The gut of healthy human neonates is usually devoid of viruses at birth, but quickly becomes colonized, which—in some cases—leads to gastrointestinal disorders. Here we show that the assembly of the viral community in neonates takes place in distinct steps. Fluorescent staining of virus-like particles purified from infant meconium or early stool samples shows few or no particles, but by one month of life particle numbers increase to $10^9$ per gram, and these numbers seem to persist throughout life. We investigated the origin of these viral populations using shotgun metagenomic sequencing of virus-enriched preparations and whole microbial communities, followed by targeted microbiological analyses. Results indicate that, early after birth, pioneer bacteria colonize the infant gut and by one month prophages 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 who were exclusively fed on formula milk compared with those fed partially or fully on breast milk, paralleling reports that breast milk can be protective against viral infections. Bacteriophage populations also differed depending on whether or not the infant was breastfed. We show that the 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; this second phase is modulated by breastfeeding.

To investigate early-life viral colonization, we first analysed stool samples from 20 healthy infants longitudinally (Supplementary Table 1). Samples included meconium and/or early stool samples collected 0–4 days after birth (median of 17 h after birth, range 11–152 h; hereafter ‘month 0’) and stool samples collected at 1 and 4 months of life. The cohort consisted of self-identified African-American mothers from an urban US setting and their infants. As an initial step, virus-like particles (VLPs) were purified from meconium or stool and stained with SYBR gold, which binds nucleic acids. VLPs were subsequently visualized by epifluorescence microscopy (Fig. 1a). VLPs were undetectable in most of the meconium samples; only 3 out of 20 samples had detectable counts of VLPs (Fig. 1b). By month 1, most samples were positive, and VLP counts averaged $1.6 \times 10^4$ per gram of stool; values at month 4 were similar to the 1-month samples. We also tested the VLP counts from 12 2–5-year-old children, which had an average of $9.4 \times 10^4$ per gram of stool; these results were not distinguishable from month-1 and month-4 samples ($p = 0.48$, Wilcoxon rank-sum test). This number is also close to that reported for adults. Therefore we conclude that the high VLP counts seen in 1-month-old infants typically persist into adulthood.

To characterize bacterial content, DNA was purified from whole meconium or stool samples and analysed by qPCR to quantify the copy numbers of the bacterial 16S rRNA gene (Fig. 1c). Some published studies have suggested that microbial colonization of the infant begins in utero, but recent studies indicate that colonization more likely begins with the rupture of membranes and delivery. 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 6 (median, $3.3 \times 10^6$ per gram of stool). By contrast, for months 1 and 4, most samples were positive for 16S rRNA gene sequences, with a median value of $3.1 \times 10^7$ per gram of stool. To analyse 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 before colonization (Extended Data Fig. 1a). Some of the 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 $p = -0.45$; Extended Data Fig. 1b), consistent with bacterial...
Fig. 1 | Detection and characterization of VLPs in infant gut samples.

a, Representative fields of fluorescently stained VLPs from infant stool sampled at months 0, 1 and 4. Scale bars, 10 μm. b, Quantification of VLP counts per gram faeces. The minimum level of quantification was 6.6 × 10^9 particles per gram. Per sample, 5–10 fields were quantified. c, Copy numbers of bacterial 16S rRNA genes analysed using qPCR. The minimum level of quantification was 2,000 copies per gram. Per sample, 5–10 fields were quantified. d, VLP richness was assessed using VLP metagenomic sequence data. Sequence reads were assembled into contigs, and contigs with viral characteristics (at least 50% of open-reading frames annotated as viral) were enumerated. Viral species were called present if at least 10 reads per million reads from one sample aligned to that contig. e, Taxonomic assignments of VLP sequences. Reads were associated with viral lineages based on the annotation of viral contigs. In b–d, violin plots represent the distribution of the individual datasets; samples were compared using two-sided Wilcoxon signed-rank tests.

colonization. Bacterial DNA predominated in month-1 and month-4 samples (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.

To investigate the origin of the early life virome, DNA and RNA were purified from preparations of VLPs from stool of 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 bacteriophages (hereafter phages)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 that half of all reading frames within a contig be annotated as viral to assign that contig as a virus. Quantification of viral species richness showed low values at month 0, but higher richness after 1 and 4 months (Fig. 1d). After taxonomic assignment of viral contigs, we found that, despite the difficulty of identifying phages, the majority of VLP classifications were in fact phage families (Fig. 1e). Most were from DNA phages, consistent with the reported rarity of RNA phages20. For DNA phages, 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 reads probably represent unstudied phages. Values were lower at month 0 (11%), which probably reflects a relative increase in background due to the low numbers of recovered particles (Extended Data Fig. 2a–h).

To assess community interactions, we compared bacterial abundances from 16S rRNA gene qPCR data, bacterial richness and bacterial diversity from sequence data with VLP counts, viral richness and viral diversity, and found strong positive correlations (Extended Data Fig. 3a, b and 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 4, 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. Lytic phages only grow by infection, replication and lysis (Extended Data Fig. 4a). Previous reports, which focused on older infants and adults, have suggested that lytic growth and predator–prey interactions between phages and bacteria were prominent in early-life communities. Temperate phages can also replicate using a second strategy, which involves the 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 prophages to excise and resume lytic growth. Induction can also take place at low levels spontaneously. The prophage state is commonly maintained by repressor proteins, which also serve to exclude infection by similar or identical (homoimmune) phage strains.

To test whether the early-life virome is composed of strictly lytic phages, we purified 24 bacterial strains from the 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 that
lytic replication occurs. Note that temperate phages are not expected to form plaques on host cells that already contain those phages as an integrated prophage due to repressor-mediated homimmune exclusion. Analysis of our DNA VLP contigs using PHACTS, a random forest-based approach to classifying phage lifestyles, indicated that most of the genomes more closely resembled temperate phages than lytic phages (Extended Data Fig. 4b).

We next investigated whether virome populations resulted from the induction of prophages by quantifying VLP production from the 24 infant bacterial strains described above. Bacterial strains were analysed for spontaneous VLP production during growth in liquid culture, and after induction with the DNA-damaging agent mitomycin C, using the fluorescent staining assay for VLP particles. Experiments were carried out under both aerobic and anaerobic conditions. We detected spontaneous VLP production in 32% of strains. After induction with mitomycin C, 80% of strains produced VLPs (Fig. 2a). In total, 16 strains showed VLP production of at least $10^7$ particles per ml under at least one condition. We therefore conclude that bacteria in the infant gut 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 samples should be detectable as integrated prophages in bacterial genome sequences. We therefore sequenced genomes of the 24 infant bacterial strains described above and of 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 in the VLP samples from infant stool. Examples are shown in Fig. 2b, c, in which the VLP sequence reads are shown aligned to bacterial contigs, so that the spikes indicate VLP detection after mitomycin C induction, in the absence of induction and in purified stool VLPs. 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 before 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 the production of phages in the infant gut by prophage induction.

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

**Fig. 2** | Prophage induction as the dominant contributor to the early-life virome. a, Heat map quantifying VLP production from 24 strains isolated from the faeces 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 (according to the scale at the bottom). Columns compare particle production with and without an inducer (mitomycin C) as well as growth under aerobic and anaerobic conditions. b, Draft genome (horizontal line) of *Enterococcus faecalis* from one of the infants studied, showing the alignment frequency of reads from VLP preparations. Reads were aligned to the bacterial genomes 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 isolated from the stool of the infant from which the bacterial strain was isolated (green). Peaks indicate the detection of integrated prophages. One putative bacteriophage genome is shown below, with gene types colour-coded as indicated. c, As in b, but for a *Klebsiella pneumoniae* isolate. d, Correlation between the abundance of VLPs present in infant stool and the abundance of the bacteria harvesting those prophages in the same stool sample ($n = 33$ phage contigs from 16 bacterial isolates from month-1 and month-4 strains). The black line shows the linear regression line and the grey-shaded region shows the 95% confidence interval for the slope (two-sided Spearman’s rank-order correlation).
The abundant crAssphages, which infect Bacteroides and do not integrate during replication\textsuperscript{25,26}, were scarce in samples from month 1, but more common by month 4 and in samples from 2–5-year-old children (Extended Data Fig. 6). Evidently this group of lytic phages colonizes children predominantly later in life, potentially reflecting sequential acquisition of Bacteroides strains and later crAssphages from the environment.

Our findings support the idea that prophage induction from pioneer bacteria is the main source of the observed virome community by month 1. This is supported by the findings that: (1) replication of lytic phages 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 prophages in bacteria isolated from these infants; (4) stool VLP genome sequences could be identified as integrated prophages 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 were annotated primarily as lysogenic phages and not lytic phages.

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). We found a strong influence of breastfeeding, which was associated with the lower accumulation of viruses that replicate in human cells. Taking a conservative threshold for detection, requiring coverage of 33% of the viral genome, viruses that infect human cells were only found in those infants who were exclusively fed formula milk (Fig. 3a and Extended Data Fig. 7a). Statistically, this achieved a $P$ value of only 0.11 (Fisher's exact test) owing to the small sample size and unbalanced distribution (Fig. 3a and Extended Data Fig. 7b). Delivery type (spontaneous vaginal delivery versus caesarean section) did achieve statistical significance ($P=0.01$, Fisher's exact test; Extended Data Fig. 7f, g).

To challenge these conclusions, we analysed 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 US 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 that infect human cells, whereas 9% of babies who were fed breast milk or breast milk together with formula were positive ($P=0.003$, Fisher's exact test; Fig. 3a). Repeating the analysis requiring 0.1% coverage 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 viruses that infect animal cells (data not shown), indicating that these viruses were unlikely to originate as contamination from 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, Astroviridae, Sapovirus and Norovirus sequences (Supplementary Table 4). The qPCR analysis also showed enrichment of viruses that infect human cells in the exclusively formula-fed cohort ($P=0.0002$, Fisher's exact test; Fig. 3b).

**Fig. 3** Breastfeeding and viral colonization of the infant gut. **a**, Quantification of the percentage of infants who were positive for viruses of human cells in metagenomic virome sequence data. Sample sizes and cohorts studied are indicated at the top. The two feeding types are colour-coded as indicated. The summation over all viral families is shown at the bottom (total). **b**, Comparison of human virus colonization based on feeding type using qPCR. Three technical replicates were compared for each sample. **c**, Relative abundancies of Bifidobacterium and Lactobacillus. **d**, VLP abundance in stool was proportional to the abundance of the host bacteria in the same sample; and (6) VLP contigs were annotated primarily as lysogenic phages and not lytic phages.

**Table 1.** Metagenomic virome community composition of infants fed with formula and fed with breast milk or mixed feeding are, respectively, 14 and 6 in the discovery cohort, 46 and 79 in the validation cohort from US urban areas, and 30 and 70 in the validation cohort from Africa.

**Fig. 3** Breastfeeding and viral colonization of the infant gut. **a**, Quantification of the percentage of infants who were positive for viruses of human cells in metagenomic virome sequence data. Sample sizes and cohorts studied are indicated at the top. The two feeding types are colour-coded as indicated. The summation over all viral families is shown at the bottom (total). **b**, Comparison of human virus colonization based on feeding type using qPCR. Three technical replicates were compared for each sample. **c**, Relative abundancies of Bifidobacterium and Lactobacillus. **d**, VLP abundance in stool was proportional to the abundance of the host bacteria in the same sample; and (6) VLP contigs were annotated primarily as lysogenic phages and not lytic phages.

**Table 1.** Metagenomic virome community composition of infants fed with formula and fed with breast milk or mixed feeding are, respectively, 14 and 6 in the discovery cohort, 46 and 79 in the validation cohort from US urban areas, and 30 and 70 in the validation cohort from Africa.
Both populations studied above were from urban cohorts in the United States. To investigate whether our results hold more broadly, we analysed samples from a cohort of African newborns from Botswana. Infants who were 4 months of age were sampled using rectal swabs, so only qPCR assays and not sequence-based assays were carried out. All infants were delivered vaginally. We again found an association between exclusive formula feeding and colonization of viruses that grow in 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 the population structure of phages could be detected based on feeding mode ($P = 0.001$, permutation multivariate analysis of variance (PERMANOVA); Extended Data Fig. 8a, c), but not for another 12 metadata variables (Extended Data Fig. 8a–e). To investigate the taxa involved, the shotgun sequencing analysis of whole stool from pooled discovery and validation cohorts was queried, which showed that whole-stool samples from breastfed infants contained a higher abundance of Bifidobacterium (false-discovery rate (FDR)-corrected $P = 0.02$; Fig. 3c) and Lactobacillus (FDR-corrected $P = 0.03$; Fig. 3c). Consistent with host abundances, VLP sequences that aligned to the 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 on phage populations.

In summary, our data indicate that viral colonization in early life is stepwise, with the first phase characterized by the induction of prophages from pioneer bacteria, and a second phase involving colonization with viruses that infect 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\(^\text{19,20}\). Mixed feeding of formula and breast milk is also reported to be protective compared with feeding of formula only\(^\text{8}\), as was seen in the metagenomic analysis reported here. Factors in breast milk that are known to inhibit viral colonization include maternal antibodies, human milk oligosaccharides, lactoferrin and additional breast milk proteins\(^\text{25–28}\). The work reported here further develops our understanding of the protection conferred by breastfeeding in several respects. The metagenomic data (1) document the extent of subclinical infections with potentially pathogenic viruses; (2) highlight the potency of viral inhibition by breastfeeding; and (3) reveal the diversity of viruses affected, including viral families that cannot be grown in the laboratory and for which inhibition by breastfeeding is unstudied. In the African cohort, we not only found viruses that grow in 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 opportunities to intervene to reduce viral transmission to infants. The metagenomic methods described here will be useful for the assessment of the effects of different feeding strategies in diverse global settings to optimize the protection of infants from viral infections in the gut.

Online content
Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2192-1.

1. Breitbart, M. et al. Viral diversity and dynamics in an infant gut. Res. Microbiol. 159, 367–373 (2008).
2. Lim, E. S. et al. Early life dynamics of the human gut virome and bacterial microbiome in infants. Nat. Med. 21, 1228–1234 (2015).
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 (2016).
4. Oude Munnik, B. B. & van der Hoeck, L. Viruses causing gastroenteritis: the known, the new and those beyond. Viruses 8, 42 (2016).
5. Kim, M. S., Park, E. J., Roh, S. W. & Bae, J. W. Diversity and abundance of single-stranded DNA viruses in human feces. Appl. Environ. Microbiol. 77, 8062–8070 (2011).
6. Lepage, P. et al. Dysbiosis in inflammatory bowel disease: a role for bacteriophages? Gut 57, 424–425 (2008).
7. Hoyles, L. et al. Characterization of virus-like particles associated with the human faecal and caecal microbiota. Res. Microbiol. 165, 803–812 (2014).
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 (2005).
9. Arifeen, S. et al. Exclusive breastfeeding reduces acute respiratory infection and diarrhea deaths among infants in Dhaka slums. Pediatrics 108, e67 (2001).
10. Victora, C. G. et al. Infant feeding and deaths due to diarrhea. A case–control study. Am. J. Epidemiol. 129, 1032–1041 (1989).
11. Aagaard, K. et al. The placenta harbors a unique microbiome. Sci. Transl. Med. 6, 237ra65 (2014).
12. Lauder, A. P. et al. Comparison of placenta samples with contamination controls does not provide evidence for a distinct placenta microbiota. Microbiome 4, 29 (2016).
13. Thes, K. R. 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.E1–267.E139 (2019).
14. de Goffau, M. C. et al. Human placenta has no microbiome but can contain potential pathogens. Nature 572, 329–334 (2019).
15. Baumann-Dudenhoefler, A. M., D’Souza, A. W., Tarr, P. I., Warner, B. B. & Dantas, G. Infant diet and maternal gestational weight gain predict early metabolic maturation of gut microbiomes. Nat. Med. 24, 1822–1829 (2018).
16. Reyes, A. et al. Viruses in the faecal microbiota of monzygotic twins and their mothers. Nature 466, 334–338 (2010).
17. Minot, S., Grunberg, S., Wu, G. D., Lewis, J. D. & Bushman, F. D. Hypervariable loci in the human gut virome. Proc. Natl Acad. Sci. USA 109, 3962–3966 (2012).
18. Reyes, A., Semenkovich, N. P., Whiteson, K., Rohwer, F. & Gordon, J. I. Going viral: next-generation sequencing applied to phage populations in the human gut. Nat. Rev. Microbiol. 10, 607–617 (2012).
19. Aggarwala, V., Liang, G. & Bushman, F. D. Viral communities of the human gut: metagenomic analysis of composition and dynamics. Mob. DNA 8, 12 (2017).
20. Wolf, Y. I. et al. Origins and evolution of the global RNA virome. mBio 9, e02329-18 (2018).
21. Jacob, F., Sussman, R. & Monod, J. On the nature of the repressor ensuring the immunity of lysogenic bacteria (in French). C. R. Acad. Sci. 254, 317–319 (1962).
22. Plasme, M. A Genetic Switch (Blackwell, 1986).
23. Jacob, F. & Wollman, E. Spontaneous induction of the development of bacteriophage lambda during genetic recombination in Escherichia coli K12 (in French). C. R. Acad. Sci. 239, 317–319 (1954).
24. McNair, K., Bailey, B. A. & Edwards, R. A. PHACs, a computational approach to classifying the lifestyle of phages. Bioinformatics 28, 634–638 (2012).
25. Dutile, B. E. et al. A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. Nat. Commun. 5, 4498 (2014).
26. Shkoporov, A. N. et al. @CaAss001 represents the most abundant bacteriophage family in the human gut and infects Bacteroides intestinalis. Nat. Commun. 9, 4781 (2018).
27. Turn, C. G. & Ochoa, T. I. The role of maternal breast milk in preventing infantile diarrhoea in the developing world. Curr. Trop. Med. Rep. 1, 97–105 (2014).
28. Newburg, D. S., Ruiz-Palacios, G. M. & Morrow, A. L. Human milk glycans protect infants against enteric pathogens. Annu. Rev. Nutr. 25, 37–58 (2005).
29. Lewis, E. D., Richard, C., Larsen, B. M. & Field, C. J. The importance of human milk for immunity in preterm infants. Clin. Perinatol. 44, 23–47 (2017).

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

**Data reporting**
No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

**Experimental model and human participants**
Three cohorts of newborn infants were studied. Detailed information is provided in Supplementary Table 1. All experimentation complied with ethical regulations, and written informed consent was obtained from all human participants or their parents.

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 (IRB4-010833). 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 0–4 days after birth (meconium samples, month 0), month 1 and 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 healthy infants were collected from diapers and aliquoted into faeces collection tubes (Sarstedt). All samples were stored at −80 °C. Metadata regarding delivery mode, infant feeding and health outcomes were 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, receiving preventive healthcare 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 infants enrolled were less than 120 h of age, greater than 36 weeks gestation, heavier than 2,000 g and spent fewer than 120 h in the neonatal care unit. Mothers were over the age of 18 and spoke English. A total of 39 healthy, 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 infant 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 faecal 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. Maternal and baby clinical and metadata 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.

**Purification of VLPs from stool samples**
VLPs were purified as previously described30. 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 MgSO₄), spun down and filtered through a 0.2-μm-pore-size filter (Thermo Fisher Scientific). The filtrate was concentrated using a 100-kDa-molecular-mass Amicon Ultra-15 Centrifugal Filter (Millipore), resuspended in 30 ml SM buffer and concentrated for the second time to a final volume of around 500 μl. The concentrate was treated with DNase I and RNase (Roche) 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 for up to 3 months. To detect enveloped viruses, no chloroform was used to treat the VLPs. To test the purification efficiency of VLPs, 16S qPCR was used to quantify the 16S copy number before (total microbial DNA) and after purification of VLPs (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, approximately 90% of plaque-forming units could be recovered using the above methods.

**VLP enumeration**
Purified VLPs (35 μl) were diluted in 10 ml SM buffer and filtered onto a 0.02-μm Anodisc polycarbonate filter (Whatman). Filters were stained with 2× SYBR Gold (Thermo Fisher Scientific) for 15 min, then washed with H₂O once. After drying, the filter was mounted on a glass slide with 15 μl of mountant buffer (100 μl 10% ascorbic acid, 4.9 ml pH 7.4 PBS, 5 ml 100% glycerol; filtered through a 0.02-μm filter). For each filter, viruses were counted in 5–10 randomly selected fields of view. The filter was visualized using a motorized inverted system microscope IX81 (Shinjuku) for fluorescence. VLPs were counted using ImageJ. Stained particles of less than 0.5 μm in diameter were regarded as VLPs (larger particles were not counted). Purified lambda phages with known plaque-forming unit counts per ml were used as a positive control to adjust image colour, saturation, level and contrast. VLPs mock-purified from SM buffer were used as negative controls. At least one count per microscope field was set as a threshold for a positive detection, which was equal to around 6.6 × 10⁶ counts per gram faeces. Lower than this threshold, the VLP counts were considered to be below the limit of detection. The results are listed in Supplementary Table 2.

**Extraction and amplification of viral nucleic acids**
Viral DNA and RNA were extracted from VLPs using the AllPrep DNA/RNA Mini kit (Qiagen) 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). Viral RNA was treated with DNase, reverse transcribed and PCR-amplified as previously described30. 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 the SuperScript III First Strand Synthesis kit (Thermo Fisher Scientific) and primer A (5’-GTTTCCCAGTCACGATCN)-N-N-N-N-N-N-N-N-N-N-N-N-N-N-3’), to allow for random priming30. ‘N’ indicates a mixture of all four bases. Second-strand synthesis was performed using DNA Polymerase I, Large (Klenow) Fragment (New England Biolabs). 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 H₂O, 10 μl of 10× PCR buffer I, 4 μl of 50 mM MgCl₂, 2.5 μl 10 mM dNTPs, 5 μl of 10× PCR buffer II, 0.25 μl Taq High Fidelity DNA Polymerase (3 units), 0.05 μl of each primer, and 3 μl of cDNA. The cycling parameters were as follows: 95 °C for 3 min, then 40 cycles of 94 °C for 30 s, 60 °C for 30 s, 72 °C for 30 s, and 72 °C for 10 min.
Sigma-Aldrich was then added to a final concentration of 5 μg ml⁻¹. The overnight cultures of isolated bacterial strains were diluted 1/100 into In vitro prophage induction and induced VLP sequencing was as described above. Sequencing was performed using the Illumina Miseq (250-bp paired-end reads, Illumina). All 24 bacterial strains were cultured in Lysogeny broth overnight in anaerobic conditions, which were incubated at 37 °C for up to 72 h. Approximately 200 mg of stool was used for total microbial DNA extraction. Total microbial DNA was purified from each sample using the Mo Bio PowerSoil kit (Mo Bio) 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 the construction of the shotgun libraries. The DNA concentration was measured using the Quant-it PicoGreen dsDNA Assay kit (Thermo Fisher Scientific) and the fluorescence was detected by an EnVision Multilabel Plate Reader (Waltham). Libraries were made using an Illumina Nextera XT Samples Prep kit (Illumina), quantified using both the Quant-it PicoGreen dsDNA Assay kit and the KAPA Library Quantification kit (Kapa Biosystems). The size distribution of the libraries was checked by 5300 Fragment Analyzer (Agilent). Libraries were pooled for sequencing. The concentration of the pooled libraries was measured using Qubit (Invitrogen) 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 was isolated from the stool samples (19 samples from 12 infants) using three types of medium: Lysogeny broth medium in aerobic conditions, Bifidus selective medium (Sigma-Aldrich) and eosin methylene blue medium (Sigma-Aldrich) in anaerobic conditions, which were incubated at 37 °C for up to 72 h. Single colonies were picked and re-streaked in medium 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 a MALDI–TOF BD instrument (BD) 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 Lysogeny broth overnight or until an optical density at 600 nm (OD₆₀₀) > 1. DNA was extracted using the phenol–chloroform method. DNA quality control was as described above. The TruSeq DNA PCR-Free kit (Illumina) was used to make genomic DNA-sequencing libraries; quality control and pooling was as described above. Sequencing was performed using the Illumina Miseq (250-bp paired-end reads, Illumina).

In vitro phage induction and induced VLP sequencing Overnight cultures of isolated bacterial strains were diluted 1/100 into 10 ml medium and grown until log phase (OD₆₀₀ = 0.6). Mitomycin C (Sigma-Aldrich) was then added to a final concentration of 5 μg ml⁻¹. The OD₆₀₀ values were measured and VLPs were purified after 6 h of culture. VLPs were purified from the bacterial culture using the same method as described for 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 phages are generally not thought to form prophages; RNA was interrogated for three samples, and no phages were identified.

In vitro phage infection Overnight cultures of isolated bacterial strains were diluted 1/100 into 10 ml medium, grown to log phase (OD₆₀₀ = 0.6), and then 100 μl bacteria was mixed with 100 μl of serial dilutions of isolated infant stool VLPs with the addition of MgSO₄ (with a final concentration of 10 mM). 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 Lysogeny broth 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). Primer, probe sequences and the PCR program were described previously and are included 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 around 2,000 copies per gram faeces.

Stool VLP sequence read quality control and taxonomic classification Quality control for the stool VLP reads was performed using the Sunbeam pipeline with a custom Sunbeam extension (https://github.com/guaxianxiangli/sbx_dedu). In brief, low-quality reads and adaptor sequences were removed by Trimmomatic, 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 that included all complete human, bacterial, archaeal and viral genomes in RefSeq release 89 (released on 9 July 2018), with low-complexity regions masked before building the database.

To investigate environmental contamination or experimental reagent contamination, negative control samples were analysed, 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. Taxa including Klebsiella phage 0507-KN21, Choristoneura occidentalis granulovirus, Vibrio phage pYD38-A, Pseudomonas phage PpPW-4, Burkholderia phage ST79, Burkholderia phage KS9, Bacillus virus phiZ9, Simbu orthobunyavirus, and Shamona 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 for each individual. To quantify contigs in each sample, quality-controlled reads were mapped back to the contigs using Bowtie2, and the number of mapped reads was 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 3,000 bp were selected to predict open-reading frames (ORFs)
using Prodigal in 'meta' mode\textsuperscript{38}. The predicted ORFs were mapped to the viral protein database in UniProt Knowledgebase (TrEMBL and Swiss-Prot)\textsuperscript{39} using BLASTP with \(E < 10^{-3}\).

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 10 kb of VLP contig and (2) 50% of the predicted ORFs were viral ORFs. The taxonomy of each contig was classified as described previously\textsuperscript{40} modified to compile attributions over multiple reading frames to generate a single taxonomic assignment. The ORFs were assigned to taxa based on the best-hit viral protein in the UniProt Knowledgebase. The majority taxonomic assignment over all ORFs within a contig was given to the contig. Contigs that could not be assigned to any taxa were classified as ‘Others’. Contigs that were not assigned as ‘Bacteriophage’ were mapped to the NCBI nt database with a threshold of 80% coverage and 80% identity to further remove contigs from non-viral genomes. In total, we identified 2,552 viral contigs among all 20 infants (Extended Data Fig. 2a, b). Contigs that shared 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 of all nonhuman reads (Extended Data Fig. 2c–e). Other reads come from contamination, genomes of other microorganisms and unassigned categories. We think that some of the unassigned reads represent unstudied bacteriophages, for which insufficient ORFs matched the viral ORFs in the database to label the contig as viral. For the RNA virome data, assembly from 12 out of 20 infants yielded contigs larger than 3,000 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 all nonhuman reads (Extended Data Fig. 2f–h).

Viral contigs were scored as temperate or lytic bacteriophages using PHACTS\textsuperscript{24}. In order to obtain strong predictions, only viral contigs with at least 10 ORFs were analysed. Of 2,552 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 crAssphages in the infant gut, we mapped the stool VLP reads to 37 genomes which belong to the crAssphage family\textsuperscript{26,27,41,42}. 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, \(n = 21\); 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 Neighbour 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\textsuperscript{27}. The output sam files were process by Samtools\textsuperscript{43}, Bedtools\textsuperscript{44} and custom code (https://github.com/guanxiangfang/liang2019) to quantify the fraction of the genome covered. We favour use of percent coverage as a metric for genome detection\textsuperscript{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\textsuperscript{46}, Human Torque teno viruses\textsuperscript{47}, Enteroviruses\textsuperscript{48}, Astroviruses\textsuperscript{49}, Sappovirus GI strains\textsuperscript{50} and Norovirus GII strains\textsuperscript{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 standard 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 comparison between virome sequencing and qPCR data are presented in 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\textsuperscript{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\textsuperscript{16}.

### 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\textsuperscript{53}, followed by SSPACE to make scaffolds\textsuperscript{34}. The quality of scaffolds (completeness and contamination) was evaluated by CheckM\textsuperscript{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 analyse whether stool viruses from the stool VLP sequences matched sequences of induced VLPs from purified infant gut bacterial strains, reads from induced VLPs and stool VLPs were mapped to the corresponding bacterial scaffolds using Bowtie2\textsuperscript{12}. The number of mapped reads and bedgraph files for coverage plots were generated using Samtools\textsuperscript{43}, Deeptools\textsuperscript{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 infants from whom the bacterial strain originated than in VLPs from unmatched individuals. The prophage genome annotation was performed by PHASTER\textsuperscript{44} targeting bacterial genomic scaffolds longer than 100,000.

VLP contigs for analysis were identified as follows. First, we asked whether assembled contigs from the induced VLPs of the 24 purified bacterial strains could be identified in the 24 sequenced bacterial genomes. We required that more than 50% of the induced VLP contig length was matched to a bacterial genome scaffold or vice versa (Blastn with \(E < 10^{-10}\)). Second, we asked whether induced VLP contigs recognized as candidate prophages 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 < 10^-5). Third, for 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 from which the bacterial strain originated than 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 bacterial species in the infant gut community and the proportion of prophages from that bacterial species in the infant gut virome.

Stool VLP sequences were mapped to the induced VLP contigs identified above using Bowtie2. The proportion of mapped reads from stool VLPs were divided by the total number of stool VLP reads 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 could be mapped to the isolated bacterial genome divided by all nonhuman reads. The abundance of bacterial prophages 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 analyze the 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 3,000 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 the Pfam database using HmmScan (HMMER 3.1; http://hmmner.org/) with E < 10^-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 to evaluate the abundance of each Pfam entry. The Pfam annotations for each sample were catalogued and a matrix was generated for annotation over all samples. Clustering was evaluated using Bray–Curtis dissimilarities. Bray–Curtis dissimilarities were plotted using principal coordinate analysis (PCoA), and differences among groups (infant age, infant feeding type, infant delivery type, infant gender, mother body type, formula type, mother pregnancy induction, hypertension or diabetes and mother Chorioamnionitis) were tested using PERMANOVA. 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 EnviFit function. P values were determined using 999 permutations. The analysis was carried out using the vegan R package.

Bifidobacterium and Lactobacillus phage analysis
We downloaded 42 Lactobacillus phage genomes from RefSeq and used these for comparison. RefSeq did not contain any Bifidobacterium phage genome sequences, but two Bifidobacterium phage genomes were available in NCBI (accession numbers 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 contain annotated ‘Integrase’ proteins, suggesting temperate replication cycles.

Quantification and statistical analysis
Statistical tests were conducted using R. Nonparametric tests were used to compare two independent groups (Wilcoxon rank-sum test), two related groups (Wilcoxon signed-rank test) and multiple groups (Kruskal–Wallis test with Bonferroni correction). Nonparametric correlation was performed using Spearman’s rank-order correlation (Rrepresents Spearman’s p). 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 FDR method. P < 0.05 or FDR-corrected P < 0.05 was considered significant. All reported P values are from two-sided comparisons. All acquired data were included in analyses.

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

Possible contribution of human endogeneous retroviruses
Human endogeneous retroviruses (HERVs) 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 that HERV DNA were probably contributed by contaminating human DNA, and were present in proportions predicted given the frequencies of other human genomic repeated sequences.

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

Data 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 numbers WVTFO0000000--WVUC00000000 (Supplementary Table 3).

Code availability
All bioinformatic scripts are available on Github (https://github.com/guaxiangliang/liang2019).

References
30. Chehood, C. et al. Transfer of viral communities between human individuals during fecal microbiota transplantation. mBio 7, e00322-16 (2016).
31. Wang, D. et al. Viral discovery and sequence recovery using DNA microarrays. PLoS Biol. 1, e2 (2003).
32. Hill, D. A. 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 (2010).
33. Clarke, E. L. et al. Sunbeam: an extensible pipeline for analyzing metagenomic sequencing experiments. Microbiome 7, 46 (2019).
34. Boilger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114–2120 (2014).
35. Davis, N. M., Proctor, D. M., Holmes, S. P., Relman, D. A. & Callahan, B. J. Simple statistical identification and removal of contaminant sequences in marker gene and metagenomics data. Microbiome 6, 226 (2018).
36. Li, D., Liu, C. M., Luo, R., Sadakane, K. & Larr, T. MEGAHI: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics 31, 1674–1676 (2015).
37. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359 (2012).
38. Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11, 119 (2010).
39. Li, D., Liu, C. M., Luo, R., Sadakane, K. & Larr, T. MEGAHI: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics 31, 1674–1676 (2015).
40. Wang, D. et al. Viral discovery and sequence recovery using DNA microarrays. PLoS Biol. 1, e2 (2003).
41. Davis, N. M., Proctor, D. M., Holmes, S. P., Relman, D. A. & Callahan, B. J. Simple statistical identification and removal of contaminant sequences in marker gene and metagenomics data. Microbiome 6, 226 (2018).
42. Li, D., Liu, C. M., Luo, R., Sadakane, K. & Larr, T. MEGAHI: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics 31, 1674–1676 (2015).
43. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359 (2012).
44. Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11, 119 (2010).
39. Pandir, S., Magrane, M., Martin, M. J., O’Donovan, C. & The UniProt Consortium. Searching and navigating UniProt databases. Curr. Protoc. Bioinformatics 50, 1.27.1-1.27.10 (2015).
40. Minot, S. et al. Rapid evolution of the human gut virome. Proc. Natl Acad. Sci. USA 10, 12450–12455 (2013).
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 (2018).
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 (2018).
43. Li, H. et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079 (2009).
44. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842 (2010).
45. Abbas, A. A. et al. Bidirectional transfer of Anelloviridae lineages between graft and host during lung transplantation. Am. J. Transplant. 19, 1086–1097 (2019).
46. Otth, K. 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 (2005).
47. Abbas, A. A. et al. The peripheroperative 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 (2017).
48. Verstrepen, W. A., Kuhn, S., Kockx, M. M., Van De Vyvere, M. E. & Mertens, A. H. 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 (2001).
49. van Maarseveen, N. M., Wessels, E., de Brouwer, C. S., Vossen, A. C. & Claas, E. C. 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 (2010).
50. Oka, T. et al. Detection of human sapovirus by real-time reverse transcription-polymerase chain reaction. J. Med. Virol. 78, 1347–1353 (2006).
51. Rolle, K. J. et al. An internally controlled, one-step, real-time RT-PCR assay for norovirus detection and genotyping. J. Clin. Virol. 39, 318–321 (2007).
52. Truong, D. T. et al. MetaPHAN2 for enhanced metagenomic taxonomic profiling. Nat. Methods 12, 902–903 (2015).
53. Bankevich, A. et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. 19, 455–477 (2012).
54. Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D. & Pirovano, W. Scaffolding pre-assembled contigs using SSPACE. Bioinformatics 27, 578–579 (2011).
55. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 25, 1043–1055 (2015).
56. Arndt, D. et al. PHASTER: a better, faster version of the PHAST phage search tool. Nucleic Acids Res. 44, W6–W21 (2016).
57. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 30, 923–930 (2014).
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 (2012).
59. Young, G. R., Kassiotis, G. & Stoye, J. P. Emv2, the only endogenous ecotropic murine leukemia virus of C57BL/6J mice. Retrovirology 9, 23 (2012).

Acknowledgements We thank J. You and co-workers for help with imaging, M. Goulian and co-workers for bacterial strains, and F. Rohwer and associates for assistance with phage induction, members of the Bushman laboratory for help and suggestions, and L. Zimmerman for artwork and help with the manuscript. The Botswana Infant Microbiome Study team thank Copan Italia for their donation of the eNAT medium and flocked swabs used for the collection of rectal swab specimens. This work was supported by NIH grants R51-HL137063 (F.D.B.), R01-HL113252 (F.D.B.) and R01DK107565 (G.D.W. and E.F.). The project described was also supported by the Penn Center for AIDS Research (P30 AI 045008) (F.D.B.), the PennCHOP Microbiome Program (F.D.B., G.D.W. and R.N.B.) and a Tobacco Formula grant under the Commonwealth Universal Research Enhancement (CURE) program (grant number SAP 4100068710) (F.D.B. and R.N.B.). 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 (C.Z., K.B., E.F., J.S.G. and B.S.Z.). The project was also supported by the National Center for Advancing Translational Sciences, National Institutes of Health, through grants UL1TR000003 and UL1TR001878 (E.F.). Funding for the Botswana Infant Microbiome Study was provided by the Duke Center for AIDS Research (SF30 AI064518) and the NIH (K23 AI113509).

Author contributions G.L. carried out biochemical analysis, sequencing and bioinformatic analysis; L.R.K., H.Z. and L.M. assisted with biochemical manipulations; C.Z., S.S.-M. and K.B. assisted with imaging; L.R.K., H.Z. and L.M. assisted with biochemical manipulations; C.Z., S.S.-M. and K.B. assisted with imaging; F.D.B. and R.N.B. assisted with bioinformatic analysis; G.D.W. and E.F. assisted with bioinformatic analysis; G.D.W., R.N.B., P.D., E.F., M.A.E., J.S.G., A.K., B.S.Z., M.S.K., M.Z.P. and T.M. carried out sample collection; G.L. and F.D.B. conceived the project and wrote the paper.

Competing interests The authors declare no competing interests.

Additional information Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2192-1. Correspondence and requests for materials should be addressed to F.D.B.

Reprints and permissions information is available at http://www.nature.com/reprints.
Extended Data Fig. 1 | Overview of total stool microbial shotgun metagenomic sequencing. **a**, Percentage of reads mapped to human or microbial genomes or that were unassigned. The types of DNA detected are indicated on the right. **b**, Correlation between the percentage of human DNA and 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 grey-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 $\rho$). **c**, Taxonomic composition of bacteria at the phylum level. The total read number is shown on the y axis; the x axis shows different samples. **d**, Bacterial richness. The y axis shows the richness calculated as the number of observed species. **e**, Bacterial diversity. A two-sided Wilcoxon rank-sum test was used to test the difference between different age groups ($n = 20$ infants at three time points). The horizontal lines in the box plots represent the third quartile, median and first quartile; whiskers extend to $\pm 1.5 \times$ the interquartile range. The dots represent the outliers.
Extended Data Fig. 2 | Summary of virome sequencing of infant stool. a, Heat map summarizing the representation of the top five most-abundant DNA viral contigs in each sample. Samples are grouped sequentially by infant on both the x-axis and y-axis. The last group of infants on the x-axis are negative control samples. Circularity indicates whether a contig is circular (orange color) or not (light-green color). The heat map color represents the abundance (log-transformed reads per million total reads) of each contig in each sample. b, Contig read abundance compared between different infants and within the same individuals. Time points were pooled for each individual. c–e, Percentage of DNA virome reads assigned to viruses (c), unassigned (d) and contamination (e). f–h, Percentage of RNA virome reads assigned to viruses (f), unassigned (g) and contamination (h). b–h, n = 20 infants at three time points were tested. The horizontal lines in box plots represent the third quartile, median and first quartile; whiskers extend to ±1.5× the interquartile range. The dots represent the outliers.
Extended Data Fig. 3 | Correlation between viral and bacterial communities. 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 colour indicates a positive correlation, and red colour indicates a negative correlation. Samples from different time points were pooled ($n = 60$). A 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 $\rho$) values are shown.
Extended Data Fig. 4 | Life cycles of bacteriophages. a, Diagram of lytic and lysogenic bacteriophage replication (based on a previous study[22]). Not shown are additional phage replication strategies, such as chronic infection and pseudolysogeny. b, Prediction of replication modes from contig sequences using PHACTS. The x axis shows the probability that a contig belongs to a lytic or temperate phage predicted by PHACTS. The y axis shows the viral contig number. In total, 1,029 phage contigs with at least 10 open-reading frames were used in this analysis. Of 1,029 contigs, 233 were predicted to be lytic and 794 were predicted to be temperate. Probability values obtained from PHACTS were standardized between −1 and 1, which was presented as a probability to be 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 with VLP sequences from stool samples. Contigs were generated from mitomycin-C-induced VLPs from purified bacterial strains from stool (n = 33 phage contigs from 16 bacterial isolates), then VLP reads from faeces were aligned to these contigs and quantified. ‘Within infants’ indicates matching stool VLPs to induced VLPs from purified bacteria for samples all from the same infant. ‘Between infants’ indicates alignment of stool VLPs versus induced VLPs from different infants. The horizontal lines in box plots represent the third quartile, median and first quartile. The dots represent the outliers. Samples were compared using a two-sided Wilcoxon rank-sum test.

**b.** Correlation between the proportion of each bacterium in the infant gut community and the proportion of prophages from that bacterial species 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 grey-shaded region shows the 95% confidence interval for the slope. The correlation was tested using a two-sided Spearman’s rank-order correlation (R represents Spearman’s ρ).
Extended Data Fig. 6 | Colonization by crAssphages in different age groups. The percentage of crAssphage-positive infants (as scored by requiring that the crAssphage genome was more than 33% covered by sequence reads from stool VLPs).
Extended Data Fig. 7 | See next page for caption.
Extended Data Fig. 7 | Profiling of animal-cell viruses by virome sequencing. a, c, f, h. Percentage of infants positive for animal cell-associated viruses using different viral genome coverage cut-offs in the discovery cohort (a, f) and validation cohort (c, h). The green line shows the data from infants who were formula fed (a, c) or born by caesarean (C)-section delivery (f, h), and the yellow line shows the data from infants fed with breast milk or who were mixed fed (a, c) or were born by spontaneous vaginal delivery (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 cut-offs in the discovery cohort (b, g) and validation cohort (d, i). The horizontal red line indicates \( P = 0.05 \). e, j. Comparison of the 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 the y axis. A two-sided Wilcoxon rank-sum test was used to test the difference. The horizontal lines in box plots represent the third quartile, median and first quartile; whiskers extend to \( \pm 1.5 \times \) the interquartile range. The dots represent the outliers. k. Genome coverage fraction of negative control samples for animal-cell viruses. The maximal fraction of animal viral genome coverage for each negative control sample \( n = 25 \) is shown on the y axis. Different negative control samples are shown on the x axis. Note that coverage never exceeds 10%. a, b, f, g, n = 20 samples from the discovery cohort were used; c–e, h–j, n = 125 samples from the validation cohort were used.
Extended Data Fig. 8 | Phage population structure. a, Statistical tests 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 columns. All categorized variables, such as infant age, infant feeding type, infant delivery type, infant gender, mother body type, formula type, mother pregnancy induced hypertension 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, PCoA plot based on phage Pfam counts per sample, coloured 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 tests. c–e, PCoA plots of phage Pfam components, coloured 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 as in a, PERMANOVA was used to test the differences. FDR-corrected P values are shown.
Extended Data Fig. 9 | 16S qPCR before and after VLP purification. Red and light-blue dots show before and after separately, and the horizontal lines represent the means (n=20 infants at three time points were tested). A two-sided Wilcoxon signed-rank test was used to test the difference.
Extended Data Fig. 10 | Percentage of DNA aligning to sequences of HERVs in each sample. The percentage of HERV sequences in stool VLPs 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 long interspersed nuclear elements and short interspersed nuclear elements, indicating that they are derived from human DNA contamination. Data are mean ± s.e.m.; n = 20 infants at three time points were tested.
Reporting Summary

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

Statistics

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

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

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

Software and code

Policy information about availability of computer code

| Data collection | Commercial softwares were not used for data collection. |
| Data analysis | qPCR was conducted and analyzed on a 7500 Fast Real Time qPCR system. Raw sequence data was analyzed in Linux environment (Ubuntu 16.04.6 LTS). Softwares under Linux system included Sunbeam (v2.0), Trimomatic (v0.36), Complexity (v0.3.0), BBmap (v38.22), Bowtie2 (v2.2.6), megahit (v1.1.3), prodigal (v2.6.3), Sambtools (v1.7), HMMER (v3.1), Redtools (v2.25), Blastp (v2.2.29), Blast (v2.2.29), SPADES (v3.12.0), SPSPACE (v2.0), and checkM (v1.0.12). RStudio (v1.1.423) was used for downstream and statistical analysis. Softwares used in R studio included R (v3.6.4), MASS (v7.3-51), ggplot2 (v3.3.1), viridis (v0.5.1), viridisLite (v0.3.0), ggplot (v2.2.3), vegan (v2.5-5), lattice (v0.20-35), permute (v0.9-5), taxonomizr (v0.5.3), data.table (v1.12.2), qlimr (v0.9.4), randomcolor (v1.1.0), RColorBrewer (v1.1-2), pheatmap (v1.0.12), tidygr (v0.8.3), reshape (v0.8.8), stringr (v1.4.0), dplyr (v0.8.3), plyr (v1.8.4), scales (v1.0.0), ggbeeswarm (v0.6.0), ggplot2 (v3.3.1), phyloseq (v1.22.3), MALDI Biotyper Realtime Classification and Biotyper software (v3.0) was used for MALDI-TOF 80 instrument. |

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

Data

Policy information about availability of data

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

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

Sample information and raw sequences are available in the National Center for Biotechnology Information Sequence Read Archive under BioProject ID [insert ID].
Field-specific reporting

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

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

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

Life sciences study design

All studies must disclose on these points even when the disclosure is not.

Sample size

No initial sample size was calculated for our first cohort (20 infants) due to the exploratory nature of the study. We then sought to validate findings on the protective effect of breastfeeding, which did not achieve significance in the small initial cohort. We acquired a cohort that we calculated would likely provide sufficient resolution to yield a significant result given similar effect sizes. Assuming the breast milk-exposed group had a proportion of 10% virus positive samples and our formula-only group had 40% virus positive samples (a conservative estimate from our discovery cohort), our collection of 78 breast milk-fed samples and 46 formula fed samples would provide a 96% chance of detecting a significant difference at \( p = 0.05 \). For our third cohort (African samples), we followed the power analysis used in assessing our second cohort.

Data exclusions

No available data was excluded.

Replication

For the analysis of the effects of breastfeeding, three separate cohorts were studied. Protective effect of breastfeeding was found in the discovery cohort, which did not achieve significance due to the small sample size. Two larger validation cohorts showed significant protective effects of breastfeeding.

Randomization

The observational cohort was segregated into formula fed and breast milk-exposed infants. Subjects were chosen to provide as balanced representation of the groups as possible given the cohort compositions. Subjects were subsampled randomly within each group where possible.

Blinding

Blinding was not possible in the context of analyzing breastfeeding.

Reporting for specific materials, systems and methods

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

Materials & experimental systems

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

Methods

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

Animals and other organisms

Policy information about studies involving animals: ARRIVE guidelines recommended for reporting animal research

Laboratory animals: Fecal samples were obtained as discarded animal by-products from a centralized germ-free mouse facility (Protocol 805449; end date 8/28/2020). Species Mus musculus, strain C57BL/6 strain, male and female, 10 weeks of age.

Wild animals: No wild animals were used in this study.

Field-collected samples: This study did not involve samples collected from the field.

Ethics oversight: No ethical approval or guidance was required. Fecal samples were obtained as discarded animal by-products from a centralized germ-free mouse facility (Protocol 805449; end date 8/28/2020) in the University of Pennsylvania.

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

Policy information about studies involving human research participants

Population characteristics

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. 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). 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 healthy, 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. 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-HPOM 13/6/1) and Institutional Review Boards at the University of Pennsylvania (IRB 827692) 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 Caesarean 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. Metadata, including data regarding infant feeding practices, were collected by medical chart review and in-person interview by trained research personnel.

Recruitment

The Infant Growth and Microbiome Study (iGram) recruited African-infant born at the Hospital of the University of Pennsylvania. Study visits were conducted at The Children's Hospital of Philadelphia. The Microbiome, Antibiotic, and Growth Infant Cohort (MAGIC) 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.

The cohorts above are US/urban cohorts which may not perfectly reflect acquisition of viruses in other settings. Therefore, we included another cohort from developing country. The Botswana Infant Microbiome Study recruited mother-infant pairs 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 Caesarean 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.

Ethics oversight

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). 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 Botswana infant Microbiome Study was approved by the Botswana Ministry of Health (IRB HPOM 13/6/1) and Institutional Review Boards at the University of Pennsylvania (IRB 827692) and Duke University (IRB 319561).

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