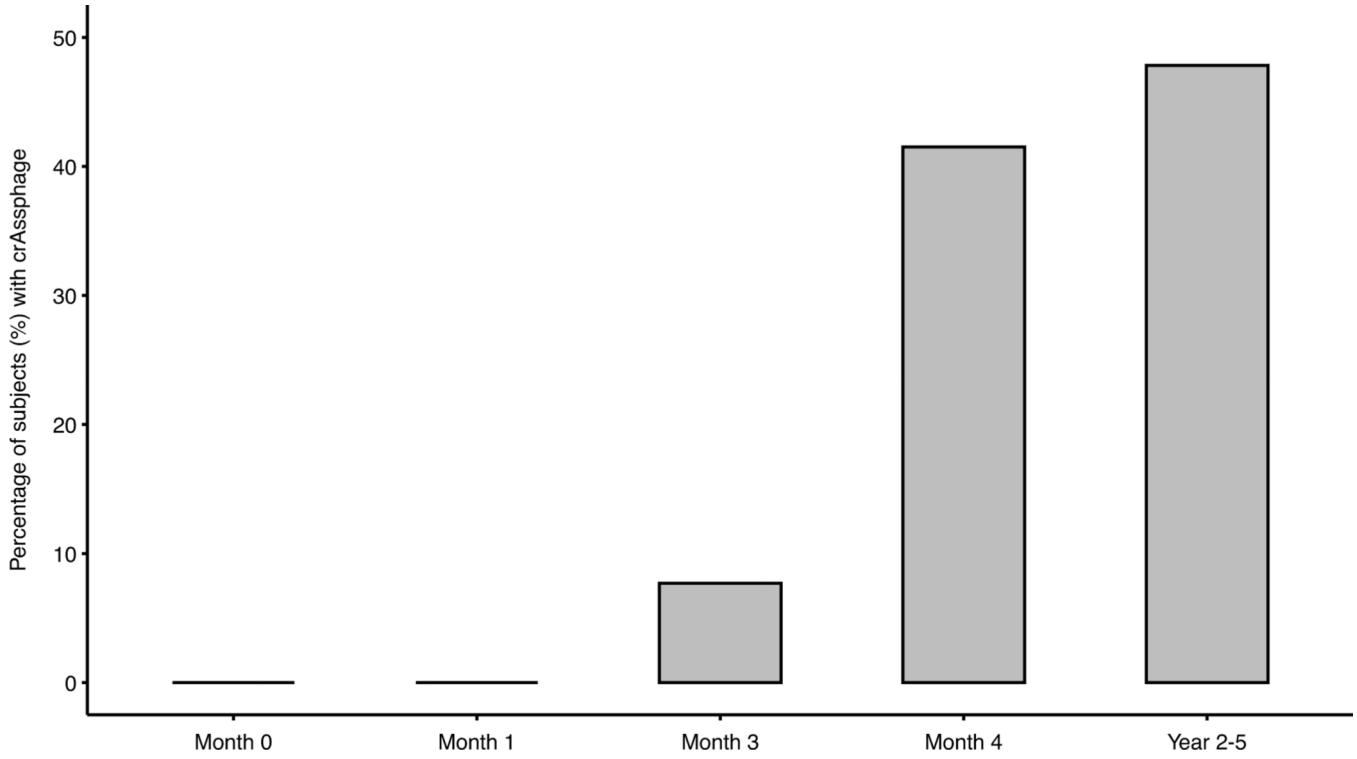
a, Diagram of lytic and lysogenic bacteriophage replication (based on Ptashne[22]). Not shown are additional phage replication strategies including chronic infection and pseudolysogeny. **b**, Prediction of replication modes from contig sequences using PHACTS. The X-axis indicates the probability that a contig belongs to a lytic or temperate phage predicted by PHACTS. The Y-axis indicates the viral contig number. In total, 1029 bacteriophage contigs with at least 10 open read frames were used in this analysis. Of 1029 contigs, 233 were predicted as lytic and 794 were predicted as temperate. Probability values

obtained from PHACTS were standardized between −1 and 1, which was presented as probability to “Lytic” or “Temperate”.

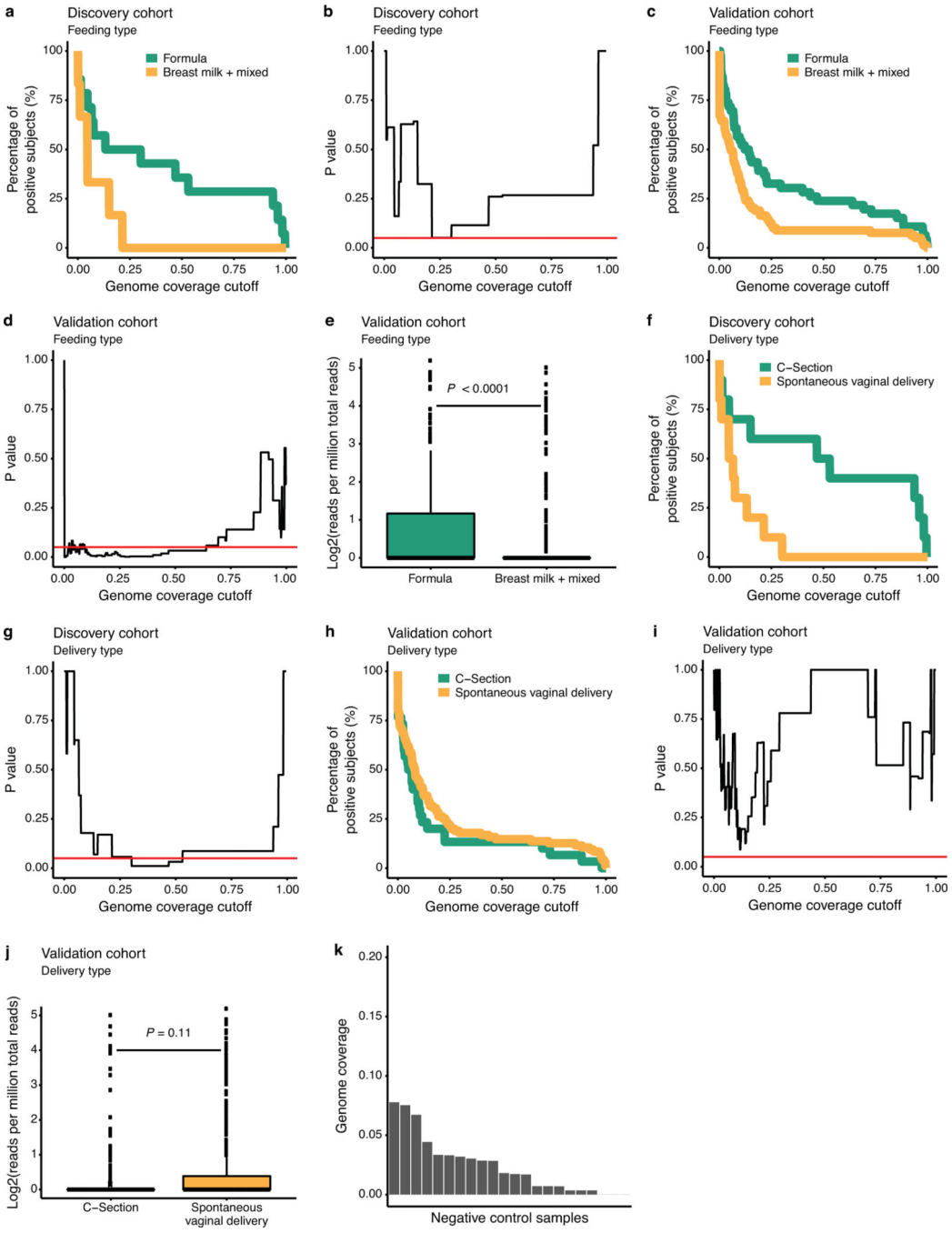


Extended Data Fig. 5 |. Prophage induction in the early life virome.

a, Comparison of the extent of sequence alignment of induced VLP sequences from bacterial strains compared to VLP sequences from stool. Contigs were generated from mitomycin C induced VLPs from purified stool bacterial strains (n = 33 phage contigs from 16 bacterial isolates), then VLP reads from feces aligned to these contigs and quantified. “Within infants” indicates matching stool VLP to induced VLP from purified bacteria for samples all from the same infant, and “Between infants” indicates alignment of stool VLP versus induced VLP from different infants. The horizontal lines in boxplots represent the third quartile, median and first quartile. The dots represent the outliers. Samples were compared using the two-sided Wilcoxon rank-sum test. **b**, Correlation between the proportion of each bacteria in the infant gut community and the proportion of that bacteria’s prophages in the infant’s gut virome. This plot is based on VLP sequences of phages produced by spontaneous induction (n = 42 phage contigs from 20 bacterial isolates). This is different from Fig. 2d, which is based on VLP sequences of phages produced after induction with mitomycin C. The black dashed line shows the linear regression line and the gray-shaded region shows the 95% confidence interval for the slope. Correlation was tested using two-sided Spearman’s rank-order correlation (*R* represents Spearman’s ρ).



Extended Data Fig. 6 |. Colonization by crAssphage in different age groups.
Grey color bars indicate the percentage of crAssphage positive subjects (as scored by requiring that the crAssphage genome was more than 33% covered by sequence reads from stool VLPs).



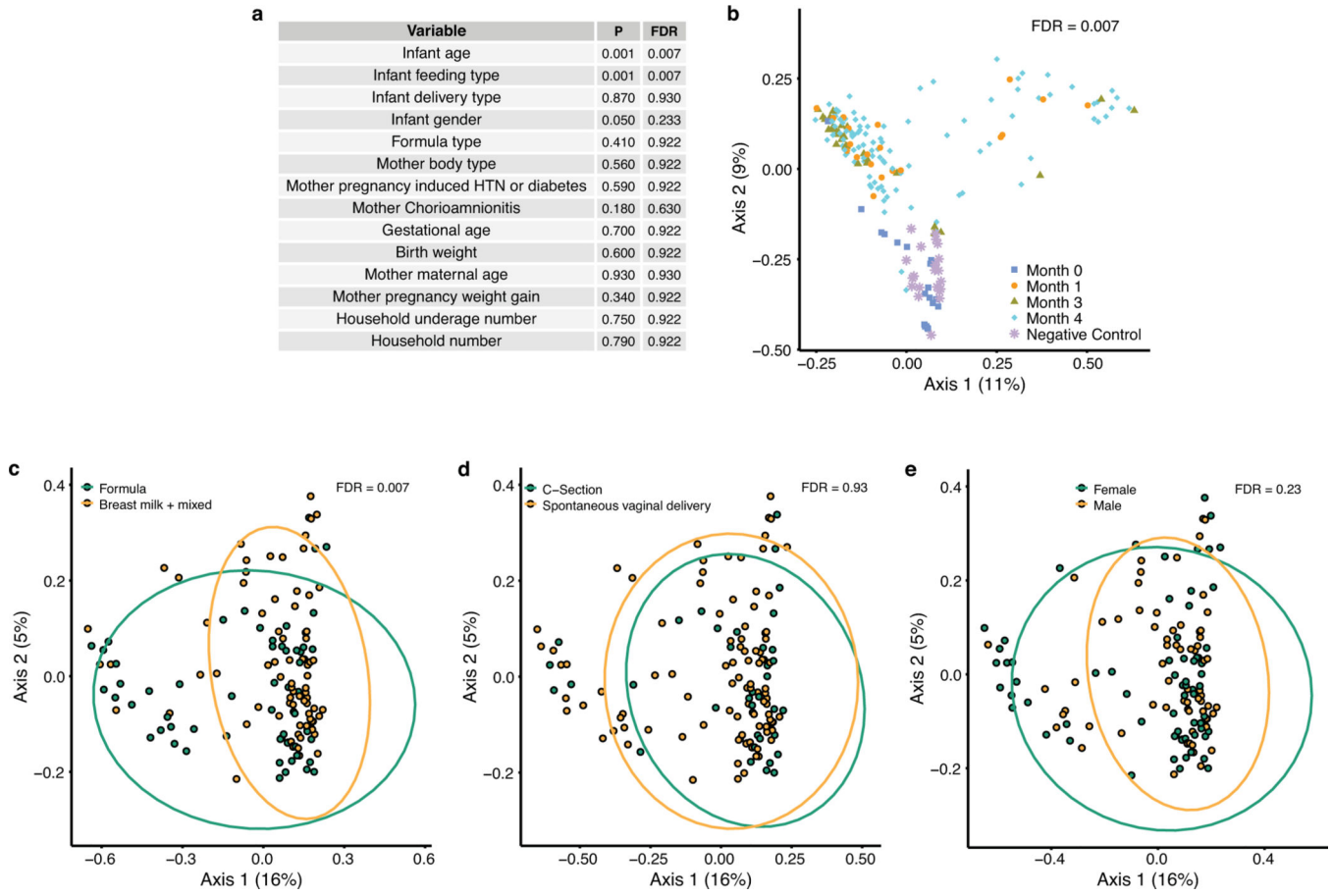
Extended Data Fig. 7 |. Animal cell viruses profiling by virome sequencing.

a, **c**, **f**, **h** Percentage of animal cell virus positive subject using different viral genome coverage cutoffs in the discovery cohort (**a**, **f**) and validation cohort (**c**, **h**). The percentage of animal cell virus positive subjects is shown on the y-axis, and different viral genome coverage cutoffs are shown in x-axis. The green line presents the data from subjects with Formula feeding type (**a**, **c**) or C-section delivery (**f**, **h**), the and yellow line presents the data from subjects with Breast milk or mixed feeding type (**a**, **c**) or spontaneous vaginal delivery type (**f**, **h**). **b**, **d**, **g**, **i** Two-sided Fisher's exact test on infant feeding types (**b**, **d**) and delivery

types (**g**, **i**) using different viral genome coverage cutoff in the discovery cohort (**b**, **g**) and validation cohort (**d**, **i**). The *P* values are shown on y-axis, and different viral genome coverage cutoffs are shown in x-axis. The horizontal red line indicates $P = 0.05$. **e**, **j**, Comparison of relative abundance of animal cell viruses between different feeding types (**e**) and delivery types (**j**). The abundance (reads per million total reads after log transformation) is shown on y-axis. Two-sided Wilcoxon rank-sum test was used to test the difference. The horizontal lines in boxplots represent the third quartile, median and first quartile. The dots represent the outliers. **K**, Genome coverage fraction of negative control samples for animal cell viruses. The maximal animal viral genome coverage faction for each negative control sample (n = 25) is shown on Y-axis. Different negative control samples are shown on x-axis. Note that coverage never exceeds 10%. In **a**, **b**, **f**, and **g**, n = 20 samples from discovery cohort were used, and in **c**, **d**, **e**, **h**, **I**, and **j**, n = 125 samples from validation cohort were used.

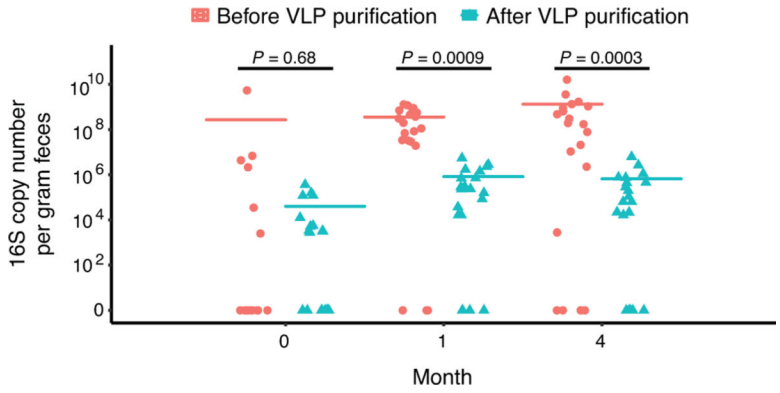
a

Variable	P	FDR
Infant age	0.001	0.007
Infant feeding type	0.001	0.007
Infant delivery type	0.870	0.930
Infant gender	0.050	0.233
Formula type	0.410	0.922
Mother body type	0.560	0.922
Mother pregnancy induced HTN or diabetes	0.590	0.922
Mother Chorioamnionitis	0.180	0.630
Gestational age	0.700	0.922
Birth weight	0.600	0.922
Mother maternal age	0.930	0.930
Mother pregnancy weight gain	0.340	0.922
Household underage number	0.750	0.922
Household number	0.790	0.922



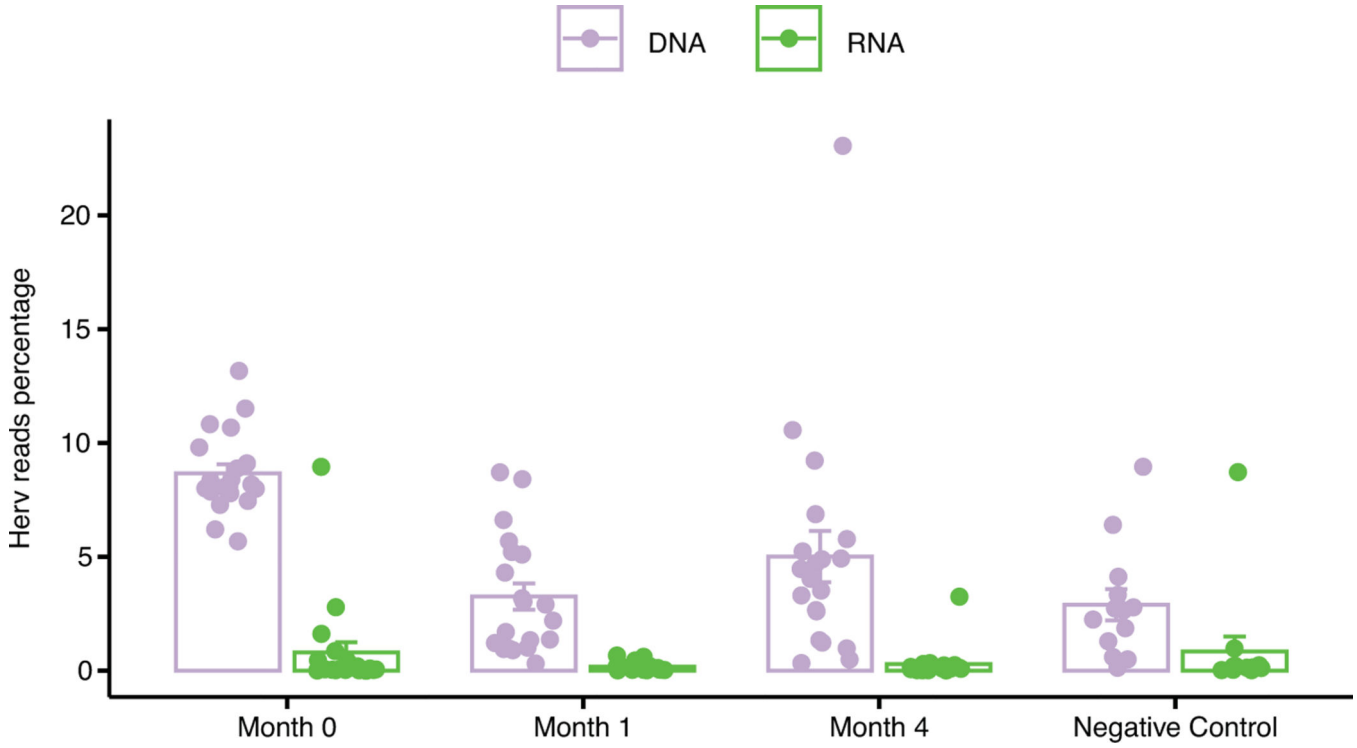
Extended Data Fig. 8 |. Phage population structure.

a, Statistical test of the association of clinical variables with phage population structure. Variables are shown in the first column. *P* values and FDR corrected p values are shown in the second and third column. All categorized variables, such as infant age, infant feeding type, infant delivery type, infant gender, mother body type, formula type, mother pregnancy induce HTN or diabetes and mother Chorioamnionitis were tested by PERMANOVA. Continuous variables, including gestational age, infant birth weight, household underage number, household number, and mother pregnancy weight gain were tested by Envfit. All samples from both discovery US and validation US cohorts (n = 185) were used to test infant age effects, and pooled samples at month 3 and month 4 from both discovery US and validation US cohorts (n = 145) were used to test other variables. **b**, Principal Coordinate Analysis (PCoA) plot based on phage Pfam counts per sample, colored by infant ages. This analysis is based on the Bray-Curtis dissimilarity index for all stool samples from both discovery US and validation US cohorts (n = 185). Negative control samples were not included for Bray-Curtis dissimilarity assessment and statistical test. **c**, **d**, **e**, Principal Coordinate Analysis (PCoA) plot of phage Pfam components, colored by infant feeding types (**c**), delivery type (**d**), and infant gender (**e**). This analysis is based on pooled samples at month 3 and month 4 from both discovery US and validation US cohorts (n = 145), and mentioned in **a**, PERMANOVA was used to test the differences. FDR corrected p values are shown.



Extended Data Fig. 9 |. 16S qPCR before and after VLPs purification.

Red and light blue dots show before and after separately, and the horizontal lines represent means (n = 20 subjects at three time points were tested). Two-sided Wilcoxon signed-rank test was used to test the difference.



Extended Data Fig. 10 |. Percentage of DNA aligning to sequences of human endogenous retroviruses in each sample.

The percentage of HERV sequences in stool VLP is shown on the y-axis. Sample type and time point is shown on the x-axis. The proportion of HERV sequences paralleled those of LINE and SINE elements, indicating they are derived from human DNA contamination. Barplot shows the mean ± s.e.m., and n = 20 subjects at three time points were tested.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank Jian You and coworkers for help with imaging, Mark Goulian and coworkers for bacterial strains, and Forest Rohwer and associates for assistance with phage induction. We are grateful to members of the Bushman laboratory for help and suggestions, and Laurie Zimmerman for artwork and help with the manuscript. The Botswana Infant Microbiome Study team would like to thank Copan Italia for their donation of the eNAT® media and flocked swabs used for the collection of rectal swab specimens.

Funding: This work was supported by NIH grants R61-HL137063 (FDB), R01-HL113252 (FDB), and R01DK107565 (GDW, EF). The project described was also supported by the Penn Center for AIDS Research (P30 AI 045008) (FDB), the PennCHOP Microbiome Program (FDB, GDW, RNB), and a Tobacco Formula grant under the Commonwealth Universal Research Enhancement (C.U.R.E) program (grant number SAP # 4100068710) (FDB, RNB). Funding was also provided by an unrestricted donation from the American Beverage Foundation for a Healthy America to the Children's Hospital of Philadelphia to support the Healthy Weight Program (CZ, KB, EF, JSG, BSZ). The project described was also supported by the National Center for Advancing Translational Sciences, National Institutes of Health, through Grant UL1TR000003 and UL1TR001878 (EF). Funding for the Botswana Infant Microbiome Study was provided by the Duke Center for AIDS Research (5P30 AI064518) and the NIH (K23 AI135090).

References

1. Breitbart M et al. Viral diversity and dynamics in an infant gut. Res Microbiol 159, 367–373, doi:10.1016/j.resmic.2008.04.006 (2008). [PubMed: 18541415]
2. Lim ES et al. Early life dynamics of the human gut virome and bacterial microbiome in infants. Nat Med 21, 1228–1234, doi:10.1038/nm.3950 (2015). [PubMed: 26366711]
3. Liu L et al. Global, regional, and national causes of under-5 mortality in 2000–15: an updated systematic analysis with implications for the Sustainable Development Goals. Lancet 388, 3027–3035, doi:10.1016/S0140-6736(16)31593-8 (2016). [PubMed: 27839855]
4. Oude Munnink BB & van der Hoek L Viruses Causing Gastroenteritis: The Known, The New and Those Beyond. Viruses 8, doi:10.3390/v8020042 (2016).
5. Kim MS, Park EJ, Roh SW & Bae JW Diversity and abundance of single-stranded DNA viruses in human feces. Appl Environ Microbiol 77, 8062–8070, doi:10.1128/AEM.06331-11 (2011). [PubMed: 21948823]
6. Lepage P et al. Dysbiosis in inflammatory bowel disease: a role for bacteriophages? Gut 57, 424–425, doi:10.1136/gut.2007.134668 (2008). [PubMed: 18268057]
7. Hoyles L et al. Characterization of virus-like particles associated with the human faecal and caecal microbiota. Res Microbiol 165, 803–812, doi:10.1016/j.resmic.2014.10.006 (2014). [PubMed: 25463385]
8. Bahl R et al. Infant feeding patterns and risks of death and hospitalization in the first half of infancy: multicentre cohort study. Bull World Health Organ 83, 418–426, doi:/S0042-96862005000600009 (2005). [PubMed: 15976892]
9. Arifeen S et al. Exclusive breastfeeding reduces acute respiratory infection and diarrhea deaths among infants in Dhaka slums. Pediatrics 108, E67 (2001). [PubMed: 11581475]
10. Victora CG et al. Infant feeding and deaths due to diarrhea. A case-control study. Am J Epidemiol 129, 1032–1041 (1989). [PubMed: 2705424]
11. Aagaard K et al. The placenta harbors a unique microbiome. Sci Transl Med 6, 237ra265, doi:10.1126/scitranslmed.3008599 (2014).
12. Lauder AP et al. Comparison of placenta samples with contamination controls does not provide evidence for a distinct placenta microbiota. Microbiome 4, 29, doi:10.1186/s40168-016-0172-3 (2016). [PubMed: 27338728]
13. Theis KR et al. Does the human placenta delivered at term have a microbiota? Results of cultivation, quantitative real-time PCR, 16S rRNA gene sequencing, and metagenomics. Am J Obstet Gynecol 220, 267 e261–267 e239, doi:10.1016/j.ajog.2018.10.018 (2019). [PubMed: 30832984]
14. de Goffau MC et al. Human placenta has no microbiome but can contain potential pathogens. Nature 572, 329–334, doi:10.1038/s41586-019-1451-5 (2019). [PubMed: 31367035]
15. Baumann-Dudenhoeffer AM, D'Souza AW, Tarr PI, Warner BB & Dantas G Infant diet and maternal gestational weight gain predict early metabolic maturation of gut microbiomes. Nat Med 24, 1822–1829, doi:10.1038/s41591-018-0216-2 (2018). [PubMed: 30374198]
16. Reyes A et al. Viruses in the faecal microbiota of monozygotic twins and their mothers. Nature 466, 334–338, doi:10.1038/nature09199 (2010). [PubMed: 20631792]
17. Minot S, Grunberg S, Wu GD, Lewis JD & Bushman FD Hypervariable loci in the human gut virome. Proc Natl Acad Sci U S A 109, 3962–3966, doi:10.1073/pnas.1119061109 (2012). [PubMed: 22355105]
18. Reyes A, Semenkovich NP, Whiteson K, Rohwer F & Gordon JI Going viral: next-generation sequencing applied to phage populations in the human gut. Nat Rev Microbiol 10, 607–617, doi:10.1038/nrmicro2853 (2012). [PubMed: 22864264]
19. Aggarwala V, Liang G & Bushman FD Viral communities of the human gut: metagenomic analysis of composition and dynamics. Mob DNA 8, 12, doi:10.1186/s13100-017-0095-y (2017). [PubMed: 29026445]
20. Wolf YI et al. Origins and Evolution of the Global RNA Virome. MBio 9, doi:10.1128/mBio.02329-18 (2018).

21. Jacob F, Sussman R & Monod J On the nature of the repressor ensuring the immunity of lysogenic bacteria. C R Hebd Seances Acad Sci 254, 4214–4216 (1962).
22. Ptashne M A Genetic Switch. (Blackwell Scientific Publications and Cell Press, 1986).
23. Jacob F & Wollman E Spontaneous induction of the development of bacteriophage lambda during genetic recombination in Escherichia coli K12. C R Hebd Seances Acad Sci 239, 317–319 (1954).
24. McNair K, Bailey BA & Edwards RA PHACTS, a computational approach to classifying the lifestyle of phages. Bioinformatics 28, 614–618, doi:10.1093/bioinformatics/bts014 (2012). [PubMed: 22238260]
25. Dutilh BE et al. A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. Nat Commun 5, 4498, doi:10.1038/ncomms5498 (2014). [PubMed: 25058116]
26. Shkoporov AN et al. PhiCrAss001 represents the most abundant bacteriophage family in the human gut and infects Bacteroides intestinalis. Nat Commun 9, 4781, doi:10.1038/s41467-018-07225-7 (2018). [PubMed: 30429469]
27. Turin CG & Ochoa TJ The Role of Maternal Breast Milk in Preventing Infantile Diarrhea in the Developing World. Curr Trop Med Rep 1, 97–105, doi:10.1007/s40475-014-0015-x (2014). [PubMed: 24883263]
28. Newburg DS, Ruiz-Palacios GM & Morrow AL Human milk glycans protect infants against enteric pathogens. Annu Rev Nutr 25, 37–58, doi:10.1146/annurev.nutr.25.050304.092553 (2005). [PubMed: 16011458]
29. Lewis ED, Richard C, Larsen BM & Field CJ The Importance of Human Milk for Immunity in Preterm Infants. Clin Perinatol 44, 23–47, doi:10.1016/j.clp.2016.11.008 (2017). [PubMed: 28159208]
30. Chehoud C et al. Transfer of Viral Communities between Human Individuals during Fecal Microbiota Transplantation. mBio 7, e00322–00316, doi:10.1128/mBio.00322-16 (2016). [PubMed: 27025251]
31. Wang D et al. Viral Discovery and Sequence Recovery Using DNA Microarrays. PLOS Biology 1, e2, doi:10.1371/journal.pbio.0000002 (2003). [PubMed: 14624234]
32. Hill DA et al. Metagenomic analyses reveal antibiotic-induced temporal and spatial changes in intestinal microbiota with associated alterations in immune cell homeostasis. Mucosal Immunol 3, 148–158, doi:10.1038/mi.2009.132 (2010). [PubMed: 19940845]
33. Clarke EL et al. Sunbeam: an extensible pipeline for analyzing metagenomic sequencing experiments. Microbiome 7, 46, doi:10.1186/s40168-019-0658-x (2019). [PubMed: 30902113]
34. Bolger AM, Lohse M & Usadel B Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114–2120, doi:10.1093/bioinformatics/btu170 (2014). [PubMed: 24695404]
35. Davis NM, Proctor DM, Holmes SP, Relman DA & Callahan BJ Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. Microbiome 6, 226, doi:10.1186/s40168-018-0605-2 (2018). [PubMed: 30558668]
36. Li D, Liu CM, Luo R, Sadakane K & Lam TW MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics 31, 1674–1676, doi:10.1093/bioinformatics/btv033 (2015). [PubMed: 25609793]
37. Langmead B & Salzberg SL Fast gapped-read alignment with Bowtie 2. Nat Methods 9, 357–359, doi:10.1038/nmeth.1923 (2012). [PubMed: 22388286]
38. Hyatt D et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11, 119, doi:10.1186/1471-2105-11-119 (2010). [PubMed: 20211023]
39. Pundir S, Magrane M, Martin MJ, O'Donovan C & UniProt C Searching and Navigating UniProt Databases. Curr Protoc Bioinformatics 50, 1 27 21–10, doi:10.1002/0471250953.bi0127s50 (2015).
40. Minot S et al. Rapid evolution of the human gut virome. Proc Natl Acad Sci U S A 110, 12450–12455, doi:10.1073/pnas.1300833110 (2013). [PubMed: 23836644]
41. Yutin N et al. Discovery of an expansive bacteriophage family that includes the most abundant viruses from the human gut. Nat Microbiol 3, 38–46, doi:10.1038/s41564-017-0053-y (2018). [PubMed: 29133882]

42. 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 e656, doi:10.1016/j.chom.2018.10.002 (2018). [PubMed: 30449316]
43. Li H et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079, doi:10.1093/bioinformatics/btp352 (2009). [PubMed: 19505943]
44. Quinlan AR & Hall IM BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842, doi:10.1093/bioinformatics/btq033 (2010). [PubMed: 20110278]
45. Abbas AA et al. Bidirectional transfer of Anelloviridae lineages between graft and host during lung transplantation. Am J Transplant 19, 1086–1097, doi:10.1111/ajt.15116 (2019). [PubMed: 30203917]
46. Jothikumar N et al. Quantitative real-time PCR assays for detection of human adenoviruses and identification of serotypes 40 and 41. Appl Environ Microbiol 71, 3131–3136, doi:10.1128/AEM.71.6.3131-3136.2005 (2005). [PubMed: 15933012]
47. Abbas AA et al. The Perioperative Lung Transplant Virome: Torque Teno Viruses Are Elevated in Donor Lungs and Show Divergent Dynamics in Primary Graft Dysfunction. Am J Transplant 17, 1313–1324, doi:10.1111/ajt.14076 (2017). [PubMed: 27731934]
48. Verstrepen WA, Kuhn S, Kockx MM, Van De Vyvere ME & Mertens AH Rapid detection of enterovirus RNA in cerebrospinal fluid specimens with a novel single-tube real-time reverse transcription-PCR assay. J Clin Microbiol 39, 4093–4096, doi:10.1128/JCM.39.11.4093-4096.2001 (2001). [PubMed: 11682535]
49. van Maarseveen NM, Wessels E, de Brouwer CS, Vossen AC & Claas EC Diagnosis of viral gastroenteritis by simultaneous detection of Adenovirus group F, Astrovirus, Rotavirus group A, Norovirus genogroups I and II, and Sapovirus in two internally controlled multiplex real-time PCR assays. J Clin Virol 49, 205–210, doi:10.1016/j.jcv.2010.07.019 (2010). [PubMed: 20829103]
50. Oka T et al. Detection of human sapovirus by real-time reverse transcription-polymerase chain reaction. J Med Virol 78, 1347–1353, doi:10.1002/jmv.20699 (2006). [PubMed: 16927293]
51. Rolfe KJ et al. An internally controlled, one-step, real-time RT-PCR assay for norovirus detection and genogrouping. J Clin Virol 39, 318–321, doi:10.1016/j.jcv.2007.05.005 (2007). [PubMed: 17604686]
52. Truong DT et al. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nat Methods 12, 902–903, doi:10.1038/nmeth.3589 (2015). [PubMed: 26418763]
53. Bankevich A et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 19, 455–477, doi:10.1089/cmb.2012.0021 (2012). [PubMed: 22506599]
54. Boetzer M, Henkel CV, Jansen HJ, Butler D & Pirovano W Scaffolding pre-assembled contigs using SSPACE. Bioinformatics 27, 578–579, doi:10.1093/bioinformatics/btq683 (2011). [PubMed: 21149342]
55. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P & Tyson GW CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res 25, 1043–1055, doi:10.1101/gr.186072.114 (2015). [PubMed: 25977477]
56. Arndt D et al. PHASTER: a better, faster version of the PHAST phage search tool. Nucleic Acids Res 44, W16–21, doi:10.1093/nar/gkw387 (2016). [PubMed: 27141966]
57. Liao Y, Smyth GK & Shi W featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 30, 923–930, doi:10.1093/bioinformatics/btt656 (2014). [PubMed: 24227677]
58. Kao D et al. ERE database: a database of genomic maps and biological properties of endogenous retroviral elements in the C57BL/6J mouse genome. Genomics 100, 157–161, doi:10.1016/j.ygeno.2012.06.002 (2012). [PubMed: 22691267]
59. Young GR, Kassiotis G & Stoye JP Emv2, the only endogenous ecotropic murine leukemia virus of C57BL/6J mice. Retrovirology 9, 23, doi:10.1186/1742-4690-9-23 (2012). [PubMed: 22439680]

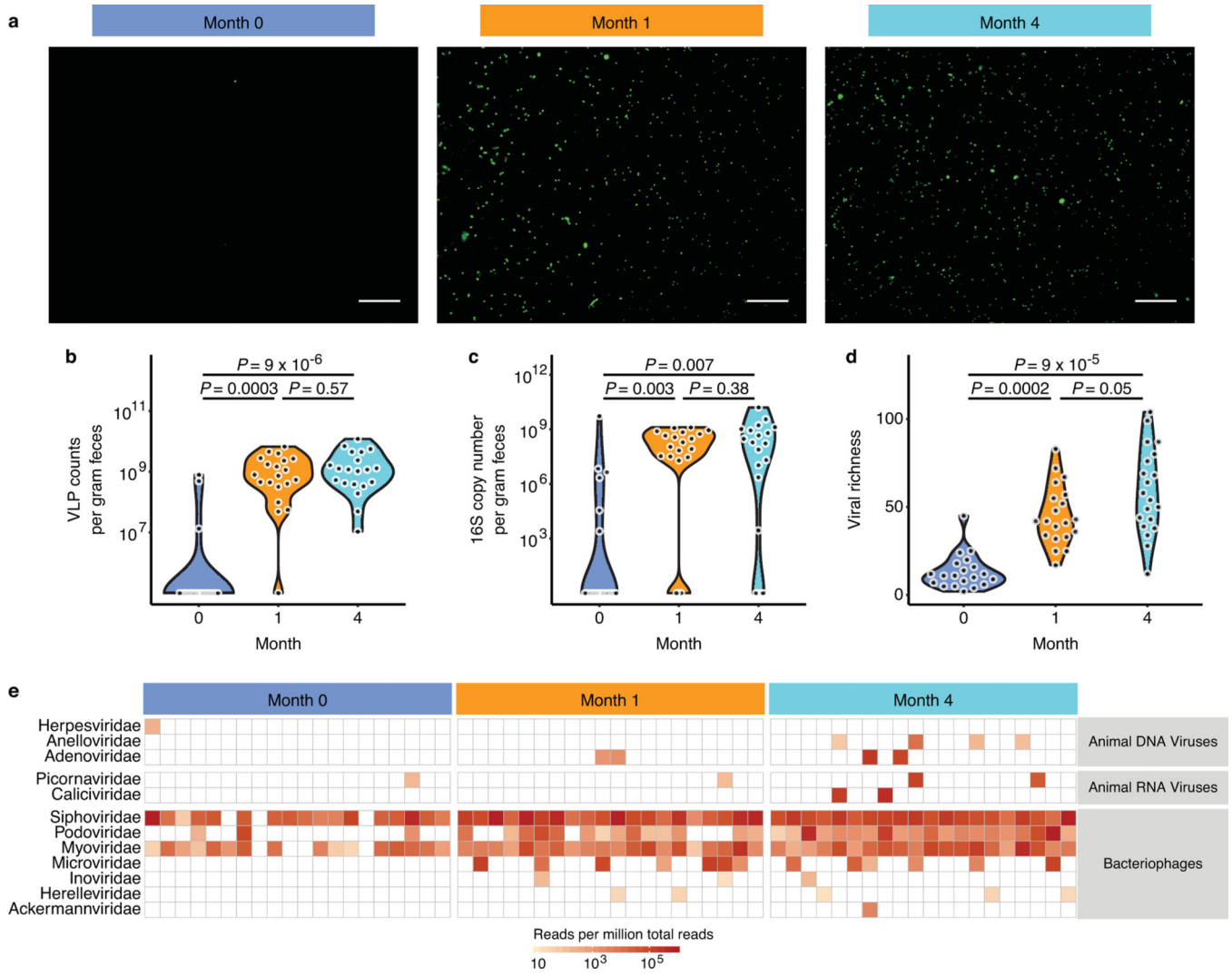


Fig. 1|.

Detection and characterization of virus-like particles (VLPs) in infant gut samples. **a**, Representative fields of fluorescently stained VLPs from infant stool sampled at month 0, 1, and 4. Scale bar = 10 μm. **b**, Quantification of VLP counts per gram. The minimum level of quantification was 6.6×10^6 particles per gram (5 to 10 fields quantified per sample). **c**, Copy numbers of bacterial 16S rRNA genes analyzed using qPCR. The minimum level of quantification was 2000 copies per gram. **d**, VLP richness assessed using VLP metagenomic sequence data. Sequences reads were assembled into contigs, and contigs with viral character (at leat 50% of open reading frames annotating as viral) enumerated. Viral species were called present if at least 10 reads per million from one sample aligned to that contig. **e**, Taxonomic assignments of VLP sequences. Reads were associated with viral lineages based on annotation of viral contigs. In **b**–**d**, violin plots represent the actual distribution of the individual data sets, and samples were compared using two-sided Wilcoxon signed-rank tests.

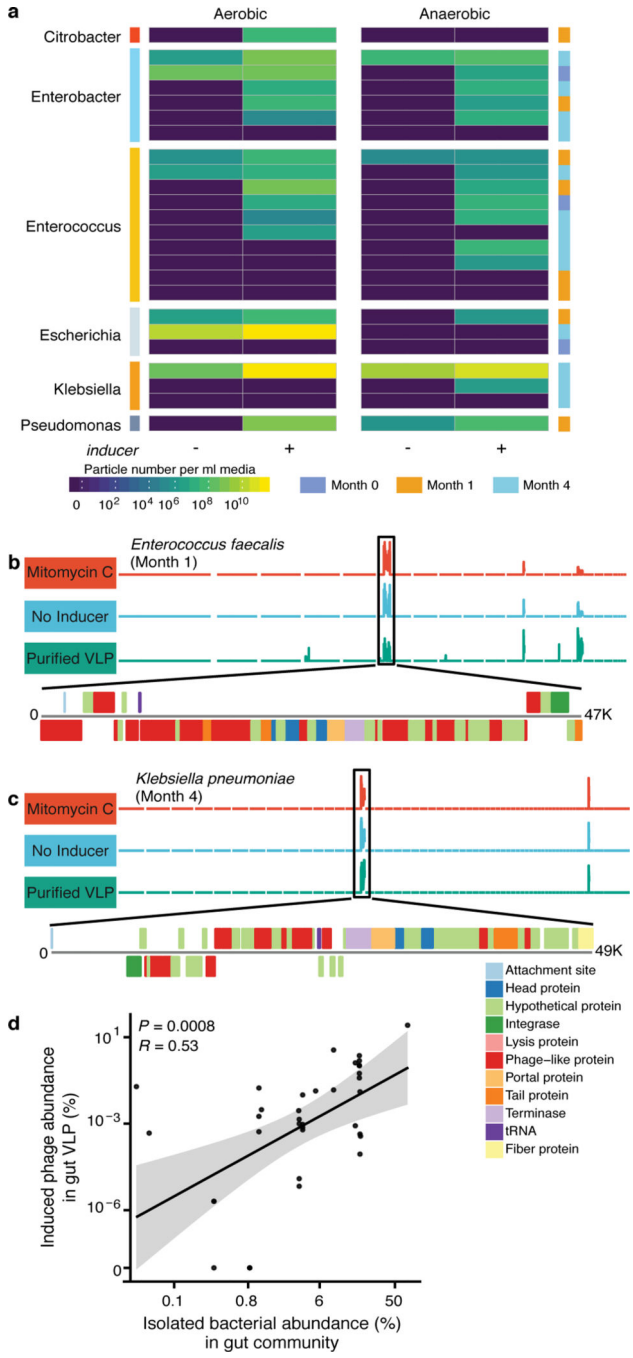


Fig. 2|.

Prophage induction as the dominant contributor to the early life virome. **a**, Heatmap quantifying VLP production from 24 strains isolated from feces of the infants studied. The bacterial genera are summarized on the left; columns summarize the numbers of fluorescent particles produced per ml of stationary phase culture (scale at bottom). Columns compare particle production with and without inducer (mitomycin C), and growth under aerobic and anaerobic conditions. **b**, Draft genome (horizontal line) from *Enterococcus faecalis* from one of the infants studied, showing alignment frequency of reads from VLP preparations. Reads

were aligned to the bacterial genome that were generated from VLPs from pure culture after mitomycin treatment (red), from VLPs from pure culture in the absence of any inducer (blue), and from VLPs isolate from stool of the the infant from which the bacterial strain was isolated (green). Peaks indicate detection of integrated prophages. One putative bacteriophage genome is shown below, with gene types color coded as indicated. **c**, As in **b**, but showing a *Klebsiella pneumoniae* isolate. **d**, Correlation between abundance of VLPs present in infant stool and the abundance of the bacteria harboring those prophages in the same stool sample (n = 33 phage contigs from 16 bacterial isolates from month 1 and month 4 srtrains). The black dashed line shows the linear regression line and the gray-shaded region shows the 95% confidence interval for the slope (two-sided Spearman's rank-order correlation).

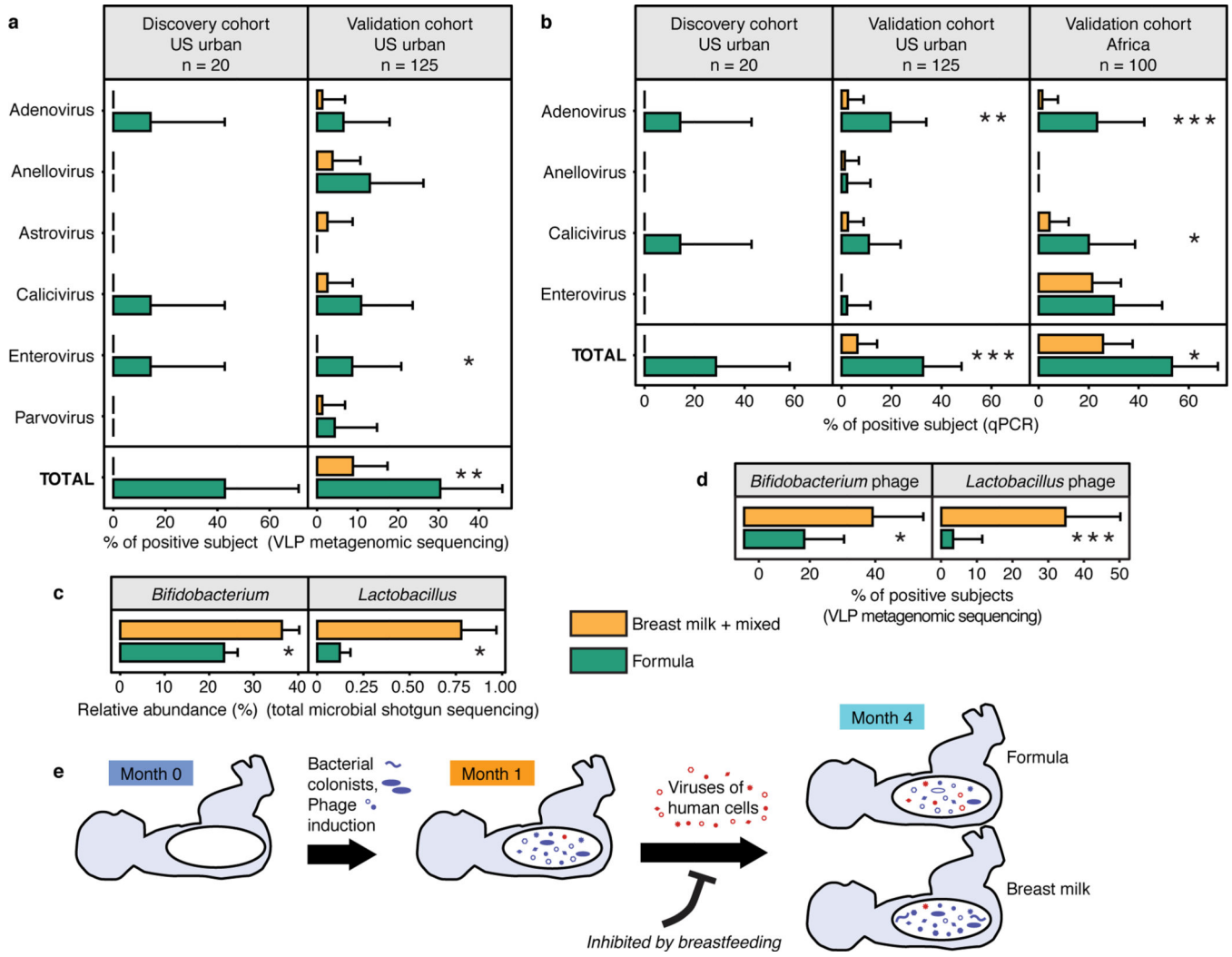


Fig. 3|.

Breastfeeding and viral colonization of the infant gut. **a**, Quantification of the percentage of subjects positive for viruses of human cells in metagenomic virome sequence data. Virus types are shown along the y-axis, percent of subjects positive is on the x-axis. Sample sizes and cohorts studied are indicated at the top. The two feeding types are color coded. Summation over all families is at the bottom. **b**, Comparison of human virus colonization based on feeding type using quantitative PCR. Three technical replicates were compared for each sample. In **a**, **b**, the numbers of infants with formula and breast milk or mixed are 14 and 6 in discovery cohort, 46 and 79 in validation cohort from US urban, and 30 and 70 in validation cohort from Africa. **c**, Abundances of the *Bifidobacterium* and *Lactobacillus* bacterial genera separated by feeding type. Samples were compared using two-sided Wilcoxon rank-sum tests with FDR correction. Bars represent mean ± s.e.m. **d**, Percentage of positive subjects with bacteriophages annotated as infecting *Bifidobacterium* or *Lactobacillus*. In **c**, **d**, 103 (Formula, n = 59; Breast milk or mixed, n = 44) samples from both discovery and validation cohorts were used for which whole stool shotgun sequence data was available. In **a**, **b**, **d**, samples were compared using two-sided Fisher's exact tests.

Error bars represent 95% confidence intervals. In **a**-**d**, ***$P < 0.001$, **$P < 0.01$, *$P < 0.05$. **e**, Summary of the findings in this study.