

A meta-analysis of infant gut viromes reveals global patterns in bacteriophage community assembly and functional capacity over the first three years of life

Received: 3 March 2026

Accepted: 8 June 2026

Cite this article as: Shamash, M., Maurice, C.F. A meta-analysis of infant gut viromes reveals global patterns in bacteriophage community assembly and functional capacity over the first three years of life. *Nat Commun* (2026). <https://doi.org/10.1038/s41467-026-74609-5>

Michael Shamash & Corinne F. Maurice

We are providing an unedited version of this manuscript to give early access to its findings. Before final publication, the manuscript will undergo further editing. Please note there may be errors present which affect the content, and all legal disclaimers apply.

If this paper is publishing under a Transparent Peer Review model then Peer Review reports will publish with the final article.

A meta-analysis of infant gut viromes reveals global patterns in bacteriophage community assembly and functional capacity over the first three years of life

Michael Shamash ¹, Corinne F. Maurice ^{1,2 *}

¹ Department of Microbiology & Immunology, McGill University, Montreal, QC, Canada

² McGill Centre for Microbiome Research, Montreal, QC, Canada

* Corresponding author: corinne.maurice@mcgill.ca

Abstract

Assembly of the gut microbiota in infancy is an important health determinant, yet the viral component, the virome, remains poorly described. Here, we conduct a meta-analysis of 12 infant cohorts across 8 countries to model gut virome development over the first three years of life. Our results revealed distinct diversity patterns, where viral richness increased significantly over the first eight months of life alongside a loss of community evenness. We define virome developmental velocity, the rate of change between sequential samples, which significantly decreased over time. This was mirrored by community-wide convergence, as infant viromes became more similar with age. Functionally, the transition to a stable state involved a significant decrease in extracellular temperate phages and shifts in phage-encoded auxiliary metabolic genes, which can modulate host metabolism. Our work provides a baseline characterization of gut virome development, marked by a conserved successional trajectory, providing a framework to identify disease-associated perturbations.

Introduction

The human gut microbiota begins its development at birth which continues throughout the first few years of life ^{1,2}. Many studies have focused on the gut bacteriome and its associations with infant health ³⁻⁶. While several studies have begun to shine light on the early life gut virome, primarily composed of bacteria-infecting viruses known as bacteriophages (phages), much remains unknown about these highly abundant members of the gut microbial ecosystem ⁷⁻¹⁰. Current evidence suggests that this viral community is not randomly acquired but instead linked to the arrival of the first bacterial hosts. The initial phages in the gut are induced prophages from pioneer gut bacteria and remain as prominent members of the gut virome for several months ¹¹. While bacterial colonizers follow a predictable successional trajectory ¹², it remains unclear whether assembly of the gut virome follows a similar structured assembly.

The gut nutritional landscape undergoes important shifts during infancy, primarily driven by the transition from a diet rich in human milk oligosaccharides (HMOs) to one which is complex and contains plant-derived polysaccharides ¹³. While the gut bacteriome adapts to these shifts through changes to its structure and composition, the role of the virome in supporting this transition remains unexplored. Indeed, phages can carry auxiliary metabolic genes (AMGs) which modulate host metabolism ¹⁴. These genes are best characterized in aquatic and soil ecosystems, where phages can carry photosynthetic genes ¹⁵; however, studies have also identified enrichment in nitrogen and carbon utilization genes among phages in the adult gut ¹⁶⁻¹⁸. AMGs found in the gut virome during early life have yet to be characterized, leaving a gap in our understanding of how phages may actively contribute to shifts in bacterial metabolism during this important developmental window.

Defining the healthy developmental program of the infant gut virome is an important step in identifying alterations linked with disease; however, a major hurdle is the extreme sequence-level variation between

individuals, where two individuals may share very few viral sequences¹⁹. To address this, we recently proposed a novel analytical approach to allow for improved inter-study and inter-individual virome comparisons²⁰. By grouping viral sequences based on their predicted bacterial host family (PHF), we can overcome the highly individualized nature of gut viromes to identify broader associations based on phage functional potential.

Using this approach, we conduct a meta-analysis of infant gut viromes to identify global patterns in phage community assembly and functional capacity. We focus on published studies which have free virome (or virus-like particle (VLP)-enriched) sequencing data, rather than bulk metagenomes. While bulk metagenomics can offer higher viral detection rates due to the inclusion of integrated prophages, VLP enrichment prioritizes the extracellular fraction of the virome, which represent a distinct subset of the gut virome with minimal overlap (~10%) with bulk metagenomic populations²¹. Additionally, VLP enrichment yields significantly longer viral contigs²¹, important for robust taxonomic and functional characterization. In our study, the focus on the free virome aims to capture and characterize the viral fraction most likely involved in active predation and horizontal gene transfer at the time of sampling. We identify 12 studies from 8 countries for analysis, consisting of 1,893 samples from 999 infants^{11,22–32}. Most of these studies are limited by small sample sizes or focus on specific geographic regions, while also using distinct analytical approaches. By employing a unified bioinformatic pipeline across these 12 diverse cohorts, we mitigate technical biases to reveal a conserved, global developmental program of the infant gut virome. We describe the development of the gut virome from birth to three years of age, finding that despite rapid fluctuations in the first few months of life, the extracellular virome converges to a stable state by two years of age. This transition from a highly dynamic to stable community is also characterized by a decrease in the abundance of extracellular temperate phages. Using a regression model, we accurately predict infant age from gut virome composition, identifying associations between groups of phages and infant chronological age. Finally, we observe enrichment in certain metabolic pathways among certain

groups of phages, suggesting that the virome's functional potential changes in response to major developmental milestones.

ARTICLE IN PRESS

Results

Integrating 12 cohorts to study the healthy infant gut virome

To characterize the developmental dynamics of the gut virome during early life, we curated a dataset composed of 12 independent studies across eight countries (**Figure 1A, Supplementary Data 1**). All included studies used viral particle enrichment protocols, such as filtration or ultracentrifugation, to sequence the extracellular virome rather than bulk metagenomes, ensuring a high-resolution view of the viral community. The combined dataset included 1,893 enriched fecal viromes from 999 healthy infants, from birth (during the first few days of life) to 38 months of age (**Figure 1B**).

All raw sequencing data were re-processed through a standardized bioinformatic pipeline (**Figure 1C**) to ensure cross-study comparability. This involved *de novo* assembly of individual samples, viral identification, and generation of a non-redundant viral operational taxonomic unit (vOTU) catalog containing 49,745 sequences $\geq 3\text{kb}$ in length. Following read mapping and the application of coverage detection thresholds ($\geq 1\times$ depth and $\geq 75\%$ breadth of coverage), 1,869 samples from 991 infants were retained for downstream analysis (24 samples from 8 infants had insufficient sequencing depth for analysis). While sequencing depth varied across studies (mean $10,494,039 \pm 27,272,299$ raw reads [1.165 ± 2.63 Gbp] per sample), this was accounted for in all downstream modelling to control for potential sampling bias (see Methods).

This integrated dataset represents one of the most comprehensive descriptions of the healthy infant gut virome to date, providing the statistical power necessary to resolve key developmental and functional milestones which may be obscured in smaller, single-cohort studies.

The infant gut virome is highly individualized but with conserved bacterial hosts

To assess the distribution of vOTUs across infant gut viromes, we calculated their prevalence in five discrete age bins based on sampling density and their alignment with key infant developmental milestones: 0 months (neonatal), 1-3 months (early infancy), 4-6 months (transitional), 7-12 months (late infancy), >1 year (toddler). In general, vOTU prevalence, defined as the fraction of infants carrying a given vOTU, was low, with a long tail of rare vOTUs (**Figure 2A**). The maximum prevalence for any single vOTU remained low across all developmental stages, ranging from 15.4% (at 1-3 months) to 48% (at 4-6 months). Notably, no single vOTU was ever prevalent in more than half of the infants in any age bin, suggesting a lack of a core virome signature at the sequence-level.

While this highlights the individuality of the infant virome, we hypothesized that more consistent patterns may emerge at a broader level, where unique vOTUs may converge upon a core set of shared bacterial hosts. Phage host families (PHFs), groupings of vOTUs based on their predicted bacterial host family, provide a functionally relevant unit of classification that can improve such comparisons²⁰. After computationally predicting hosts of vOTUs (36,800 of 49,745 vOTUs [74%] were assigned a host, consistent with²⁰) and grouping the data at the PHF level, prevalence increased substantially (**Figure 2B**). The top PHFs in each age bin were found in a majority of infants, reaching as high as 97.5% for *Bacteroidaceae* at 7-12 months. We also observed clear temporal successions in host targets. For example, phages targeting *Bifidobacteriaceae*, a dominant bacterial family in the neonate gut and key consumer of human milk oligosaccharides¹³, reached peak prevalence during the first 6 months of life, mirroring the typical period of exclusive milk consumption. Across all age bins, the number of core PHFs detected in more than half of infants, increased steadily from 2 at birth to 9 in toddlers (>12 months).

Together, these results highlight the importance of host-based analysis, suggesting that while prevalence of individual vOTUs is highly variable, potentially limiting downstream analyses, the virome maintains functional redundancy by consistently targeting specific bacterial families throughout development.

Gut viral diversity expands rapidly during the first year of life

To model vOTU-level alpha diversity trajectories while controlling for inter-study heterogeneity and repeated measures, we implemented generalized additive mixed models (GAMMs). After accounting for fixed effects (sequencing depth, study) and random effects (individual), only observed richness had a significant association with age over the entire sampling period ($p = 0.042$), increasing significantly during the first 8 months of life, before reaching a plateau (**Figure 2C**, **Supplementary Supplementary Figure 2**). In contrast, the Shannon index remained stable throughout the three-year sampling period. This combination of increasing richness and stable Shannon diversity implies a shift in community structure. Indeed, Pielou's evenness decreased, though not significantly, during the first 18 months before returning to baseline levels. These successional trajectories were largely unaffected by sex, delivery mode, and diet, though analysis of metadata-complete subsets of samples revealed specific divergences later in infancy (**Supplementary Figure 3A**). Cesarean-born infants exhibited lower richness and Shannon diversity after 15 and 18 months, respectively, and female infants had higher richness between 21 and 27 months. Additionally, formula and solid food diets significantly increased community richness compared to exclusive breastfeeding (**Supplementary Figure 3B**).

To evaluate between-sample diversity, we computed PHF-level beta diversity using weighted UniFrac distances (**Figure 2D**). Infant age and study of origin were both significant drivers of community composition.

The infant gut virome converges to a common functional state

These age-related shifts in diversity prompted a deeper analysis of the dynamics of gut virome assembly. Specifically, we sought to quantify the rate of change or developmental velocity, and the level of inter-individual variation or beta dispersion to characterize how the virome stabilizes during the first years of

life. GAMMs were implemented to account for fixed (sequencing depth, study) and random effects (individual).

Virome developmental velocity (VDV) was defined as the distance (PHF-level weighted UniFrac, or vOTU-level Bray-Curtis) between two consecutive samples from the same infant, divided by the time interval (in months) between samples. VDV had a significant negative association with age at both the PHF and vOTU levels ($p < 0.001$), highest at birth and stabilizing by approximately 6 months (**Figure 3A**). Combined with the richness findings, these data suggest that the first 6 to 8 months of life represent the most dynamic period of gut virome development.

To evaluate gut virome convergence and changes in inter-individual variation over time, beta dispersion was calculated as the distance of each sample to its age-specific centroid in multivariate ordination space. At the PHF level, beta dispersion significantly decreased with age ($p = 0.0032$), continuing this downward trend beyond 36 months (**Figure 3B**). In contrast, beta dispersion at the vOTU level had no association with age ($p = 0.23$), suggesting that the gut virome remains highly individualized at the sequence level throughout early life.

The overall patterns in VDV and beta dispersion were largely robust to sex, delivery mode, and diet variations. However, there were differences which emerged upon analysis of metadata-complete subsets of samples: females exhibited significantly lower beta dispersion between 6 and 21 months (**Supplementary Figure 3A**), and formula-fed infants had significantly higher VDV compared to those exclusively breastfed (**Supplementary Figure 3B**).

The discordance between PHF- and vOTU-level beta dispersion patterns suggests that infant gut viromes converge toward a common structure of phage-host interactions, while remaining highly individualized at the sequence level.

A subset of phage host families enables accurate prediction of infant age

Given the successional patterns observed thus far, we next investigated whether gut virome composition could serve as a predictor for infant age. We trained a random forest regression model to predict infant age based on PHF abundances, allowing us to identify a set of age-discriminatory features that define healthy virome maturation. A 70%/30% ratio was used to split the dataset into train and test subsets, respectively.

Performance was evaluated using root mean square error (RMSE) and R^2 metrics. On training data, the model had high accuracy (RMSE = 2.49 months, and $R^2 = 0.93$), while maintaining predictive power on the test set (RMSE = 5.14 months, $R^2 = 0.555$; **Figure 3C**). Using a three-month sliding window to assess local performance, we found that model RMSE was lowest during the first 20 months of life, the period of highest sampling density and most rapid virome turnover, before increasing significantly as the virome stabilized in older infants (**Supplementary Figure 4A**). Leave-one-study-out cross-validation (LOSOCV) was also performed to ensure that the identified patterns were not driven by study-specific biases. In general, removal of any single study from the training dataset reduced the resulting model's performance, indicating that each study provides important information for capturing global successional trends in virome development (**Supplementary Figure 4B**, **Supplementary Figure 4C**). The robustness of the model across the LOSOCV, where predictive power was maintained even when a cohort was entirely new to the model, provides evidence that these successional shifts are not cohort-specific but represent a conserved global program.

To identify the drivers of the age predictor model, we ranked PHFs by their predictive importance and evaluated the direction of their associations with age (**Figure 3D**). The most age-discriminatory features included *Ruminococcaceae*, *Rikenellaceae*, *Bacteroidaceae*, and *Bifidobacteriaceae*. Early-life age prediction was primarily driven by the decline of *Bifidobacteriaceae* and *Veillonellaceae* (highest during first 3 months), while transition to toddler age was marked by increasing abundances of *Rikenellaceae* and *Bacteroidaceae* (highest after 13 months) (**Figure 3D**).

The ability to accurately predict infant age based on PHF-level community composition confirms that the gut virome follows a predictable successional pattern.

Temperate phages dominate the extracellular gut virome during early infancy

Several studies have hypothesized that the gut virome is quickly populated after birth by the induction of temperate phages from colonizing bacteria, though the specific triggers remain unknown. To test this hypothesis, we predicted vOTU lifestyles and modelled the cumulative relative abundance of temperate phages across the first few years of life, using GAMMs to account for fixed (sequencing depth, study) and random effects (individual). Temperate phage abundances were significantly associated with age ($p = 0.002$), highest at birth and decreasing rapidly thereafter (**Figure 4A**). Specific shifts in temperate phage abundances were identified in metadata-complete subsets of infants. Cesarean-born infants had significantly higher temperate phage abundances at birth compared to vaginally-delivered infants (**Supplementary Figure 3A**). Additionally, diet significantly influenced the temperate virome, with exclusively breastfed infants having significantly higher temperate phage abundances compared to those who were formula-fed (**Supplementary Figure 3B**).

The compositional diversity of the temperate virome also varied with time. At birth, most of the detected temperate vOTUs were predicted to infect *Bacteroidaceae*, *Enterobacteriaceae*, *Bifidobacteriaceae*, and *Streptococcaceae* hosts (**Figure 4B**). By one year of age, the temperate virome was primarily composed of phages infecting *Lachnospiraceae*, *Bacteroidaceae*, and *Ruminococcaceae*.

Functional specialization of the gut virome mirrors key developmental milestones

To further characterize the functional potential of the infant virome, protein-coding sequences were predicted from the vOTU library and then clustered into a non-redundant catalog based on $\geq 95\%$ average amino acid identity (AAI) and $\geq 90\%$ coverage (**Figure 1C**). A total of 797,941 protein-coding sequences

were predicted from the library of 49,745 vOTUs, resulting in 485,806 non-redundant protein clusters (reduction of 39.1%). We performed functional annotation of these non-redundant gene catalog against the KOfam database and filtered for auxiliary metabolic genes (AMGs) using a publicly available curated list of KEGG orthologs (KOs) ³³.

We conducted evaluated enrichment of KEGG pathway categories among PHFs, to determine if a functional category is significantly overrepresented relative to the size of that PHF. Among the 10 most prevalent PHFs, *Ruminococcaceae* and *Bacteroidaceae* PHFs were the most functionally diverse PHFs, enriched in 37 and 30 pathway categories, respectively (**Figure 5**). These AMGs included genes for carbohydrate metabolism, biosynthesis of secondary metabolites, and amino acid metabolism.

The relative abundance of different AMGs also followed a temporal succession, reflecting the evolving requirements of the infant gut (**Supplementary Figure 5**). The neonatal period (0-3 months) was characterized by a high abundance of genes involved in energy, glycan, and amino acid metabolism. As infants transitioned to solid food (7-12 months), we observed a peak in AMGs associated with carbohydrate, lipid, xenobiotic, and nucleotide metabolism, which is sustained into the third year of life. We attempted to classify vOTUs into PHFs using a random forest classification model based on AMG profiles (presence/absence) and age at detection. The model's poor performance (mean F1 score across all classes/PHFs of 0.58) suggests a high degree of functional redundancy across PHFs, where specific metabolic repertoires are not necessarily informative of host.

Characterization of the non-bacteriophage fraction of the infant gut virome

While the primary focus of our analysis was the bacteriophage communities, we also characterized the non-phage viral fraction to provide a more comprehensive overview of the infant gut virome. Using geNomad, we identified 534 vOTUs (1.1% of the total dataset) representing 31 non-phage viral families. These were classified by their predicted host, including those infecting animals (n = 466; e.g.

Anelloviridae, *Parvoviridae*, *Adenoviridae*), plants ($n = 13$; e.g. *Tombusviridae*, *Virgaviridae*), fungi/protists ($n = 49$; e.g. *Mimiviridae*, *Genomoviridae*), and archaea ($n = 4$; e.g. *Smacoviridae*, *Sphaerolipoviridae*), as well as virophages ($n = 2$; *Lavidaviridae*) (**Supplementary Figure 6A**). Consistent with previous reports³⁴, these viruses accounted for a small median fraction of the viral community (0.32%); however, their distribution was highly sporadic (mean relative abundance = $7.4\% \pm 19.1\%$), likely reflecting transient dietary inputs, isolated infection events, or study-specific technical effects.

To identify drivers of these populations, we modelled the abundance of each host-grouped category using GAMMs. Animal-infecting viruses were the only group significantly associated with infant age ($p < 0.001$), increasing until reaching a peak at approximately 9 months of age before declining (**Supplementary Figure 6B**). Given this age-dependent signal, we further evaluated the effects of delivery mode, sex, and diet on the animal-infecting fraction. Notably, cesarian-born infants exhibited significantly higher animal-infecting virus abundances from 6 to 12 months of age, compared to vaginally-delivered infants (**Supplementary Figure 6C**). In contrast, animal viral abundances were not significantly influenced by sex or diet, and plant-infecting virus abundances also remained consistent across feeding modes (data not shown).

Discussion

In recent years, there has been growing interest in the human gut virome, especially in the context of early life development. Here, we conducted a meta-analysis of twelve infant cohorts across 8 countries, with 1,893 samples from 999 infants, to model gut virome trajectories during the first three years of life. Although the studies have each individually identified patterns within their respective datasets, geographical and analytical differences have hindered cross-study comparisons. Using a unified bioinformatic pipeline and statistical modelling to control for cohort- and individual-specific effects, we identified significant increases in viral richness during the first eight months of life, consistent with findings from four of the longitudinal cohorts included in this analysis^{11,22,24,28}. By grouping phages together by the hosts family (PHF) they are predicted to infect, we could overcome the issue of virome interindividuality to resolve common successional trends²⁰. While PHFs are linked to the bacteriome, they represent distinct ecological information, providing a functional measurement of the viral predatory landscape rather than a proxy for studying host dynamics²⁰.

We observed high prevalence of multiple PHFs among infants, with *Bacteroidaceae* phages being the most prevalent (detected in 97.5% of infants at 7-12 months). Conversely, the most prevalent vOTU was detected in only 48% of infants at a single timepoint (4-6 months). This contrast suggests that while vOTUs are highly individualized, their ecological niches (at a phage-host family level) are conserved across individuals and cohorts.

To evaluate virome temporal dynamics, we define a novel metric, the virome developmental velocity (VDV). VDV was highest after birth and decreased with time, stabilizing by six months of age. Similarly, beta dispersion, a measure of sample relatedness, decreased with time, suggesting that infant gut viromes converge toward a common structure, despite remaining highly individualized at the sequence level. Taken together with the richness results, these data suggest that the first year of life represents the most active phase of gut virome development, with high viral turnover and continuous introduction of new taxa.

Using a machine learning approach to determine if conserved successional dynamics exist within the infant gut virome across geographically diverse cohorts, we identified several predictors of healthy virome development. Phages targeting *Ruminococcaceae*, *Rikenellaceae*, *Bacteroidaceae*, and *Bifidobacteriaceae* were among the most important predictors of infant age. Interestingly, the model's predictive ability declined as infants aged, with model error increasing rapidly beyond 20 months of age. This suggests that gut virome assembly is most strictly regulated during the first 20 months of life, likely driven by the consistent introduction of new foods and infant developmental milestones. Once the virome stabilizes, as evidenced by our VDV and beta dispersion results, the community enters a state of equilibrium where its composition is no longer primarily governed by age, but likely by other individualized factors such as long-term diet, and unique environmental exposures. Alongside these changes in virome structure is a clear shift in phage-host interactions as the gut environment stabilizes. A recent meta-analysis of 3,154 gut metagenomes from 1,827 healthy infants demonstrated that, the gut bacteriome also follows a universal, globally conserved developmental trajectory¹². In contrast to the stable Shannon diversity observed in the gut virome, that of the gut bacteriome increases consistently over the 18-month study period. Additionally, the most important predictors of age based on gut bacteriome and virome development are distinct. For example, we identify *Ruminococcaceae*, *Rikenellaceae*, *Bacteroidaceae*, and *Bifidobacteriaceae* PHFs as primary predictors of age in viromes, while the corresponding bacterial model relies heavily on *Lachnospiraceae*, in addition to *Ruminococcaceae*.

Consistent with the included studies from Garmaeva et al²², Beller et al²⁸, and Liang et al¹¹, and as described elsewhere in the literature^{35,36}, we observed decreasing abundances of temperate phages with time, suggesting a transition to an adult-like microbiota where temperate phages primarily exist as prophages within host genomes³⁷. While a decrease in temperate phage abundance implies a corresponding increase in the lytic fraction, we adopted a conservative analytical approach by focusing

exclusively on high-confidence temperate vOTUs. Indeed, a temperate phage missing key lysogeny genes due to incomplete metagenomic assembly may falsely be classified by computational tools as being lytic. Beyond changes in phage lifestyle, the functional potential of the gut virome also experiences significant shifts in early life. The high enrichment of AMGs within phages targeting *Ruminococcaceae* and *Bacteroidaceae* – two bacterial families essential for fermentation of dietary fibers¹³ – suggests that the extracellular virome provides a metabolic toolkit which supports host adaptation to a post-weaning, plant polysaccharide-rich diet, potentially enhancing their competitive fitness during dietary transitions.

To address the heterogeneity in metadata reporting from the 12 original studies, we implemented a subset-modelling strategy that maximized the use of available clinical metadata. This approach allowed us to explore the effects of the sex, delivery mode, and diet without significant loss of statistical power, or risk of selection bias, that would have resulted from restricting our analyses to the small number of samples containing all three metadata variables ($n = 280$, or 15% of all samples). Overall, the trajectories of all modelled metrics (alpha diversity, virome developmental velocity, beta dispersion, and temperate phage relative abundance) were broadly conserved regardless of sex, delivery mode, or diet, demonstrating the robust nature of the gut virome development in early life.

While the included study from McCann et al reported significantly higher Shannon diversity in vaginally-delivered infants compared to those born by cesarean section at 12 months of age²⁵, our meta-analysis suggests that this divergence becomes most statistically pronounced later, between 18-24 months of age. Our observation of significantly higher VDV and richness in formula-fed infants over exclusively breastfed infants suggests that dietary choices may accelerate the diversification of the gut virome. This mirrors patterns seen in the gut bacteriome, where formula feeding drives increased Shannon diversity and community maturity³⁸. Although these findings involve different alpha diversity metrics, both suggest that formula feeding correlates with more complex and diversified bacterial and viral communities, ahead of the successional program in breastfed infants.

While not the primary focus of our work, characterization of the non-bacteriophage fraction of the gut virome revealed a high prevalence of animal-infecting viruses, whose abundance varied with age, including *Anelloviridae* and *Parvoviridae*, consistent with several of the cohorts in our analysis^{11,23–25,28–30}.

It is important to note that all twelve studies are from the Global North, with many from high-income countries. Further virome studies from more diverse geographic regions and low- and middle-income countries are necessary to gain a better understanding of global patterns in virome development. Furthermore, all studies included here use library preparation methods which are heavily biased towards sequencing of double-stranded DNA, precluding insights into RNA and single-stranded DNA viruses. Ultimately, by integrating twelve global cohorts, our work characterizes healthy gut virome development in early life, highlighting a conserved successional program and baseline for future studies on the viral drivers of human health.

Methods

Study curation

The following inclusion criteria were used to identify studies to be included in our meta-analysis: (1) public availability of Illumina sequencing data derived from the fecal extracellular virome (enriched virus-like particles); (2) sampling within the first three years of life (0-36 months); and (3) a “healthy” clinical status as defined by the original study investigators. For datasets containing both healthy and clinical cohorts (e.g. type I diabetes or inflammatory bowel disease), only the healthy controls were retained for analysis. This selection process yielded 12 distinct studies, comprising six longitudinal^{11,22–24,28,29} and six cross-sectional^{25–27,30–32} studies (**Supplementary Data 1**). A map highlighting the studies’ sampling locations was generated in R using the *sf* (v1.0-21)^{39,40} and *rnatruearth* (v1.1.0)⁴¹ packages. The underlying geographical base map is sourced from Natural Earth and is in the public domain.

Raw data processing and viral OTU (vOTU) library construction

Raw sequencing reads retrieved for each study from the NCBI Sequence Read Archive (SRA) or European Nucleotide Archive (ENA) using study-specific accession numbers (**Supplementary Data 1**). Read quality control was performed using *fastp* (v0.23.4)⁴² with the following parameters: `-q 15 --cut_right --cut_window_size 4 --cut_mean_quality 20 --length_required 50`. For paired-end datasets, the `--detect_adapter_for_pe` parameter was added. To remove host contamination, reads were mapped against the *Homo sapiens* GRCh38 reference using *bowtie2* (v2.5.2)⁴³ in sensitive mode.

Cleaned reads were assembled *de novo* for each sample individually using *metaSPAdes* (v3.15.4)⁴⁴ with default parameters (including automatically selected k-mer sizes based on read length; parameters: `--meta -t 16 -m 16`). All resulting contigs were provided as input to *geNomad* (v1.7.4)⁴⁵ in end-to-end mode for viral identification. To generate a non-redundant viral catalogue, viral contigs were dereplicated into viral operational taxonomic units (vOTUs). An all-versus-all BLASTN (v2.14.0)⁴⁶ search was performed, and

all geNomad viral contigs were clustered using `anicalc.py` and `aniclust.py` scripts from CheckV ⁴⁷ based on $\geq 95\%$ average nucleotide identity (ANI) over 85% of the shorter contig's length. Following dereplication of viral contigs into vOTUs, a final length filter of 3kb was applied. Only vOTUs meeting this 3kb threshold were retained for downstream analyses. To quantify vOTU abundance, filtered reads from each sample were mapped back to the vOTU library using Bowtie2 (v2.5.2) ⁴³. Coverage statistics were generated using samtools (v1.18) ⁴⁸. A vOTU was considered “detected” in a sample only if it met a minimum threshold of $\geq 1X$ mean depth of coverage and $\geq 75\%$ breadth of coverage ⁴⁹. Following application of these thresholds, 24 samples (from 8 infants) were excluded from downstream analysis for failing to meet these criteria. These samples were distributed across five cohorts: Beller 2022 (n = 1), Liang 2020a (n = 8), Lim 2015 (n = 3), Maqsood 2019 (n = 1), and Walters 2023 (n = 11). Computational host prediction was performed using iPHoP (v1.0.0, db version: Sept_2021_pub) ⁵⁰, and viral lifestyles (temperate or non-temperate) were predicted using BACPHLIP (v0.9.6) ⁵¹. Viral taxonomy was assigned using geNomad in end-to-end mode. The ICTV Virus Properties database (<https://ictv.global/virus-properties>) was used to identify common hosts for these viral families.

Alpha diversity analyses and statistical modelling

Alpha diversity metrics (observed richness, Shannon index, and Pielou's evenness) at the vOTU level were calculated in R (v4.5.1) using phyloseq (1.52.0) ⁵². To model diversity trajectories while controlling for inter-study heterogeneity and repeated measures, generalized additive mixed models (GAMMs) were implemented with mgcv (v1.9-3) ⁵³. Different model distributions were chosen based on the properties of each diversity metric. Observed richness was modelled using a negative binomial distribution (family = `nb()`) to account for overdispersed count data. Shannon index was modelled using a Gaussian distribution (family = `gaussian()`). Pielou's evenness was modelled using a beta regression (family = `betar()`) to accommodate continuous values constrained in the (0, 1) interval. For Pielou's evenness, values were

transformed prior to modelling to compress the [0, 1] interval to (0, 1) following the formula $y' = \frac{y(n-1)+0.5}{n}$, as described by Smithson and Verkuilen ⁵⁴, where n is the total number of samples, to accommodate the requirements of the beta distribution.

The general model formula was defined as $y \sim s(\text{age_months}) + \log_{10}(\text{depth}) + s(\text{age_months}, \text{study}, \text{bs} = \text{"fs"}, m = 1) + s(\text{subjectID}, \text{bs} = \text{"re"})$. Sequencing depth ($\log_{10}$ transformed) was included as a fixed effect to control for varying sequencing depth across samples without the need for rarefaction. To maximize statistical power while accounting for missing metadata, additional variables (infant sex, delivery mode, and diet at the time of sampling) were evaluated separately using metadata-complete subsets (**Supplementary Data 1**). These subsets included n = 774 samples for infant sex (41% of all samples), n = 500 for delivery mode (26%), and n = 645 for diet (34%). Post-hoc pairwise differences between sexes, delivery modes, and dietary categories were calculated using the emmeans package (v2.0.0) ⁵⁵ with Tukey's HSD adjustment for multiple comparisons. ComplexHeatmap (v2.24.1) ⁵⁶ was used to visualize the emmeans estimates and significance across all pairwise dietary contrasts.

Overall model performance was evaluated based on deviance explained and the significance of the smooth terms. Visualization of the fitted smooth terms was performed using the gratia package (v0.11.1) ⁵⁷.

To identify the specific window where observed richness was increasing before stabilizing, we analyzed the first derivative of the fitted age_months smooth term using the derivatives() function from the gratia package. The point of stabilization was defined as the earliest age (in months) where the 95% confidence interval of the first derivative became negative, indicating the rate of change of richness was no longer significantly different from zero.

Prevalence of vOTUs and PHFs

To assess the distribution of viruses across cohorts, we calculated the prevalence of both individual vOTUs and phage host families (PHFs). PHFs are groupings of vOTUs based on their predicted bacterial host

family. We have previously shown that PHFs are a functionally relevant, qualitative unit of phage classification which can improve virome comparisons across studies ²⁰.

Samples were grouped into five age bins: 0 months (neonatal), 1-3 months (early infancy), 4-6 months (transitional), 7-12 months (late infancy), >1 year (toddler). These intervals were selected based on both sampling density and their alignment with key infant developmental milestones: the period of exclusive milk or formula feeding (0-3 months), the introduction of complementary solid foods (4-6 months), and the establishment of a diversified diet (>12 months) ⁵⁸⁻⁶¹. Although only six of twelve studies had infant diet metadata available, the observed dietary patterns were consistent with global pediatric guidelines, with solid food starting to be introduced as of 4 months, justifying the application of these specific cutoffs (**Supplementary Figure 1**).

Prevalence was defined as the proportion of infants within an age bin which have a vOTU/PHF at detectable levels (see above). For longitudinal studies where an infant was sampled multiple times within an age bin, the vOTU/PHF was considered present if detected in at least one sample during that period. This binning strategy was used to standardize the contribution of infants from both cross-sectional and longitudinal cohorts.

Beta diversity analyses

Beta diversity was assessed using the weighted UniFrac metric on PHF-level viromes. The GTDB tree used by iPHoP (bac120_r202.tree) was used for weighted UniFrac distance calculations. To test the impact of metadata variables (e.g., infant age, study) on viral community composition, a permutational multivariate analysis of variance (PERMANOVA) test was performed using the `adonis2()` function in `vegan` (2.7-2) ⁶². The primary model included sequencing depth ($\log_{10}$ transformed), study, and age (months). To account for the longitudinal nature of the data, permutations were constrained within individual subjects using the `strata` argument. A second PERMANOVA model was used to quantify the

variance explained (R^2) by inter-individual differences. This model's formula included only sequencing depth ($\log_{10}$ transformed), study, and subjectID. No p-value is reported for the subjectID term as the high number of levels ($n = 991$) relative to the total number of samples prevents us from conducting a valid permutation test for this variable, rather it is included only to estimate the total R^2 of subjectID.

Virome developmental velocity and beta dispersion

To quantify the rate of virome community change over time, we calculated virome developmental velocity using the six longitudinal studies ($n = 989$ samples from 120 individuals, across 6 studies). Developmental velocity was defined as the distance (PHF-level weighted UniFrac, or vOTU-level Bray-Curtis) between consecutive samples from the same infant, divided by the time interval (in months) between these consecutive samples. The age assigned to each velocity estimate was that of the second (later) sample.

Inter-individual variation was assessed with beta dispersion, calculated as the distance of each sample to its age-specific (in months) centroid in multivariate space, using the `betadisper()` function in the `vegan` package.

Both velocity and beta dispersion metrics were modelled using GAMMs with a Tweedie distribution (`family = tw()`) to accommodate the non-negative nature of the data. The velocity model included a fixed effect for `mean_log10_depth` (the average sequencing depth of the two compared samples), while the beta dispersion model included `log10(depth)`. Both models used the previously described fixed and random effect structure to account for heterogeneity among cohorts: $y \sim s(\text{age_months}) + \text{mean_log10_depth} + s(\text{age_months}, \text{study}, \text{bs} = "fs", m = 1) + s(\text{subjectID}, \text{bs} = "re")$. The effects of infant sex, delivery mode, and diet on virome developmental velocity and beta dispersion were evaluated using the same subset-modelling and post-hoc testing framework described for alpha diversity analyses above. Visualization of the fitted smooth terms was performed using the `gratia` package.

Random forest age predictor

To determine the predictability of virome maturation, we trained a random forest (RF) regression model using the randomForest (v4.7-1.2)⁶³ package to predict infant age (months) based on PHF relative abundances. The dataset was split into training (70%) and testing (30%) sets using the createDataPartition() function from the caret (v7.0-1)⁶⁴ package to maintain a balance of age across both datasets. Model performance was evaluated by calculating the root mean squared error (RMSE) and variance explained (R^2) on both training and test datasets. To evaluate the model's accuracy over the course of infant development, a sliding window approach was used, where RMSE was calculated for overlapping 3-month age windows. This higher-resolution approach provides a continuous assessment of the model's predictive performance during periods of rapid virome change, avoiding boundary effects which could arise when using the discrete age bins used for prevalence calculations.

The most important features (PHFs) were identified using the IncNodePurity metric from the importance() function, part of the randomForest package. A heatmap of the top 20 most important PHFs was generated using ComplexHeatmap to visualize their mean abundance patterns during the first years of life. For each PHF, a Pearson correlation coefficient (r) was calculated against age_months. These values were squared to determine the coefficient of determination (R^2), which represents the proportion of variance in age explained by the individual PHF. A signed R^2 value was calculated by multiplying the R^2 by the sign of the correlation coefficient, to indicate the direction of the PHF's association with age.

Auxiliary metabolic gene identification and analysis

To characterize the functional potential of the infant virome, protein-coding sequences were predicted from the vOTU library using prodigal-gv (v2.11.0)^{45,65} in meta mode. The resulting gene set was clustered into a non-redundant catalog using mmseqs2 (b804fbe384e6f6c9fe96322ec0e92d48bccd0a42)⁶⁶ based on $\geq 95\%$ average amino acid identity (AAI) and $\geq 90\%$ coverage.

Functional annotation was performed by searching the non-redundant gene catalog against the KOfam database (v20250201) using `hmmsearch` (HMMER v3.4)⁶⁷ with an E-value threshold of 1×10^{-5} . Results were processed in R using the `rhmmmer` package (v0.2.0) (<https://github.com/arendsee/rhmmmer>). Auxiliary metabolic genes (AMGs) were specifically identified by filtering annotations against a curated list of KEGG orthologs (KOs) established by the virus identification tool VIBRANT³³.

A heatmap of AMG metabolism categories was generated using `ComplexHeatmap` to visualize abundance patterns during the first years of life. AMG abundance (either at the metabolism or pathway level) within each sample was calculated as the cumulative abundance of all vOTUs carrying that specific AMG. To avoid biases introduced by potential gene duplications or functional redundancy, copy number was not considered. Relative abundances were min/max-scaled per category to facilitate comparisons of trends across categories with significantly differing abundances.

Enrichment was assessed using a hypergeometric test (Fisher's exact test). This test compares the frequency of a specific AMG category within a single PHF against its global frequency across all PHFs. By accounting for the number of unique vOTU within each PHF, this test determines if a functional category is significantly overrepresented relative to the size of that PHF. P-values were adjusted for multiple testing using the Benjamini-Hochberg method, and categories were considered significantly enriched if the adjusted p-value was less than 0.05.

Statistics & reproducibility statement

No statistical method was used to predetermine sample size, all eligible samples were included in our analyses, as described in the "Study curation" section above. 24 samples were excluded from analysis due to insufficient sequencing depth, as described in the Results section. Generalized additive mixed models were fit to all samples, and also to metadata-complete subsets of samples. Samples were randomly assigned to either training or test datasets when working with machine learning models. Models were

blinded to test data during the training process. Leave-one-study-out cross-validation of machine learning models was performed 12 times, each iteration excluding a different study.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used Google Gemini (Alphabet Inc., Mountain View, USA) to improve sentence clarity and readability. Overall sentence meaning was not altered, and there were no changes to any of the presented data, facts, or conclusions by using this tool. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Data Availability

The raw virome sequencing data for each study are available in the NCBI Sequence Read Archive (SRA) under the following BioProject accessions: Lim 2015 (PRJNA284162), Zhao 2017 (PRJNA387903), McCann 2018 (PRJNA385126), Maqsood 2019 (PRJEB33578), Liang 2020a (PRJNA524703), Beller 2022 (PRJNA693793), Walters 2023 (PRJNA916952), Shah 2023 (PRJEB46943), Pannaraj 2018 (PRJNA427187), Yan 2021 (PRJNA641593), Liang 2020b (PRJNA524703). The raw virome sequencing data from Garmaeva 2024 study were not publicly available and were made available upon request from the study authors.

Code Availability

Code used for data analysis is available at <https://zenodo.org/records/19701266>⁶⁸.

References

1. Shamash, M. & Maurice, C. F. Phages in the infant gut: a framework for virome development during early life. *The ISME Journal* <https://doi.org/10.1038/s41396-021-01090-x> (2021) doi:10.1038/s41396-021-01090-x.
2. Lim, E. S., Wang, D. & Holtz, L. R. The bacterial microbiome and virome milestones of infant development. *Trends in Microbiology* **24**, 801–810 (2016).
3. Vatanen, T. *et al.* The human gut microbiome in early-onset type 1 diabetes from the TEDDY study. *Nature* **562**, 589–594 (2018).
4. Subramanian, S. *et al.* Persistent gut microbiota immaturity in malnourished Bangladeshi children. *Nature* **510**, 417–421 (2014).
5. Smith, M. I. *et al.* Gut Microbiomes of Malawian Twin Pairs Discordant for Kwashiorkor. *Science* **339**, 548–554 (2013).
6. Arrieta, M.-C. *et al.* Early infancy microbial and metabolic alterations affect risk of childhood asthma. *Sci. Transl. Med.* **7**, (2015).
7. Fernández-Pato, A. *et al.* Early-life development of the gut virome and plasmidome: A longitudinal study in cesarean-born infants. *Cell Reports* **44**, 115731 (2025).
8. Leal Rodríguez, C. *et al.* The infant gut virome is associated with preschool asthma risk independently of bacteria. *Nat Med* **30**, 138–148 (2024).
9. Zeng, S. *et al.* A metagenomic catalog of the early-life human gut virome. *Nat Commun* **15**, 1864 (2024).
10. Kaelin, E. A. *et al.* Longitudinal gut virome analysis identifies specific viral signatures that precede necrotizing enterocolitis onset in preterm infants. *Nat Microbiol* **7**, 653–662 (2022).
11. Liang, G. *et al.* The stepwise assembly of the neonatal virome is modulated by breastfeeding. *Nature* **581**, 470–474 (2020).

12. Fahur Bottino, G. *et al.* Early life microbial succession in the gut follows common patterns in humans across the globe. *Nat Commun* **16**, 660 (2025).
13. Arrieta, M.-C. Microbiome Maturation Trajectory and Key Milestones in Early Life. *Ann Nutr Metab* **81**, 20–27 (2025).
14. Kreuze, K., Friman, V.-P. & Vatanen, T. Mobile genetic elements: the hidden puppet masters underlying infant gut microbiome assembly? *Microbiome Res Rep* **3**, (2024).
15. Breitbart, M., Bonnain, C., Malki, K. & Sawaya, N. A. Phage puppet masters of the marine microbial realm. *Nature Microbiology* **3**, 754–766 (2018).
16. Shaffer, M. *et al.* DRAM for distilling microbial metabolism to automate the curation of microbiome function. *Nucleic Acids Research* **48**, 8883–8900 (2020).
17. Brown, T. L., Charity, O. J. & Adriaenssens, E. M. Ecological and functional roles of bacteriophages in contrasting environments: marine, terrestrial and human gut. *Current Opinion in Microbiology* **70**, 102229 (2022).
18. Ji, S. *et al.* Engrafting gut bacteriophages have potential to modulate microbial metabolism in fecal microbiota transplantation. Preprint at <https://doi.org/10.21203/rs.3.rs-5259313/v1> (2024).
19. Shkoporov, A. N. *et al.* The human gut virome is highly diverse, stable, and individual specific. *Cell Host and Microbe* **26**, 527–541 (2019).
20. Shamash, M., Sinha, A. & Maurice, C. F. Improving gut virome comparisons using predicted phage host information. *mSystems* **10**, e01364-24 (2025).
21. Gregory, A. C. *et al.* The gut virome database reveals age-dependent patterns of virome diversity in the human gut. *Cell Host & Microbe* **28**, 724–740 (2020).
22. Garmaeva, S. *et al.* Transmission and dynamics of mother-infant gut viruses during pregnancy and early life. *Nat Commun* **15**, 1945 (2024).

23. Lim, E. S. *et al.* Early life dynamics of the human gut virome and bacterial microbiome in infants. *Nature Medicine* **21**, 1228–1234 (2015).
24. Zhao, G. *et al.* Intestinal virome changes precede autoimmunity in type I diabetes-susceptible children. *Proceedings of the National Academy of Sciences of the United States of America* **114**, E6166–E6175 (2017).
25. McCann, A. *et al.* Viromes of one year old infants reveal the impact of birth mode on microbiome diversity. *PeerJ* **2018**, e4694 (2018).
26. Maqsood, R. *et al.* Discordant transmission of bacteria and viruses from mothers to babies at birth. *Microbiome* **7**, 156 (2019).
27. Liang, G. *et al.* Dynamics of the Stool Virome in Very Early-Onset Inflammatory Bowel Disease. *Journal of Crohn's and Colitis* **14**, 1600–1610 (2020).
28. Beller, L. *et al.* The virota and its transkingdom interactions in the healthy infant gut. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2114619119 (2022).
29. Walters, W. A. *et al.* Longitudinal comparison of the developing gut virome in infants and their mothers. *Cell Host & Microbe* **31**, 187–198.e3 (2023).
30. Shah, S. A. *et al.* Expanding known viral diversity in the healthy infant gut. *Nat Microbiol* <https://doi.org/10.1038/s41564-023-01345-7> (2023) doi:10.1038/s41564-023-01345-7.
31. Pannaraj, P. S. *et al.* Shared and distinct features of human milk and infant stool viromes. *Frontiers in microbiology* **9**, (2018).
32. Yan, Q. *et al.* Characterization of the gut DNA and RNA Viromes in a Cohort of Chinese Residents and Visiting Pakistanis. *Virus Evolution* **7**, veab022 (2021).
33. Kieft, K., Zhou, Z. & Anantharaman, K. VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. *Microbiome* **8**, 90 (2020).

34. Liang, G. & Bushman, F. D. The human virome: assembly, composition and host interactions. *Nature Reviews Microbiology* **0123456789**, (2021).
35. Khan Mirzaei, M. *et al.* Bacteriophages isolated from stunted children can regulate gut bacterial communities in an age-specific manner. *Cell Host and Microbe* **27**, 199–212 (2020).
36. Redgwell, T. A. *et al.* Prophages in the infant gut are pervasively induced and may modulate the functionality of their hosts. *npj Biofilms Microbiomes* **11**, 46 (2025).
37. Khan Mirzaei, M. & Maurice, C. F. Ménage à trois in the human gut: interactions between host, bacteria and phages. *Nature Reviews Microbiology* **15**, 397–408 (2017).
38. Fehr, K. *et al.* Breastmilk feeding practices are associated with the co-occurrence of bacteria in mothers' milk and the infant gut: the CHILD cohort study. *Cell Host & Microbe* **28**, 285–297 (2020).
39. Pebesma, E. & Bivand, R. *Spatial Data Science: With Applications in R*. (Chapman and Hall/CRC, New York, 2023). doi:10.1201/9780429459016.
40. Pebesma, E. Simple Features for R: Standardized Support for Spatial Vector Data. *The R Journal* **10**, 439–446 (2018).
41. Massicotte, P. & South, A. *Rnaturalearth: World Map Data from Natural Earth*. (2025). doi:10.32614/CRAN.package.rnaturalearth.
42. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
43. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**, 357–359 (2012).
44. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. metaSPAdes: a new versatile metagenomic assembler. *Genome Res.* **27**, 824–834 (2017).
45. Camargo, A. P. *et al.* Identification of mobile genetic elements with geNomad. *Nat Biotechnol* **42**, 1303–1312 (2024).

46. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
47. Nayfach, S. *et al.* CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nature Biotechnology* <https://doi.org/10.1038/s41587-020-00774-7> doi:10.1038/s41587-020-00774-7.
48. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
49. Roux, S., Emerson, J. B., Eloë-Fadrosh, E. A. & Sullivan, M. B. Benchmarking viromics: An in silico evaluation of metagenome-enabled estimates of viral community composition and diversity. *PeerJ* **2017**, e3817 (2017).
50. Roux, S. *et al.* iPHoP: An integrated machine learning framework to maximize host prediction for metagenome-derived viruses of archaea and bacteria. *PLoS Biol* **21**, e3002083 (2023).
51. Hockenberry, A. J. & Wilke, C. O. BACPHLIP: predicting bacteriophage lifestyle from conserved protein domains. *PeerJ* **9**, e11396 (2021).
52. McMurdie, P. J. & Holmes, S. phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PLoS ONE* **8**, e61217 (2013).
53. Wood, S. N. Fast Stable Restricted Maximum Likelihood and Marginal Likelihood Estimation of Semiparametric Generalized Linear Models. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **73**, 3–36 (2011).
54. Smithson, M. & Verkuilen, J. A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychological Methods* **11**, 54–71 (2006).
55. Lenth, R. V. & Piaskowski, J. *Emmeans: Estimated Marginal Means, Aka Least-Squares Means*. (2025).
56. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**, 2847–2849 (2016).

57. Simpson, G. L. gratia: An R package for exploring generalized additive models. *Journal of Open Source Software* **9**, 6962 (2024).
58. Meek, J. Y. & Noble, L. Technical Report: Breastfeeding and the Use of Human Milk. *Pediatrics* **150**, e2022057989 (2022).
59. Fewtrell, M. *et al.* Complementary Feeding: A Position Paper by the European Society for Paediatric Gastroenterology, Hepatology, and Nutrition (ESPGHAN) Committee on Nutrition. *J. pediatr. gastroenterol. nutr.* **64**, 119–132 (2017).
60. National Health Commission of the People's Republic of China. Nutrition Guidelines of Complementary Feeding for Infants and Toddlers (WS/T 678—2020). *Biomedical and Environmental Sciences* **36**, 899 (2023).
61. *WHO Guideline for Complementary Feeding of Infants and Young Children 6-23 Months of Age.* (World Health Organization, Geneva, 2023).
62. Oksanen, J. *et al.* vegan: Community Ecology Package. (2019).
63. Liaw, A. & Wiener, M. Classification and Regression by randomForest. *R News* **2**, 18–22 (2002).
64. Kuhn & Max. Building Predictive Models in R Using the caret Package. *Journal of Statistical Software* **28**, 1–26 (2008).
65. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).
66. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol* **35**, 1026–1028 (2017).
67. Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Research* **39**, W29–W37 (2011).
68. Shamash, M. mshamash/infant_gut_virome_manuscript: Manuscript - v2. Zenodo <https://doi.org/10.5281/zenodo.19701266> (2026).

ARTICLE IN PRESS

Acknowledgements

We would like to acknowledge and thank the participants and authors of the 12 studies included in our analysis. This research was enabled in part by computational support provided by Calcul Québec (calculquebec.ca) and the Digital Research Alliance of Canada (alliancecan.ca).

Funding Statement

This work was funded by a Canadian Institutes of Health Research Canada Project grant (PJT-175065), Natural Sciences and Engineering Research Council of Canada Discovery Grant (RGPIN-2023-04216), and Canada Research Chair in Gut Microbial Interactions (Tier 2) to Corinne F. Maurice. Michael Shamash is supported by the Canadian Institutes of Health Research Canada Graduate Scholarship to Honor Nelson Mandela (CIHR CGS-D; #DF2-187718), and the *Fonds de recherche du Québec-Santé: Bourse de formation au doctorat* (FRQS; #311071).

Author Contributions Statement

M.S. curated data for analysis, conceived and performed the analyses, prepared figures and tables, authored and reviewed drafts of the manuscript, and approved the final draft. C.F.M. conceived the analyses, reviewed drafts of the manuscript, and approved the final draft.

Competing Interests Statement

The authors declare no competing interests.

Figure Legends

Figure 1. Integrated analysis and characterization of the global infant gut virome. (A) Geographic overview of sampling locations for the 12 cohorts included in this meta-analysis, representing 999 infants, coloured by cohort^{11,22–32}. (B) Longitudinal sampling distribution across the first three years of life. Inset plots summarize the total number of samples ($n = 1,893$) and participants contributed by each study ($n = 999$). Plots are coloured by cohort. (C) Schematic of the bioinformatic pipeline. A total of 19.865 billion reads were processed through quality control and *de novo* metagenomic assembly. Viral identification and dereplication at 95% average nucleotide identity (ANI) over 85% alignment fraction (AF) resulted in a library of 49,745 viral operational taxonomic units (vOTUs) ≥ 3 kb in length. Functional potential of a non-redundant catalog of 485,806 protein clusters (95% average amino acid identity [AAI], 90% AF) was evaluated through annotation against the KOfam database.

Figure 2. Prevalence and diversity patterns of the gut virome across the first three years of life. Prevalence of (A) 49,745 viral operational taxonomic units (vOTUs), and (B) phage host families (PHFs) across five developmental age bins. Inset plots highlight the mean prevalence of the top ten most prevalent vOTUs and PHFs. Prevalence plots are colour-coded by developmental age bin, while inset plots are colour-coded by vOTU/PHF. (C) Longitudinal vOTU-level alpha diversity trajectories modelled using generalized additive mixed models (GAMMs), colour-coded by alpha diversity metric. The fitted smooth represents the estimated mean of the non-linear relationship between each metric and infant age (in months), where the shaded ribbon represents the 95% confidence interval. (D) Community structure visualized with principal coordinates analysis (PCoA) of weighted UniFrac distances at the PHF level. Samples are coloured according to infant age (in months, left) and original study cohort (right).

Accompanying PERMANOVA results quantify the relative contribution of metadata variables to overall virome beta diversity.

Figure 3. Rapid and predictable assembly of the infant gut virome. (A) Virome developmental velocity (VDV), calculated at the viral operational taxonomic unit (vOTU) and phage host family (PHF) levels, modelled using a generalized additive mixed model (GAMM) and colour-coded by classification level (vOTU or PHF). The fitted smooth represents the estimated mean of the non-linear relationship between VDV and infant age in months, where the shaded ribbon represents the 95% confidence interval. (B) Beta dispersion, calculated at the vOTU and PHF levels, modelled using GAMMs and colour-coded by classification level (vOTU or PHF). The fitted smooth represents the estimated mean of the non-linear relationship between each model and infant age (in months), where the shaded ribbon represents the 95% confidence interval. (C) Random forest regression model performance for predicting infant chronological age from PHF-level virome composition. Predicted versus actual age (in months) is shown, colour-coded by dataset (training or testing). (D) Heatmap of the top twenty most important PHF features in the age predictor model, with mean relative abundance across 3-month age bins (min/max scaled for each PHF). Accompanying coloured strip and bar plot indicate relative overall contribution and importance of each PHF to the model, respectively.

Figure 4. Abundance and diversity of the temperate infant gut virome. (A) Cumulative temperate phage relative abundance modelled using a generalized additive mixed model (GAMM). The fitted smooth represents the estimated mean of the non-linear relationship between temperate phage relative abundance and infant age in months ($p = 0.002$), where the shaded ribbon represents the 95% confidence interval. (B) Composition of the temperate virome at the phage host family (PHF) level across five

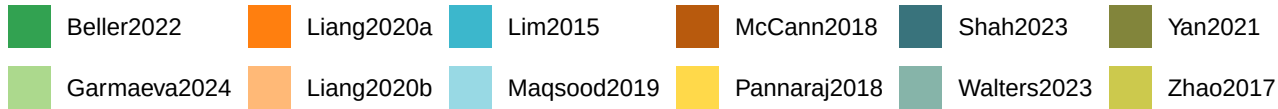
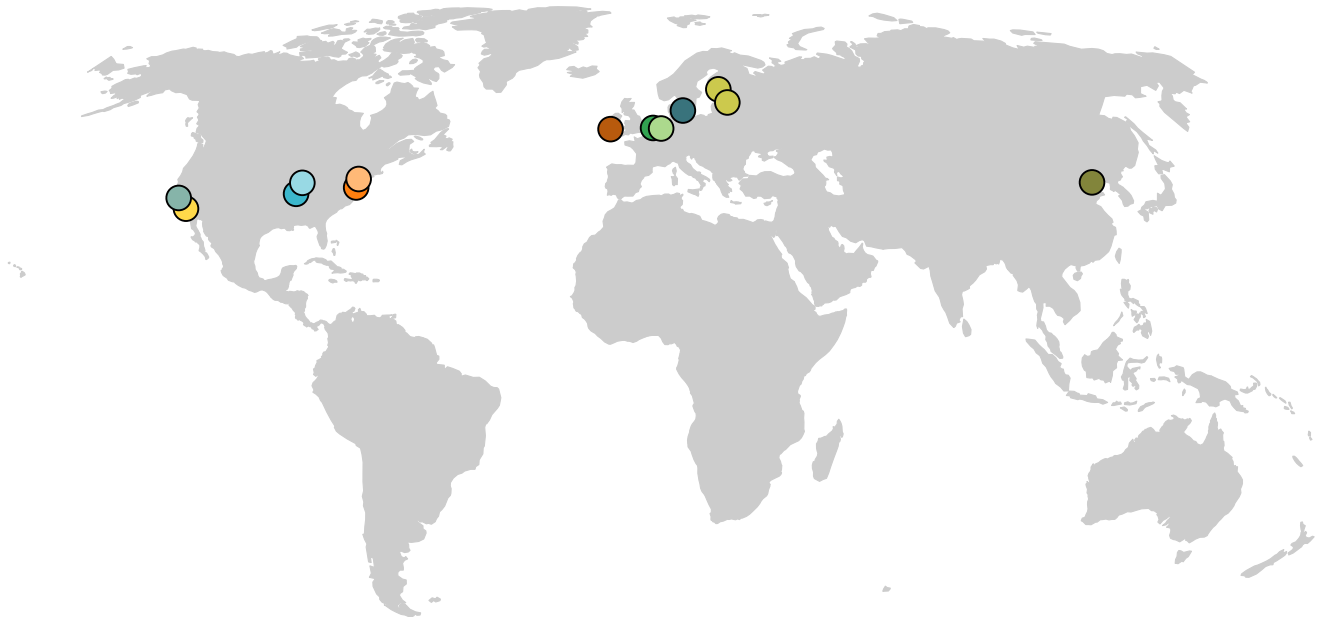
developmental age bins. Data represent the compositional richness (proportion of unique viral operational taxonomic units, vOTUs) within each bin, unweighted by read abundance, and are colour-coded by PHF.

Figure 5. Functional specialization of the infant gut virome through auxiliary metabolic gene enrichment. Enrichment of KEGG pathway categories among the 10 most prevalent phage host families (PHFs). KEGG pathways (y-axis) are organized by broader KEGG metabolism categories. Dot size indicates the enrichment ratio, while colour indicates statistical significance. Enrichment was determined using a hypergeometric test (Fisher's exact test) to determine identify significantly overrepresented functional categories relative to the total gene count and size of a PHF. Only significant ($FDR < 0.05$) results are shown.

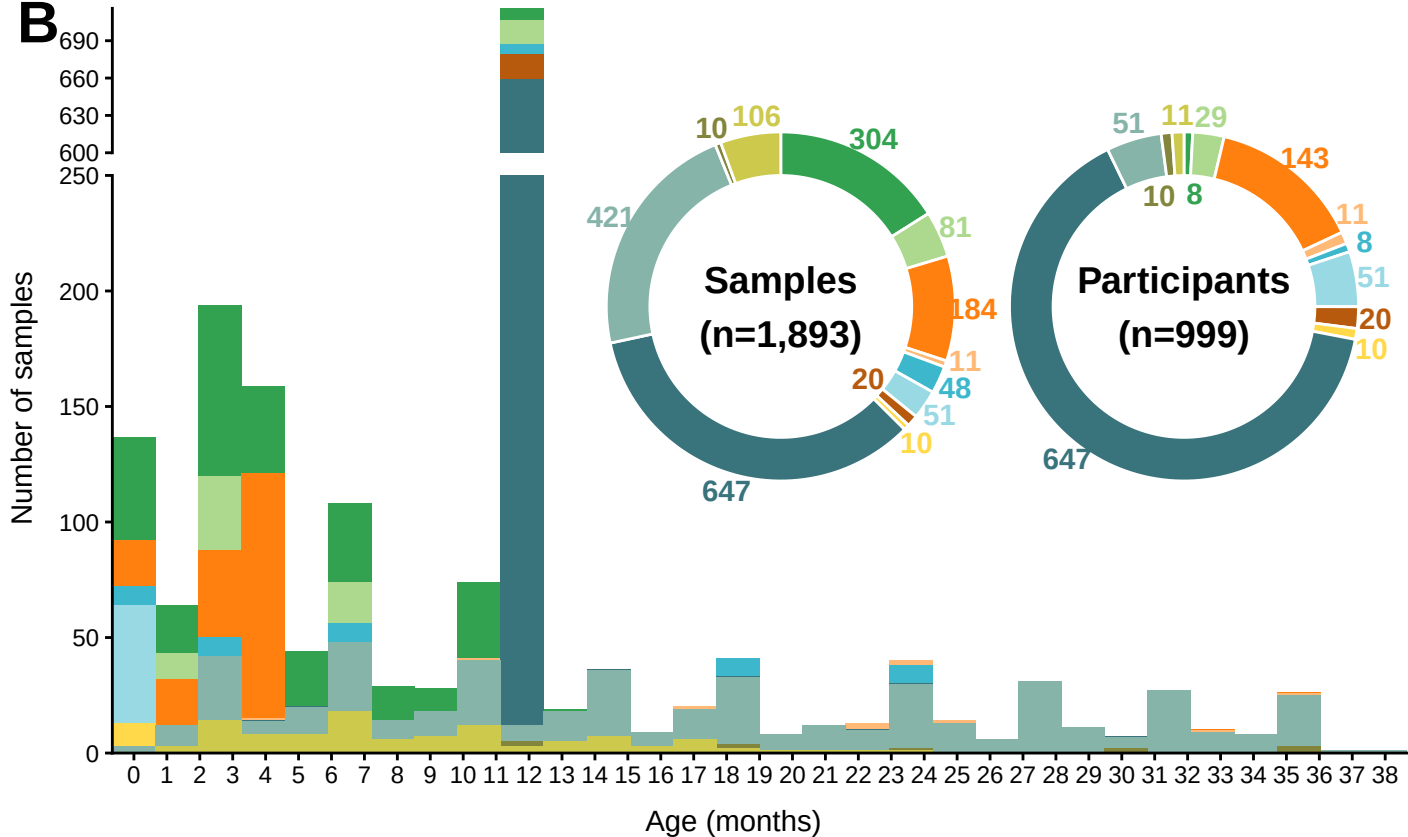
Here, via a meta-analysis of 12 cohorts, the authors reveal that the infant gut virome follows a predictable, conserved developmental program, and that despite individual variation at the sequence level, viral communities rapidly expand after birth and stabilize by two years of age.

Peer Review Information: *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work. A peer review file is available.

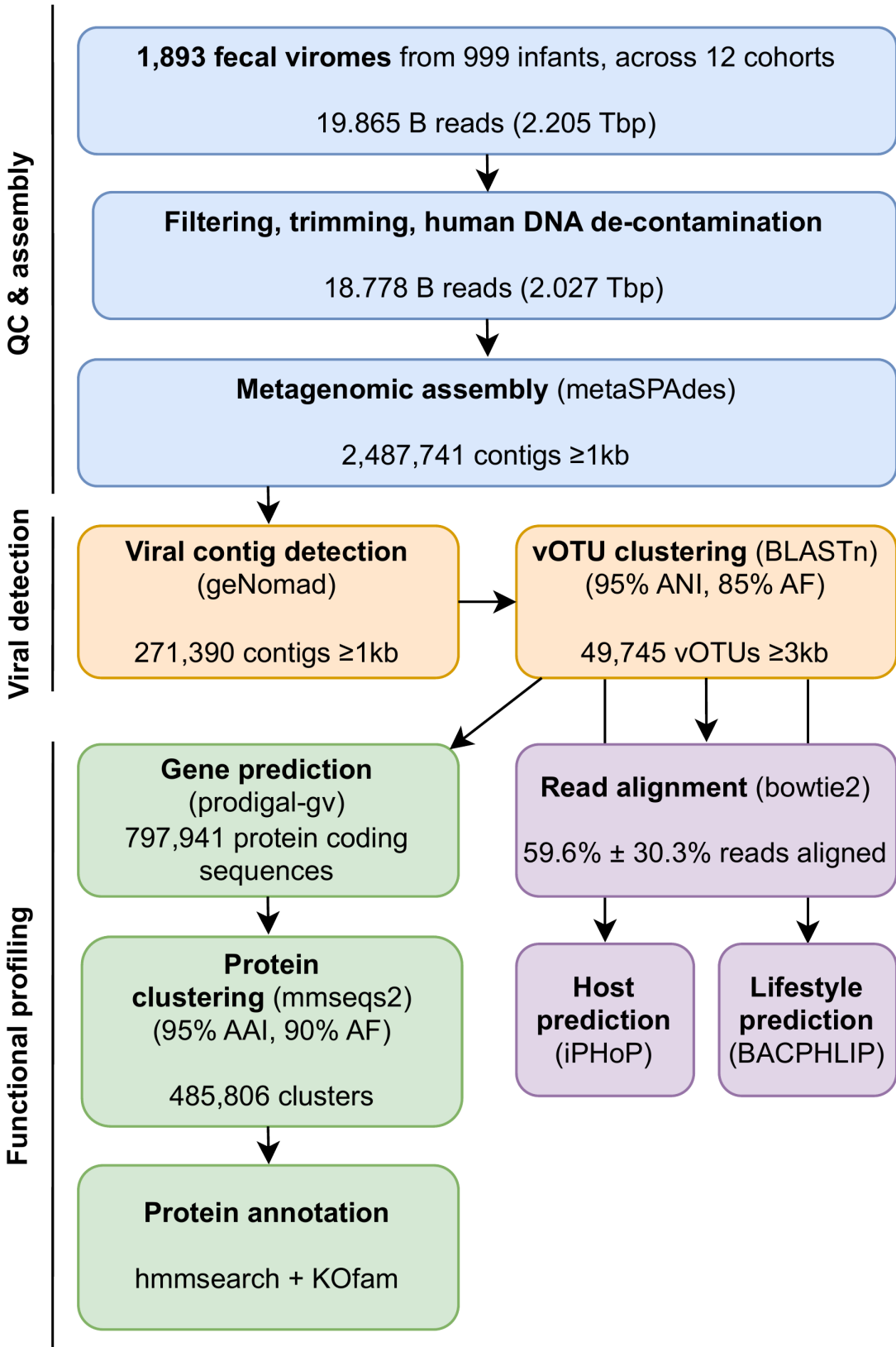
A

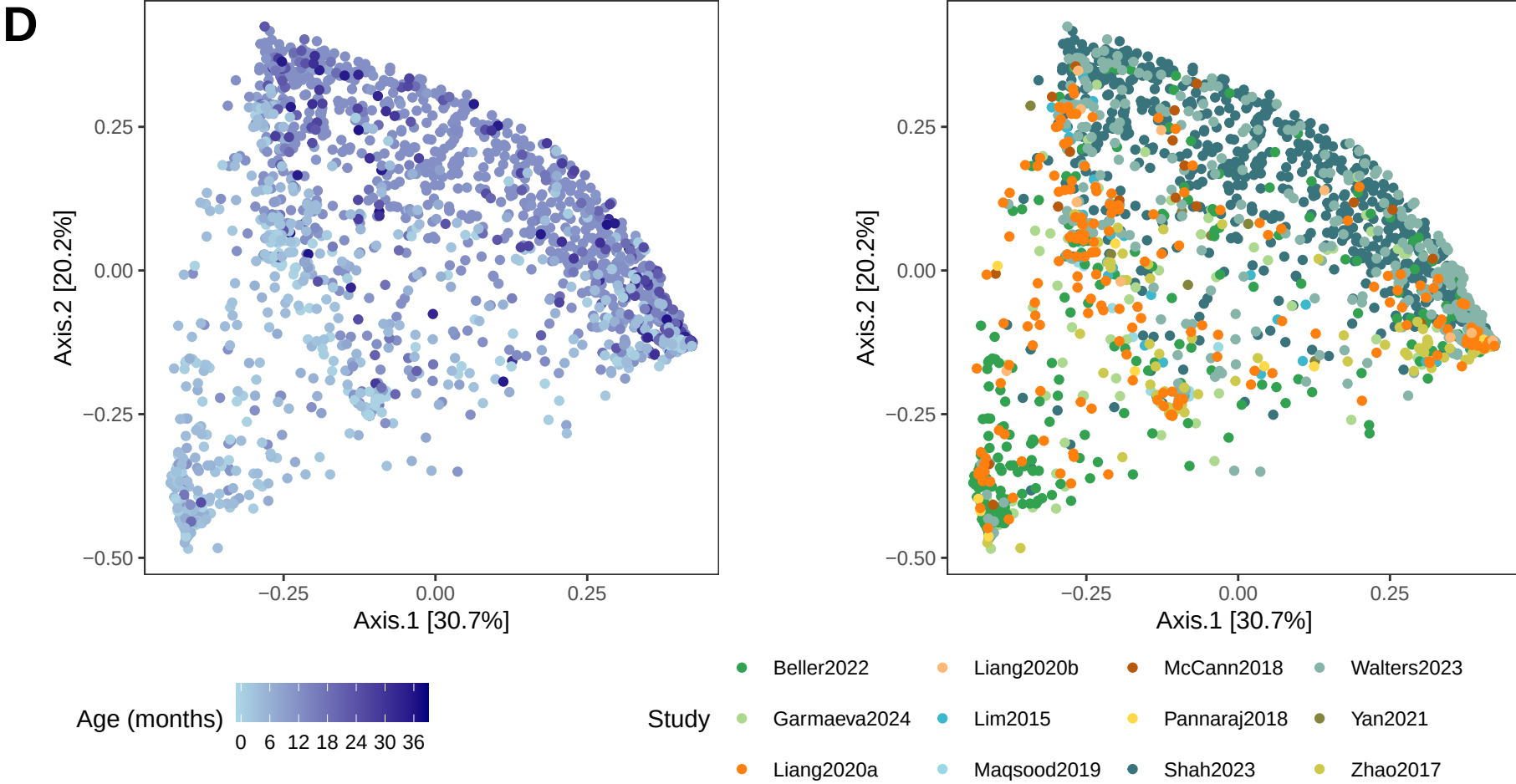
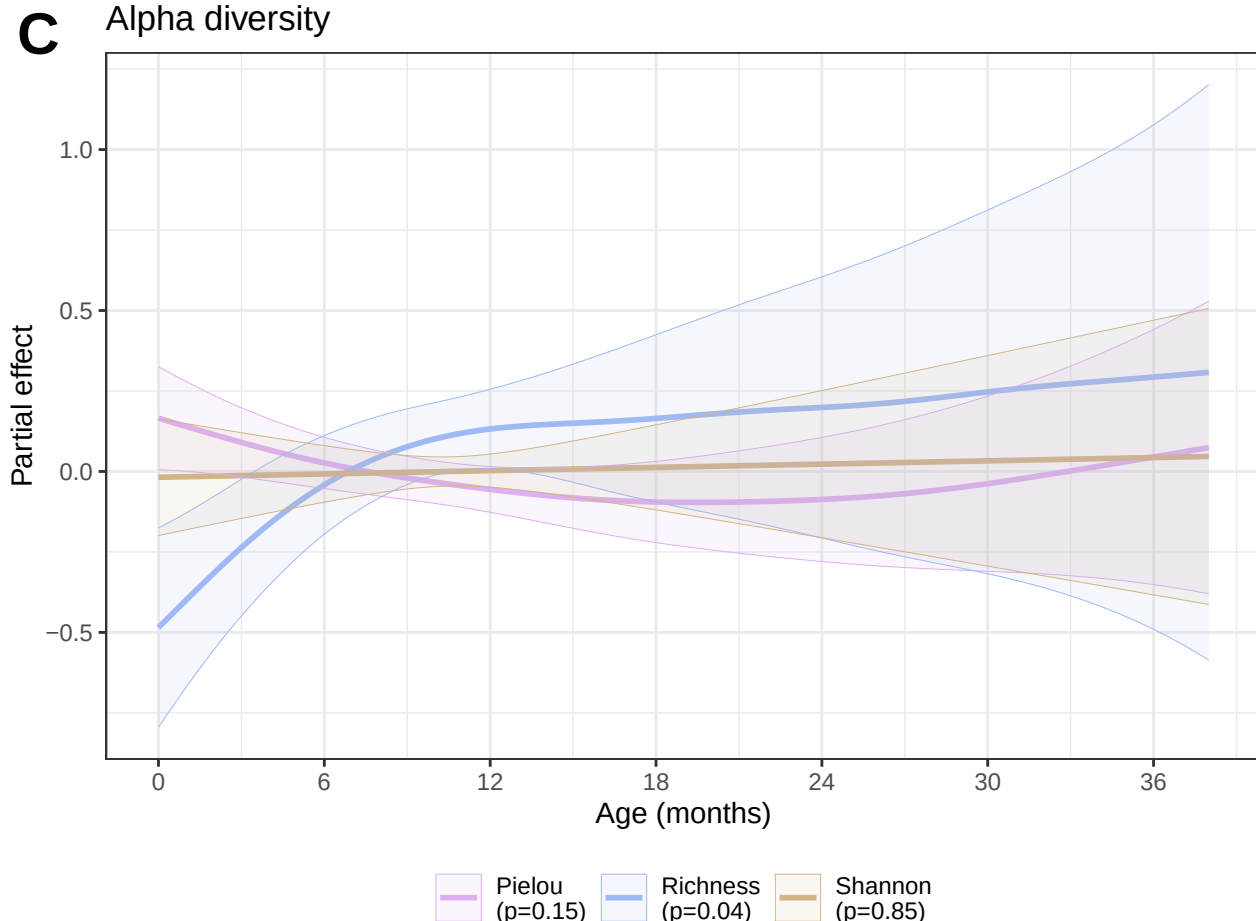
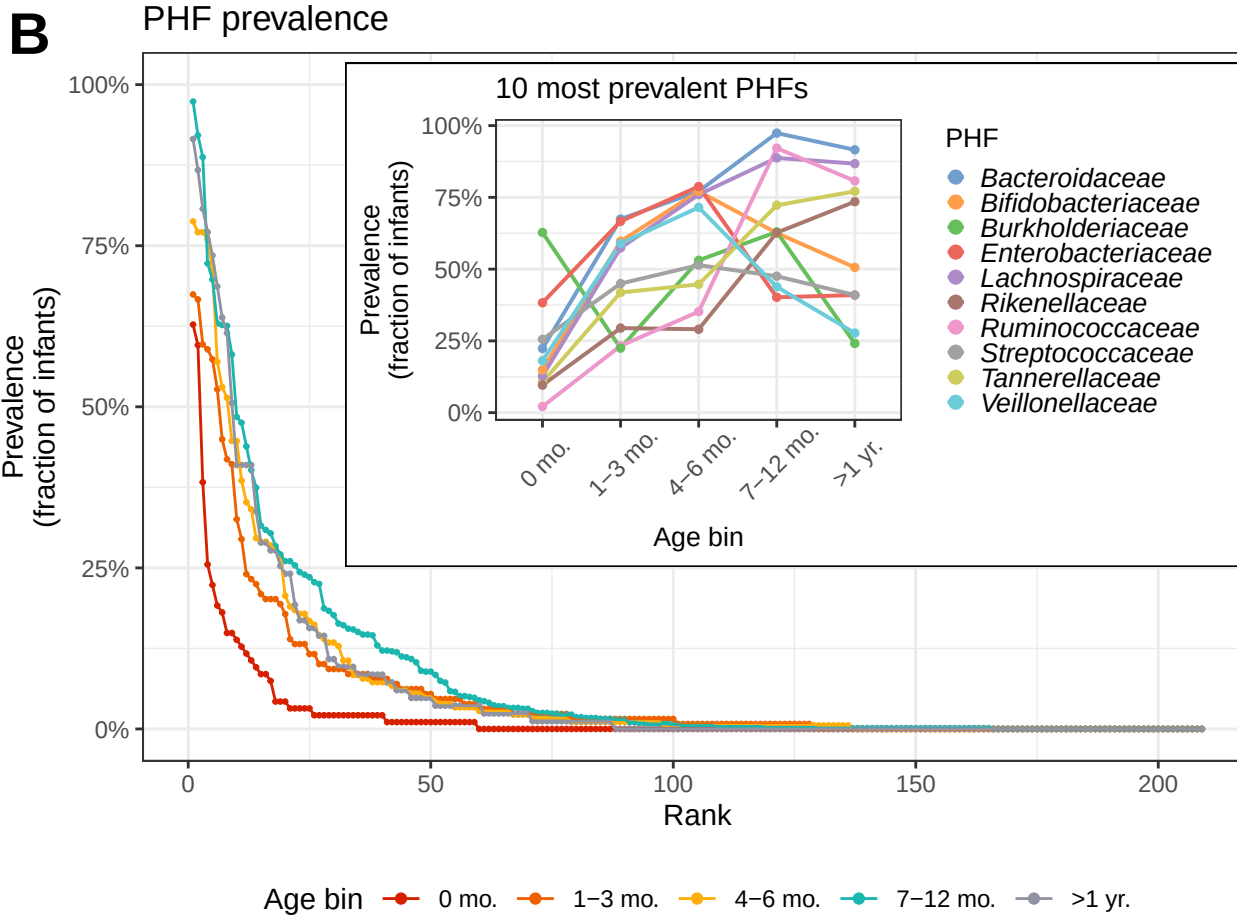
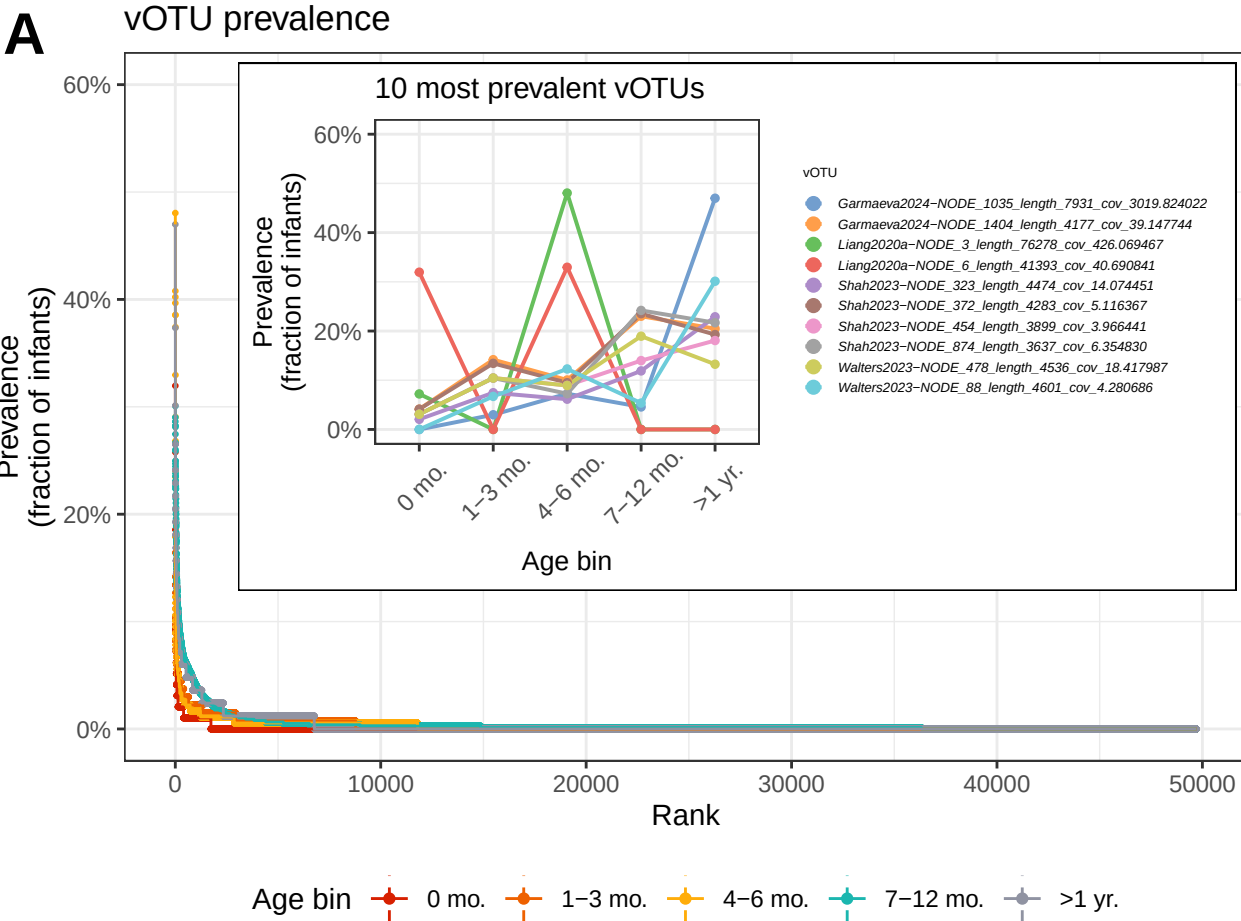


B



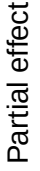
C



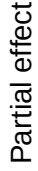


Weighted UniFrac distance ~ sequencing_depth + study + age		
Effect	R ²	P-value
Sequencing depth	0.009	0.988
Study	0.136	0.011 *
Age (months)	0.001	0.032 *
Subject ID	0.488	–
PERMANOVA (adonis2) with 999 permutations		

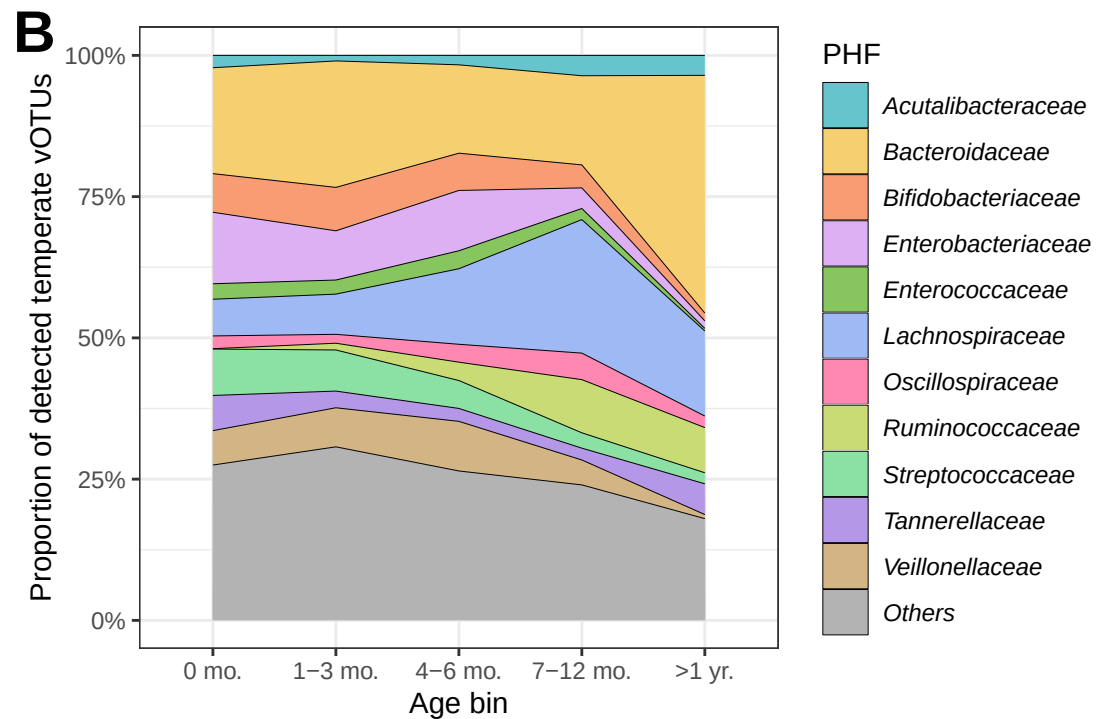
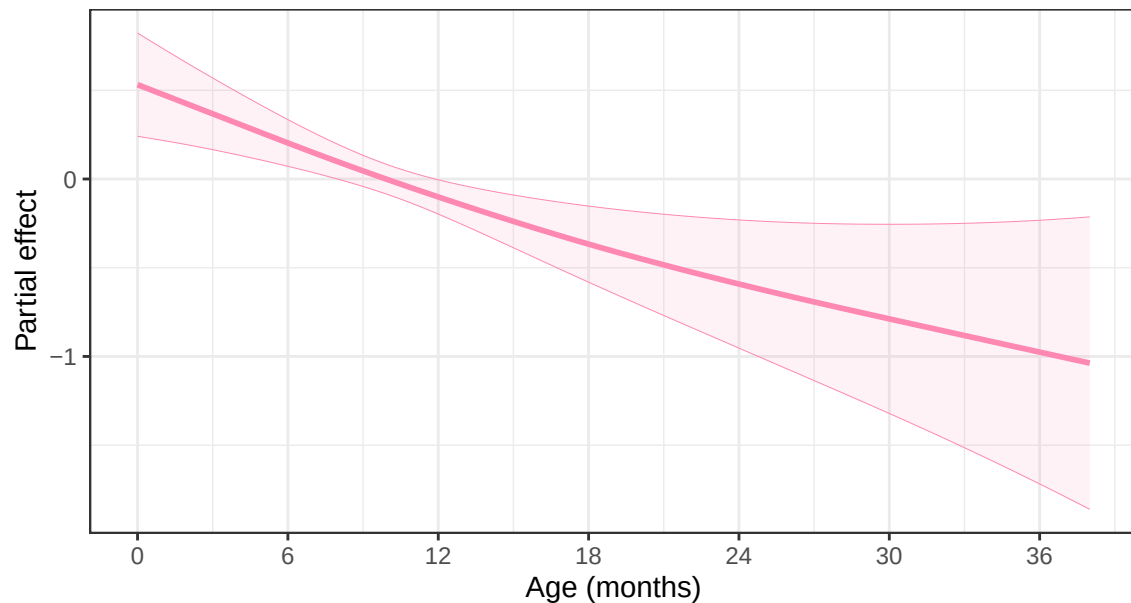
A



B



Dataset  Train  Test

A Temperate phage relative abundance

AMG enrichment among prevalent PHFs

