Mother-Infant Gut Viruses and their Bacterial Hosts: Transmission Patterns and Dynamics during Pregnancy and Early Life

Sanzhima Garmaeva[1*], Trishla Sinha[1*], Anastasia Gulyaeva[1&], Nataliia Kuzub[1&], Johanne E Spreckels[1], Sergio Andreu-Sánchez[1,3], Ranko Gacesa[1,2], Arnau Vich Vila[1,2], Siobhan Brushett[1,4], Marloes Kruk[1], Lifelines NEXT cohort study[§], Jackie Dekens[1,5], Jan Sikkema[5], Folkert Kuipers[3,6], Andrey Shkoporov[7,8], Colin Hill[7,8], Sicco Scherjon[9], Cisca Wijmenga[1], Jingyuan Fu[1,3], Alexander Kurilshikov[1]and Alexandra Zhernakova[1#]

[1]Department of Genetics, University of Groningen, University Medical Center Groningen, Groningen, the Netherlands

[2]Department of Gastroenterology and Hepatology, University of Groningen, University Medical Center Groningen, Groningen, the Netherlands

[3]Department of Pediatrics, University of Groningen, University of Groningen, University Medical Center Groningen, Groningen, the Netherlands

[4]Department of Health Sciences, University of Groningen, University Medical Center Groningen, Groningen, the Netherlands

[5]University Medical Center Groningen, Center for Development and Innovation

[6]European Research Institute for the Biology of Ageing (ERIBA), University of Groningen, University Medical Center Groningen, Groningen, the Netherlands

[7]APC Microbiome Ireland, University College Cork, Cork, Ireland

[8]School of Microbiology, University College Cork, Cork, Ireland

[9]Department of Obstetrics and Gynecology, University of Groningen, University Medical Center Groningen, Groningen, the Netherlands

[*]Shared first authors: SG and TS

[&] Shared second authors: AG and NK

[#]Corresponding author: Alexandra Zhernakova

Highlights

- Longitudinal characterisation of the gut microbiome and virome in 30 mothers during pregnancy, at birth and 3 months after birth and in 32 infants from birth across the first year of life.
- The maternal gut bacteriome changes from the first to the second trimester and then remains stable through birth and the first 3 months after birth.
- The maternal gut virome remains stable during late pregnancy, birth and the first 3 months after birth.
- The infant gut virome is highly dynamic during the first year of life and is shaped by infant feeding mode and place of delivery.
- The infant gut harbours more temperate bacteriophages than the maternal gut, but their relative abundance decreases with increasing infant age.
- Gut viral strains and their bacterial host strains are co-transmitted from mothers to their infants.
- Gut viral strains are transferred from mother to infant around birth directly or via transfer of their bacterial hosts followed by the induction of prophages.

Abstract

Early development of the gut ecosystem is crucial for lifelong health. While infant gut bacterial communities have been studied extensively, the infant gut virome remains under-explored. We longitudinally assessed the composition of gut viruses and their bacterial hosts in 322 total metagenomes and 205 metaviromes from 30 mothers during and after pregnancy and from their 32 infants during their first year of life. While the maternal gut virome composition remained stable during late pregnancy and after birth, the infant gut virome was dynamic in the first year of life and contained a higher abundance of active temperate phages compared to the maternal gut viromes. The infant gut virome composition was also influenced by infant feeding mode and place of delivery. Lastly, we provide evidence of viral-bacterial strains co-transmission from mothers to infants, demonstrating that infants acquire some of their virome from their mother's gut.

Introduction

The human early-life gut ecosystem has garnered much interest in recent years because of its links to health and disease later in life, but core aspects of its origin and development remain poorly understood.[1] Previous studies have characterised the development of the infant gut microbiome through the first 2–3 years of life, after which the gut microbiome reaches a state of high microbial richness and diversity that is similar to that of an adult[2–6]. While the focus of research thus far has been the developing gut bacteriome, the gut ecosystem also comprises viruses, archaea and eukaryotes, whose role in the early gut ecosystem is perceived to be very important but whose composition and development over time has received little attention.

Microbes from the maternal gut, skin and vaginal tract have been described as sources of the infant gut microbiota[7,8], and recent studies provide increasing support for the maternal gut bacterial reservoir as a key source of microbes transmitted from mothers to infants[9–15]. While mother-to-infant transmission of human viruses such as human immunodeficiency virus, cytomegalovirus and herpes simplex virus has been established in the context of maternal and infant morbidity[16], little is known about the transmission of bacteriophages (bacteria-infecting viruses) from the maternal to the infant gut. Studies of viral transmission have been hindered by difficulties in isolating and annotating metaviromes[17]. As environmental studies have demonstrated that bacteriophages are key players in the modulation of bacterial communities[18,19], it is crucial to study them in the context of the developing human gut ecosystem as the bacterial community is established in the months following birth. To this end, a crucial Liang et al. study examining virus-like particle (VLP) data representative of metaviromes in 20 healthy infants provided evidence that the bacteriophages colonising the infant gut arise from excisions from pioneering infant gut

bacteria[20]. However, the origin of these pioneering bacteriophages, as well as their hosts and possible roots in the maternal gut ecosystem, have remained elusive. In the limited number of studies to examine both maternal and neonatal samples, infant faecal samples collected within 4 days postpartum shared 15% of their viruses with the respective maternal samples[21]. Another study, which focused on bifidobacterial phages, showed that infants acquire Bifidobacterium phages from their mother[7]. Contrastingly, a recent study concluded that maternal exposure does not directly impact the development of the infant gut virome[22].

In this study, we investigate the composition of the maternal and infant gut virome during pregnancy and the first year after birth, associate it to host and environmental factors and investigate whether infants and their mothers share some viruses. To do so, in 30 mothers and their 32 infants, including two twin pairs, we sequenced 322 total metagenomes obtained by isolating total microbial DNA from stool and 205 metaviromes obtained using a VLP-enrichment isolation protocol. We found that the composition of the infant virome is highly dynamic during the first year of life and remains very different from the maternal virome at 1 year of age. In contrast, the maternal virome is relatively stable during late pregnancy and after birth. At early timepoints, the infant viromes are dominated by active temperate bacteriophages, the abundance of which decreases over time. Lastly, we show evidence for the transmission of viruses from mothers to infants in related mother-infant pairs, indicating that infants derive some of their virome from their mothers.

Results

Study population

We profiled the gut microbiome (primarily referred to as the bacteriome) in 322 metagenome samples and the double-strand DNA (dsDNA) gut virome in 205 metavirome samples from 30 mothers and their 32 term-born infants (including 2 twin pairs) collected longitudinally from pregnancy to 12 months after birth (Fig. 1a; Supplementary Fig. 1a,b). The infants had a median birth weight of 3705 g (range 2462–5055 g). A significant majority of the children, 87.5% (28 infants), were delivered vaginally, and 28.1% (9 infants) were born at home (Supplementary Fig. 1c). Median maternal age at childbirth was 32 years (range 24–40 years). Breastfeeding behaviour showed a gradual decline over the initial 3 months of life (Supplementary Fig. 1d,e). In the first month, half of the infants were exclusively breastfed, falling slightly to 43.8% in the second month and to 33.3% by the third month. After processing all samples through our metagenome and metavirome pipelines (Methods), we characterised the maternal gut virome during pregnancy, at birth and in the first 3 months after birth and the infant gut virome over the first year of life, along with the predicted bacterial hosts. We then related the infant virome composition with feeding mode and birth-related factors and described the infant virome acquisition by means of transmission from the mother.

The infant gut virome is dynamic during the first year of life

To characterise the infant virome directly after birth, we first attempted to sequence the VLP-enriched metavirome and total microbial DNA from meconium. However, none of the 17 virome DNA isolations from meconium could be sequenced, and only 7% (n=2) of the total metagenomic samples yielded viable sequencing data (microbial read depth >5 million reads).

These outcomes highlight the remarkably low microbial biomass present in meconium samples, further reinforcing the prevailing consensus that the gut remains sterile before birth.[23,24]

To investigate temporal changes in the infant virome after birth across the first year of life, we sequenced metaviromes from five timepoints (months 1, 2, 3, 6 and 12) and total metagenomes (mainly comprising bacteria) from six timepoints (months 1, 2, 3, 6, 9 and 12) from faecal samples from 32 infants. From 205 metaviromes, we reconstructed 102,270 species-level viral operational taxonomic units (vOTUs) from all available timepoints. Of these, 10.7% (n=10,959) of vOTUs were infant-specific (detected only in infants) and 76.1% were mother-specific. The average number of vOTUs in infants (889, 95%CI: [719, 1060]) was lower than in mothers (5182, 95%CI: [4294, 6096]; p-value=1.6e-07, beta=-4304). Over the course of infancy, the average number of vOTUs in infants increased from an average of 675 (95% confidence interval (CI, Methods): [299.0, 1234.9]) at age 1 month to an average of 1,352.9 (95% CI: [977.0, 1732.2]) at 12 months. The overall composition of both viruses and bacteria in the infants underwent significant changes over time (Bray-Curtis dissimilarity, NMDS with one dimension, p-value=1.1e-03, beta=-5.3e-04, for virome and p-value=2.0e-07, beta=-1.8e-03, for microbiome), moving towards a mother-like state (Fig. 1b,c). Both the overall virome and bacteriome composition were significantly different between mothers and infants (p-value=1.5e-10, beta=-0.2, for virome and p-value=3.6e-36, beta=-1.3, for microbiome).

We investigated the temporal changes in the maternal virome at five different timepoints (month 7 of pregnancy, birth and months 1, 2 and 3 after birth) and observed that the maternal virome did not change significantly with time (p=0.6, beta=-0.005). The bacterial composition also did not change significantly during late pregnancy, birth and the first 3

months after birth, but we did observe a significant change in the bacterial composition from month 3 to month 7 of pregnancy, which agrees with previous findings[25,26] (p-value=2.3e-4, beta=0.02, Fig. 1d, e).

Gut viral and bacterial alpha diversity were higher in mothers than in infants in the first year of life (p-value=1.1e-09, beta=2.5; p-value=5.9e-29, beta=1.8). The alpha diversity of both the infant gut virome and bacteriome increased with infant age, approximating but not reaching the alpha diversity of mothers by 1 year of life (p-value=4.1e-4, beta=4.7e-03, Fig. 2a; p-value=1.6e-12, beta=2.7e-03, Fig. 2b). However, even at age 1 year, there was still a significant difference between the viromes of infants and their mothers (p-value=1.7e-3, W-statistic=802). The alpha diversity of the maternal gut virome and bacteriome showed no significant changes, remaining stable throughout pregnancy and after delivery (p-value=1, beta=4e-04; p-value=0.3, beta=1.1e-02, Fig. 2a,b).

We then sought to determine whether the increase in viral alpha diversity in infants over time was a result of retention of the initial coloniser vOTUs paired with the introduction of new species over time. To address this, we investigated the persistence of the viruses initially detected in the infant gut at month 1. By month 2, an average of 37.8% (95% CI: [19.1, 58.8]) of the vOTUs present at month 1 were retained, and these retained vOTUs on average accounted for 42.2% of vOTUs detected at month 2 (Fig. 2c). Both the proportion of retained month 1 vOTUs and the percentage of richness these vOTUs occupied per sample decreased significantly over time, along with their relative abundance (p-value=1.3e-02, beta=-2.8e-02; p-value=9.1e-04, beta=3.4e-02; p-value=7.0e-3, beta=-3.6, Fig. 2c,d)). In contrast, the maternal virome remained relatively stable compared to that of infants, as birth and postpartum samples, on average, retained 42.9% of the vOTUs detected at month 7 of pregnancy, with the relative abundance of these retained vOTUs accounting for on average

for 74.7% (Fig. 2c,d). We observed similar trends for the bacteriomes, i.e., a dynamic infant bacteriome over the first year and a stable maternal bacteriome (Supplementary Fig. 2g-j). In summary, the infant gut virome and microbiome were highly dynamic during the first year of life, whereas the maternal gut virome and microbiome were relatively stable.

Previous studies have demonstrated that adults possess a unique and stable collections of viruses, termed the personal persistent virome (PPV), that plays a crucial role in shaping the diversity and stability of the virome between different individuals[17,27]. Considering the highly dynamic nature of the infant gut, we investigated whether the development of the PPV begins during infancy. For this analysis, we focused on the 14 infants for whom at least three stool sample metaviromes were available and categorised the vOTUs present in these infants into PPVs, i.e., vOTUs present in ≥75% of an individual's samples. The remaining, much-less-stable part of the virome can include phages of low abundance, those infecting transient microbiota members and plant viruses of dietary origin, which are together termed the transiently detected virome (TDV) [27]. We defined the TDV as those present in <75% of an individual's samples. In infants, PPVs were individual-specific, as 50.5% of all infant PPVs (n=1,054) were attributed only to a single infant. PPVs in infants accounted for an average of 16.4% (95% CI: [11.7, 21.9]) of the vOTUs present per sample (Fig. 2g), whereas the TDV accounted for an average of 83.6% (95% CI: [78.5, 88.0], Fig. 2g). The average cumulative relative abundance of the PPV per sample was 35.1% (95% CI: [26.6; 44.6]) and lower than that of the TDV (p-value=8.8e06, beta=-29.8, Fig. 2h). However, it must be noted that infant metaviromes showed a large amount of inter-individual variation in the fractions occupied by PPVs and TDVs. In contrast to infants, the PPVs in mothers accounted for an average of 43.8% (95% CI: [39.2, 48.0]) of the vOTUs detected per sample, with the cumulative

167 relative abundance per sample accounting for an average of 72.3% (Supplementary Fig. 2c,
168 d).

169 Similar to the infant gut metaviromes, transiently detected bacteriomes (TDB) were
170 the most prevalent in the infant gut bacteriomes (mean 67.6% per sample, p-value=2.1e-62,
171 beta=35.2, Supplementary Fig. 2a). However, the cumulative relative abundance of personal
172 persistent bacteria (PPBs) reaching on average 68.4% (95%CI: [63.5, 73.1]) outnumbered the
173 relative abundance of TDB on average accounting for 31.8%, 95%CI: [26.9, 36.8] (p-
174 value=3.4e-23, beta=36.6, Supplementary Fig. 2b). The maternal bacteriome, like the
175 maternal virome, showed remarkable conservation and stability. It was dominated by PPBs at
176 both the richness (p-value=2.1e-131, beta=-53.4, Supplementary Fig. 2e) and abundance (p-
177 value=7.5e-238, beta=-92.7, Supplementary Fig. 2f) levels. We thus concluded that the infant
178 gut ecosystem contains persistent viruses, but they do not fully define the infant gut virome
179 due to the large variability and high virus turnover with time.
180
181 **Virus–host interactions in the maternal and infant gut**
182 We next predicted bacterial hosts for viruses based on vOTU representatives using the iPHoP
183 framework [28]. We could assign 85,135 (82.9%) vOTUs to their 4,572 bacterial hosts at the
184 species level and 68,299 vOTUs to 826 host genera (Fig. 2i, j; Methods). We then observed
185 that the composition of the virome based on its predicted host closely resembles the bacterial
186 composition and that the dynamics of the relative abundance of viruses closely mirror that of
187 the host genera (Supplementary Fig. 3a-d). For example, over the first year of life in infants,
188 we observe an increase in the relative abundance of bacteriophages predicted to infect the
189 genus *Faecalibacterium* (Supplementary Fig. 3c, FDR=2.1e-12, beta=0.5) and a decrease in
190 bacteriophages infecting bacteria from *Bifidobacterium* (Supplementary Fig. 3c, FDR=1.8e-4,

beta=-0.3) and *Klebsiella (*Supplementary Fig. 3c, FDR=3e-04, beta=-0.2) genera. We also see similar trends for the abundance of corresponding bacteria (Supplementary Fig. 3b,d). We next checked if the predicted hosts of the PPVs were also, in fact, found in the PPBs, and we found that the overlap between predicted hosts of PPVs (210 species) and taxonomy of PPBs (69 species) was 35 species (Supplementary Fig. 3e). We noticed that despite the individual specificity of PPVs at the vOTU level, the majority of PPVs were predicted to infect bacterial species from the genus *Bacteroides* (56.3%) and *Phocaeicola* (10.1%).

To further establish the bacteria-virus dynamics, we focused on temperate bacteriophages because they have the ability to integrate their DNA into the host bacterium's genome, potentially influencing the bacterial phenotype and contributing to long-term interactions between bacteria and viruses. The relative abundance of active temperate bacteriophages (detected in VLP-enriched metaviromes) in infants was higher than that in mothers (p-value=1e-09, beta=22.2, Fig 3a). The average relative abundance of temperate phages was high in the first 3 months (45.9, 95% CI: [36.6, 55.4]) and decreased drastically by 6 months after birth (22.6, 95% CI: [13.3, 31.8], p-value=2.9e-04, beta=-8.7e-02, Fig. 3a). By 12 months, the relative abundance of active temperate bacteriophages was only slightly higher than that observed in maternal samples (on average 20.0, 95% CI: [13.4, 27.4] in infants vs on average 11.0, 95% CI: [9.8, 12.3] in mothers, p-value=0.045, W-statistic = 1731). Given that the majority of temperate bacteriophages in the adult gut are in the form of prophages[27], we compared the percentage of temperate bacteriophages in mothers and infants using the vOTUs detected in whole metaviromes, created by aligning the reads from whole metagenomes aligned to vOTU database reconstructed from metaviromes, thus representing the prophage content of the virome. In whole metaviromes, we again observed that the relative abundance of temperate bacteriophages in the form of prophages was slightly higher

in infants than in mothers (p-value=0.04, beta=2.1, Supplementary Fig. 4a). Overall, our findings demonstrate a higher relative abundance of temperate bacteriophages, both in active metaviromes and as prophages, in infants compared to mothers, indicating their large potential impact on the dynamic interplay between bacteria and viruses in the infant gut.

Infant gut virome composition is influenced by infant feeding mode and place of delivery

We next explored the effects of maternal age, infant sex, place of birth, birthweight, feeding mode and gestational age on infant gut viral and bacterial composition. The alpha diversity of vOTUs was associated with infant feeding mode throughout the first year of life, with exclusively formula-fed infants showing higher alpha diversity than breastfed infants (p-value=0.02, beta=0.9, Fig. 3b). This effect held true even after correcting for bacterial diversity ((p-value=0.01, beta=0.8). Feeding mode was also associated with the richness of active temperate bacteriophages, as exclusively formula-fed infants consistently showed a higher richness of temperate bacteriophages over time compared to breastfed infants (p-value=3.0e-3, beta=45.4, Fig. 3c), also after correcting for viral alpha diversity (p-value=0.03, beta=22.3). The richness of temperate phages in whole metaviromes also showed a significant association with formula feeding even after correction for bacterial alpha diversity (p-value=0.02, beta=38.8, Supplementary Fig. 4b). We then looked at the effect of feeding mode on vOTUs aggregated on the basis of their host. Intriguingly, for some phages there are indications that the association can be attributed to the phage itself and not just its host. Compared to breastfed infants, exclusively formula-fed infants had more active temperate phages of bacteria from the genus *Bacteroides*, with the most significant association belonging to *Bacteroides fragilis* (FDR=2.1e-3, beta=1.4, Fig. 3d), *Phocaeicola vulgatus* (FDR=0.01, beta=1.5), and *Bacteroides caccae* (FDR=0.02, beta=0.7), even after correction for

the abundance of their bacterial hosts. The hosts of these bacteriophages were not associated with feeding mode themselves (*B. fragilis*: p-value=0.3, beta=2.0; *P. vulgatus*: p-value=0.9, beta=-0.4; *B. caccae*: p-value=0.9, beta=-0.3). When corrected for both the host abundance and the estimated number of prophages from whole metaviromes, the differential prevalence of active temperate phages of *B. fragilis* remained significantly higher in formula-fed infants (FDR=0.02, beta=1.1). This suggests that formula feeding might be associated with the induction of bacteriophages in *B. fragilis*, independent of changes in bacterial abundances.

Since only two mothers gave birth to their infants by caesarean section, we did not have sufficient power to explore the effect of birth mode on virome composition (Supplementary Fig. 1c). However, as 28% of the infants were born by vaginal delivery at home, we investigated if home versus hospital delivery was associated with specific vOTUs, aggregated by the microbial host or microbial taxa themselves. Here we observed that the bacterial species *Akkermansia muciniphila* (Supplementary Fig. 4c) was more abundant in infants born at home compared to those born at the hospital (FDR=0.04, beta=2.5). Concomitantly, the phages of *A. muciniphila* were also more abundant in infants delivered at home (p=0.02, beta=1.5, Supplementary Fig. 4d). This shows the importance of accounting for the birth environment when considering the early development of the infant virome and its interactions with its host(s).

Infants can acquire gut viruses and bacterial hosts from the maternal gut

Despite the large difference in adult and infant gut microbes, related mother-infant pairs have previously been shown to share gut bacterial species[11,12], but limited information is available about sharing of viruses between maternal and infant guts. We therefore investigated if the gut ecosystem of mothers and their infants harbours the same viruses. We first compared the

263 percentage of infant vOTUs that were shared with the pooled pre-birth (Month 7, Birth) and
264 pooled post-birth (Month 1, 3) maternal samples. Infants shared a higher percentage of
265 vOTUs with post-birth maternal samples (32.3%) compared to the pre-birth maternal samples
266 (26.6%) across all infant timepoints (p=0.04, beta=3.3e-02, Fig. 3e). As pioneer viruses in the
267 infant gut are thought to be primarily temperate phages induced from the first gut bacterial
268 colonisers[20], we next assessed the sharedness of maternal to infant vOTUs while accounting
269 for prophages detected in whole metaviromes. Notably, sharedness increased significantly
270 when considering prophages in both the maternal and infant gut (p=0.001, mean increase of
271 4.9%, Fig. 3f, Supplementary Fig. 4f). We then sought to explore whether infant feeding mode,
272 place of delivery and infant gestational age influenced the percentage of shared vOTUs
273 between mother and infant, but this was not the case. On average, the relative abundance of
274 vOTUs shared with maternal gut virome in infants was 32.7% (95% CI: [28.1, 37.4]). These
275 findings suggest that sharing of vOTUs between mothers and infants is more likely attributed
276 to cohabitation rather than direct seeding of these viruses during birth. Additionally, the
277 higher degree of sharedness observed when considering the whole metavirome provides
278 support for the notion that the presence of shared bacteria containing prophages contributes
279 to the colonisation process in infants.

280 We next sought to see if there were cases of strain-sharing within mother-infant pairs.
281 To do so, we selected vOTUs that were shared between mother and infant and passed strict
282 cut-offs for completeness and coverage (Methods), resulting in 51 vOTUs for downstream
283 analysis. For these 51 vOTUs, we reconstructed consensus sequences from quality-trimmed
284 reads aligned to vOTUs from metagenomes and metaviromes and calculated pairwise genetic
285 distances (Kimura) between consensus sequences corresponding to the same vOTUs. We
286 compared these genetic distances between the viruses shared across an infant and their own

mother as compared to unrelated mothers. We found that, for 28 of the 51 vOTUs (55%), the genetic distance between related mother-infant sample pairs was significantly lower than that of unrelated mother-infant sample pairs (Fig. 4a). We then defined strain-sharing using a distance cut-off estimated assuming strain retention in longitudinal samples (Methods, Supplementary Fig. 4g). In 26 of these 28 viruses, we observed 841 strain transmission events between samples from related mothers and infants (Methods). Of the 26 transmitted viruses, 23 were shared with higher frequency within related mother-infant pairs compared to unrelated pairs (FDR<0.05, Fig. 4b). Seven of these were found among PPVs in 50% of infants with more than 3 timepoints available. These persistent transmitted colonizers were predicted to infect bacteria from genera *Phocaeicola*, *Bacteroides and Parabacteroides*. Next, we explored the transmission of the predicted bacterial hosts of the shared viruses. For the 26 transmitted viruses, we constructed 37 strains of their 29 bacterial hosts in both maternal and infant faecal samples (see Methods). Our findings indicate that, for 26 of the 30 (86.7%) reconstructed bacterial host strains present in both mother and infant, the distances between related mother-infant pairs were lower than those observed between unrelated mother-infant pairs (Fig. 4c, FDR<0.05). Of those 26 bacterial strains, 24 were shared with higher frequency within related mother-infant pairs compared to unrelated pairs (Fig. 4d, FDR<0.05, Supplementary Fig. 5). These bacterial strains mostly belong to the genera *Alistipes, Bacteroides, Bifidobacterium, Faecalibacterium, Parabacteroides, Phocaeicola* and *Sutterella.*

To establish whether viral transmission occurred during or after birth, we investigated if there was a difference in viral strain-sharing between infant samples (at all timepoints) and maternal pre-birth (pregnancy month 7, birth) versus post-birth (month 1, month 3) samples. Here, we found one significant difference in viral strain-sharing between pre-birth and post-

311 birth samples (FDR < 0.05, Table S39). In concordance with this, the host of this virus,
312 *Parabacteroides distasonis,* was also shared with a higher frequency between infant samples
313 and post-birth maternal samples as compared to pre-birth maternal samples (p-value < 0.05),
314 suggesting a higher probability that this bacterium and its phage were transmitted after birth.
315 Subsequently, we tested whether bacteriophages were preferably co-transmitted alongside
316 their bacterial hosts, as opposed to other bacteria, by correlating their strain-sharing events
317 in concurrent samples (Methods). Our observations revealed that bacteriophages were
318 predominantly co-transmitted in conjunction with their bacterial hosts (p-value=0.01, Fig. 5a–
319 b, Table S41). An example of one such bacteriophage is L85266_LS0, whose host *Bacteroides*
320 *uniformis* shows a very similar topological pattern in its phylogenetic tree (Fig 5c). We also
321 found evidence for non-random co-occurrence of phages and their bacterial host in 14/32
322 virus–host pairs (linkage FDR<0.05, Methods, Table S40), indicating a common co-
323 transmission mechanism throughout different mother-infant pairs. This co-occurrence was
324 more often seen between phage–host pairs than between phage–unrelated bacterium pairs
325 (Fisher test, p-value=0.02, Fig. 5b). This co-transmission was observed for multiple species of
326 the genus *Bacteroides (7), Bifidobacterium bifidum* and *Sutterella wadsworthensis*.
327
328 **Possible mechanisms for the origin of the early-life virome**
329 Having established that there are cases of viral strain transmission from mother to infant, we
330 sought to explore the mechanisms underlying the colonisation of bacteriophages in the infant
331 gut. Among the 26 transmitted viruses, 21 were identified as virulent bacteriophages. Based
332 on their predicted lifecycle, it is likely that they were transmitted through direct seeding of
333 VLPs from the maternal to infant gut.

We next focused on investigating the origin of temperate phages in the infant gut. One of the strongest cases of co-transmission between bacteriophages and bacteria was *B. bifidum* and its predicted temperate bacteriophage L34922_LS1 (r=1, p-value=0.02, Fig. 5a). The phylogenetic trees of this temperate phage and its bacteria are topologically very similar (Fig. 6a). We postulated that the high co-transmission rate we observed might be attributed to the temperate nature of L34922_LS1, which enables it to integrate and be transmitted within the genome of its host. As we observed L34922_LS1 in both metaviromes and metagenomes (Fig. 6b), it could also suggest possible phage induction from its transmitted host. To investigate this hypothesis, we initially reconstructed the genome of *B. bifidum* from metagenomes in which the presence of L34922_LS1 was detected. Next, we mapped the L34922_LS1 genome sequence to the genome of *B. bifidum* and observed a high identity (>99%) and coverage (100%) for the L34922_LS1 sequence (Fig. 6c), which confirmed that this phage could be observed in the prophage form within the *B. bifidum* genome. Additionally, we detected both the integrase and viral recombination genes in the L34922_LS1 sequence (Fig. 6d), further confirming the phage's ability to integrate into the host genome and undergo recombination. Metagenomic read alignment to the genome of *B. bifidum* containing L34922_LS1 revealed that 96.2% of *B. bifidum* strains do not contain L34922_LS1 at the determined region of prophage insertion (Fig. 6e). However, in *B. bifidum* genomes detected in samples from mother-infant pairs where L34922_LS1 transmission was observed, the prophage insertion region was sufficiently covered by reads. Metaviromic read-alignment profiles demonstrated a consistent increase in read coverage at the region of prophage insertion compared to the rest of the *B. bifidum* genome, confirming prophage induction in infant and maternal samples (Fig. 6f). Overall, these observations suggest that L34922_LS1 originated from the *B. bifidum* of the mother.

We next attempted to find the origin of temperate bacteriophages that were not shown to be significantly transmitted from mother to infant. One of these was a temperate bacteriophage identified in infant samples, L37775_LS1, that is predicted to infect multiple species of the *Bifidobacterium* genus. After mapping the genome sequence of L37775_LS1 to patched *Bifidobacterium* genomes reconstructed from metagenomes concurrent to metaviromes carrying L37775_LS1, we narrowed down the host range to *Bifidobacterium scardovii* (qcov 100%, e-value<0.005), which was absent in maternal samples. Metagenomic and metaviromic read-alignment profiles to the *B. scardovii* genome revealed that L37775_LS1 is present at the indicated prophage region and can be induced from its host (Supplementary Fig. 6 a,b,c). As *B. scardovii* was only present in infant gut metagenomes and metaviromes, our observations suggest that its phage L37775_LS1 does not originate from the maternal gut. Gene annotation showed that, in addition to carrying an integrase gene, L37775_LS1 also carries a CAZyme (Glycosyl hydrolases family 25, Supplementary Fig. 6d), indicating that this phage might be associated with infant feeding.

Discussion

In this study, we characterised the faecal microbiome and virome in 30 mothers and their 32 infants longitudinally during pregnancy, at birth and during the first year of life. To our knowledge, this is the only study to look at the maternal virome longitudinally during pregnancy, birth and after birth. In the maternal total microbiome, we observed a notable shift in composition between the first and second trimesters of pregnancy. This is in line with the findings of Koren et al., who proposed that hormonal shifts, immune system adaptations and dietary variations during pregnancy can impact the composition of gut bacteria[25,26]. During late pregnancy, birth and after birth, however, we observed that the overall composition of the maternal gut microbiome and virome do not change. These results suggest that once established during the second trimester of pregnancy, the maternal gut microbiome and virome remain consistent throughout this critical period of maternal and infant health.

We demonstrated that the infant gut virome during the first year of life was highly dynamic, and while it progressively transitioned to resemble an adult-like virome with time, it was still very different from that of the mother at the age of one year. We show that the infant gut has a high proportion of temperate phages in the first 3 months of life and that this proportion decreases drastically at 6 months. We thus hypothesise that temperate phages are fundamental in seeding the gut virome, most likely through prophage induction of pioneering gut bacteria. However, even at the age of one year, the abundance of prophages still remained higher than that observed in adults, indicating ongoing viral development. Altogether, our results indicate there is a high degree of prophage induction in the gut during the first 3 months of infancy that is followed by stabilisation of the gut environment at a later age as the availability of more bacterial hosts allows for more prophage integration.

Our results highlight the influence of infant feeding mode on infant virome composition. The higher alpha diversity we observed in exclusively formula-fed infants may be attributed to the different composition of formula compared to breast milk. The increased richness of active temperate bacteriophages in exclusively formula-fed infants suggests that formula feeding may provide specific nutrients or environmental conditions that promote the proliferation of temperate phages. The consistent association observed in both active metaviromes and prophages supports the notion that formula feeding has a lasting impact on the acquisition and maintenance of temperate phages in the infant gut. Our investigation into vOTUs grouped by their host bacteria reveals an intriguing finding related to *Bacteroides fragilis*. Although the relative abundance of *B. fragilis* itself was similar between feeding groups, formula feeding was associated with a higher presence of active temperate phages specifically targeting *B. fragilis*. This suggests that formula feeding may induce the production of bacteriophages that target this bacterial species, independent of changes in bacterial abundances. The mechanisms underlying this association warrant further investigation.

Only a few studies have addressed infant gut viromes in relation to maternal viromes[7,21,22]. Duranti et al. focused entirely on the transmission of Bifidobacterium phages from mother to infant gut and showed that these could be transmitted[7]. Our study showed that not only Bifidobacterium phages but numerous phages predicted to infect bacteria from other genera like *Bacteroides* were also transmitted from mother to infant. A recent study by Walters et al. in 53 infants showed that the overall infant gut virome composition was not driven by exposure to mothers but rather by dietary, environmental and infectious factors[22]. However, in this study, no direct comparison was made between infant and maternal gut microbial strains. Another study by Maqsood et al. comparing virus scaffold presence-absence between mother and infant in 28 infant twin pairs showed that, on average, 15% of the infant

virome was shared with their own mother's gut virome[21]. Our study revealed that, despite significant distinctions between the infant and maternal gut viromes, infants shared on average 32.7% of vOTUs with their mothers. This difference in sharedness may be attributed to the fact that our study encompassed longitudinal samples from both mother and infant for a longer timeframe, whereas Maqsood et al. looked at the sharing of vOTUs between mother and infant cross-sectionally around birth. However, to make definitive claims about transmission, one cannot merely rely on virus sequence co-occurrence in maternal and infant viromes, and it is essential to examine the genetic makeup of viral strains, their sequence similarity and to define a strict strain identity discrimination threshold, an aspect not explored in the above-mentioned studies. For vOTUs shared between the infant and maternal guts, we constructed strains and showed that more than half of the strains were shared more frequently within related mother-infant pairs compared to unrelated mother-infant pairs. This shows that, while it might not be the most defining factor of the infant gut virome, infants do share viral strains with their mother's gut. In microbiome transmission studies, determining the directionality of strain inheritance poses a fundamental challenge. Although it is theoretically possible for an infant to acquire a strain and transmit it to the mother, this scenario is improbable given the greater diversity and stability of the adult gut microbiota. The acquisition of a shared strain by both individuals from a common environmental source is another plausible explanation. However, the prevailing consensus in the field suggests that the primary direction of transmission is from mother to infant[12] which is what we assume in this paper. Our findings suggest that some viral and bacterial strains are co-transmitted between related mother-infant pairs, and this holds especially true for species of the genus *Bacteroides, Bifidobacterium bifidum* and *Suttertella wadsworthensis.* We then show examples of how this occurs via direct transmission of bacteriophages and prophage

induction following transmission of their bacterial hosts. We also show how infants obtain some viral strains from induction from bacteria that did not come from their mothers.

Our study has several strengths. Firstly, our quantitative approach (avoidance of amplification techniques during the isolation and sequencing of VLP DNA) allowed accurate quantification of viruses and led to minimal bias in our estimation and characterisation of dsDNA viruses. Furthermore, by utilising both total and viral-enriched metagenomes, we could characterise the whole virome, including prophages. Additionally, our dense longitudinal sampling design in both mothers and infants allowed us to study viral compositional dynamics in critical periods such as pregnancy, birth and post-birth. Our study also addressed previously unstudied factors for maternal-to-infant gut microbial transmission such as the place of delivery. As home deliveries constitute 12.7% of the total deliveries (2018) in the Netherlands and 28% in our samples, we had a unique opportunity to explore the effect of place of birth on the viral and bacterial communities of the gut, although we note that our sample size is small and larger studies are needed to make strong conclusions about this.

Our study has several limitations. Firstly, due to our isolation method, we focused solely on the dsDNA viruses present in the gut and overlooked the single-stranded DNA viruses and RNA viruses. While RNA viruses are typically perceived to have a lower abundance in the healthy human gut, it is crucial for future investigations to delve into this under-studied aspect of the infant virome. Including single-stranded DNA viruses and RNA viruses in future research will shed light on their potential roles within the gut ecosystem. Our sample size was also small, which hampers associations with phenotypes and limits the generalizability of the findings to the whole population. Due to the small number of participants who had C-sections, we could not study the effect of this important factor on the gut virome even though an impact of birth mode on the early-life gut virome has previously been described[29]. Thirdly,

while we use the term “human gut virome” throughout this study to be consistent with earlier studies, we acknowledge that this result may be biased as faecal samples do not accurately take into account the viruses residing in the gut mucosa. Hence our results regarding the gut virome are limited to faecal viromes. Finally, despite our best efforts, we cannot guarantee that the viral scaffold database is free from bacterial contamination.

In conclusion, we characterised the total gut microbiome and gut virome in 30 mothers and their 32 infants from birth to the first year of life using a complementary approach examining both total metagenomes and viral metagenomes. We show that the maternal virome composition does not change significantly during late pregnancy, birth and after birth, whereas the infant gut virome composition is highly dynamic in the first year of life and is influenced by the infant’s feeding mode and place of delivery. Initially, infants' viromes have a high proportion of active temperate bacteriophages, which decrease over time but remain higher than in adults at one year of life. Lastly, we provide evidence of viral and bacterial strain co-transmission from mothers to infants, indicating that infants acquire some of their virome from their mothers. Moving forward, future investigations should focus on elucidating the functional implications of these findings and their potential impact on the long-term health and development of infants.

Methods

Study cohort

The samples for this study were obtained from the Lifelines NEXT cohort, a birth cohort designed to study the effects of intrinsic and extrinsic determinants on health and disease in a four-generation design[30]. Lifelines NEXT is embedded within the Lifelines cohort study, a prospective three-generation population-based cohort study recording the health and health-related aspects of 167,729 individuals living in the Northern Netherlands[26]. In Lifelines NEXT, we included 1,450 pregnant Lifelines participants and intensively followed them, their partners and their children up to at least 1 year after birth. During the Lifelines NEXT study, biomaterials, including maternal and neonatal (cord) blood, placental tissue, faeces, breast milk, nasal swabs and urine are collected from the mother and child at ten timepoints. Furthermore, data on medical, social, lifestyle and environmental factors are collected via questionnaires at 14 different timepoints and via connected devices[27]. The current study is a pilot study of the first samples collected in the Lifelines NEXT project, without prior selection.

Informed consent

The Lifelines NEXT study was approved by the Ethics Committee of the University Medical Center Groningen, document number METC UMCG METc2015/600. Written informed consent forms were signed by the participants or their parents/legal guardians.

Sample collection

Mothers collected their faeces during pregnancy at weeks 12 and 28, very close to birth and during the first 3 months after birth (Fig. 1a). Faeces from infants were collected from diapers by their parents at 1, 2, 3, 6, 9 and 12 months of infant age. Parents were asked to freeze the stool samples at home at -20°C within 10 min of stool production. Frozen samples were then collected and transported to the UMCG in portable freezers and stored in a -80°C freezer until extraction of microbial and viral DNA. For this study, we collected 361 samples for total microbiome analysis.

DNA extraction from total microbiome

Total microbial DNA was isolated from 0.2–0.5 g faecal material using the QIAamp Fast DNA Stool Mini Kit (Qiagen, Germany) using the QIAcube (Qiagen) according to the manufacturer's

instructions, with a final elution volume of 100 µl. Additionally, DNA was extracted from two negative controls consisting of Milli-Q water. The exact same procedure was used for the negative controls. DNA eluates were stored at -20°C until further processing.

DNA extraction from VLPs

Out of 361 faecal samples, 259 were selected for VLP enrichment and VLP DNA extraction based on the amount of faecal material collected. This included maternal samples from 28 weeks of pregnancy, birth and months 1, 2 and 3 after delivery and infant samples at birth and months 1, 2, 3, 6 and 12 after birth (Fig. 1a). To study the gut virome, DNA was extracted from VLPs as described in Shkoporov et al. 2018. Briefly, 0.5 g faecal material was resuspended in a 10 ml SM buffer (50 ml 1 M UltraPure™ 1 M Tris-HCl Buffer, pH 7.5 (Invitrogen™ #15567027); 20 ml 5 M NaCl (Sigma-Aldrich Cat#221465); 8.5 ml 1 M $MgSO_4$ (Sigma-Aldrich Cat#230391); 921.5ml H_2O), then centrifuged at 4,800 rcf for 10 min at 4°C, followed by supernatant collection and repeated centrifugation. The supernatant was filtered twice through a 0.45-µm pore polyethersulfone membrane filter to obtain the VLPs. The VLPs were concentrated with Polyethylene glycol 8000 (Sigma-Aldrich, Cat#P2139), overnight precipitation and purification by chloroform treatment. The resulting fraction was treated with 8 U of TURBO DNase (Ambion/Thermo Fisher Scientific Cat#AM2238) and 20 U of RNase I (Thermo Fisher Scientific Cat#10568930) at 37°C for 1 h before inactivating enzymes at 70°C for 10 min. Subsequently, proteinase K (40 µg, Sigma-Aldrich, Cat#2308) and 20 µl of 10% SDS were added to the samples and incubated for 20 min at 56°C. Finally, VLPs were lysed by the addition of 100 µl of Phage Lysis Buffer (4.5 M guanidinium isothiocyanate (Sigma-Aldrich Cat#50983), 44 mM sodium citrate pH 7.0 (Sigma-Aldrich Cat#C8532), 0.88% sarkosyl (Sigma-Aldrich Cat#5125), 0.72% 2-mercaptoethanol (Sigma-Aldrich Cat#M6250) and incubated at 65°C for 10 min. Nucleic acids were extracted twice from lysates using Phenol/Chloroform/Isoamyl Alcohol 25:24:1 (ThermoFisher Scientific Cat#10308293) treatment followed by centrifugation at 8,000 g for 5 min at room temperature. The resulting aqueous phase was subjected to the final round of purification using the DNeasy Blood & Tissue Kit (Qiagen Cat#69506) with a final elution volume of 50 µl. For the negative controls, four samples of SM buffer alone were run through the VLP enrichment and DNA extraction process. The resulting viral DNA eluates were stored at -20°C until further processing.

Genomic library preparation and sequencing

Faecal microbial DNA and viral DNA samples were sent to Novogene, China, for genomic library preparation and shotgun metagenomic sequencing. Sequencing libraries were prepared using the NEBNext® Ultra™ DNA Library Prep Kit or the NEBNext® Ultra™ II DNA Library Prep Kit, depending on the sample DNA concentration, and sequenced using HiSeq 2000 sequencing with 2 × 150 bp paired-end chemistry (Illumina). On average, 30.2±5.0 million paired-end total metagenome reads, and 27.4±6.9 million paired-end VLP reads were generated per sample. Of 361 samples prepared for total microbiome analysis, 326 were successfully sequenced, and 86% (n=30) of the failed samples were meconium. Following this, low read-depth samples (<5 million reads) and one mislabelled sample were excluded, leaving 322 samples for analysis. For virome analysis, 205 of 255 samples were successfully sequenced. As with the total microbiome, the meconium samples failed sequencing. Three out of four negative metavirome controls failed to be sequenced, and the only successfully sequenced one had 24.4 million paired-end reads and was used to remove contaminating reads from the rest of metaviromes (see below in "Profiling of gut virome composition"). The positive negative control had a high genomic bacteria DNA contamination (36.5%, see below) and consisted mainly of Sphingomonas spp. and Human adenovirus C.

Profiling of total gut microbiome composition

Total microbiome sequencing reads were trimmed, and Illumina sequence adaptor sequences were removed using BBMap (v38.98) [31] and KneadData tools (v0.7.4) [32]and on average resulted in a PHRED quality score of 33. Following trimming, the KneadData-integrated Bowtie2 tool (v2.4.2)[33] was used to remove reads that aligned to the human genome (GRCh37/hg19), and the quality of the processed data was examined using the FastQC toolkit (v0.11.9)[34]. Taxonomic composition of metagenomes was profiled using the MetaPhlAn4 tool with the MetaPhlAn database of marker genes mpa_vJan21 and the ChocoPhlAn SGB database (202103)[35].

Profiling gut virome composition

Metavirome sequencing reads underwent quality trimming and human read removal, as described above. Bacterial contamination of metaviromes was assessed by aligning reads to

the single copy chaperonin gene cpn60 database[36]. On average, metaviromes contained 8.3% (95% CI: [7.2; 9.6]) of bacterial genomic DNA per sample.

We used a *de novo* assembly approach to annotate the composition of the gut virome. Specifically, SPAdes (v3.14.1)[37] was utilised in metagenomic mode (*-meta*) with default settings to perform *de novo* assembly per metavirome. The average number of assembled scaffolds was 283,893 for maternal samples and 103,192 for infant samples. Scaffolds smaller than 1 kbp were removed. Scaffolds that were at least 1 kbp underwent rigorous filtering per sample for the following gut virome annotation. The Open Reading Frames (ORFs) in these scaffolds were predicted using Prodigal v2.6.3[38] in metagenomic mode. Ribosomal proteins were identified using a BLASTp [39] search (e-value threshold of 10^{-10}) against a subset of ribosomal protein sequences from the COG database (release 2020). We used a Hidden Markov Model (HMM) algorithm (hmmsearch from HMMER v3.3.2 package)[40] to compare amino acid sequences of predicted protein products against the HMM database Prokaryotic Virus Orthologous Groups (pVOGs)[41]. Hits were considered significant at an e-value threshold of 10^{-5}. To detect viral sequences, VirSorter v1.0.3 [42] was run with its expanded built-in database of viral sequences ('–db 2' parameter) in the decontamination mode (--virome). Scaffolds larger than 1 kbp were considered viral if they fulfilled at least one of six criteria, similar to those described previously : (1) BLASTn alignments to a viral section of NCBI RefSeq (release 211) with e-value≤10^{-10}, covering >90% of sequence length at >50% Average Nucleotide Identity (ANI), (2) having at least three ORFs, producing HMM-hits to the pVOG database with an e-value≤10^{-5}, with at least two ORFs per 10 kb of scaffold length, (3) being VirSorter-positive (all six categories, including suggestive), (4) being circular[43], (5) BLASTn alignments (e-value≤10^{-10}, >90% query coverage, >50% ANI) to 1,489 Crassvirales dereplicated sequences (99% ANI and 85% length) larger than 50 kbp from the NCBI database (taxid:1978007) and published datasets[44–47] and (6) being longer than 3 kbp with no hits (alignments >100 nucleotides, 90% ANI, e-value of 10^{-10}) to the nt database (release 249). 281,789 scaffolds fulfilled at least one of these six criteria.

To remove putative cellular contamination from the virus sequences, scaffolds meeting the filtering criteria were dereplicated at 99% ANI with all negative control scaffolds with no filtration applied other than the size of the scaffold (larger than 1 kbp) using CheckV at 85% alignment fraction (relative to the shorter sequence). Sequence clusters containing negative control scaffolds were excluded from further consideration. The remaining 280,633

putative virus scaffolds were dereplicated at 95% ANI and 85% length to represent vOTUs at the species level.[48] The resulting 110,526 vOTU representatives were screened for the presence of ribosomal RNA (rRNA) genes using a BLASTn search in the SILVA 138.1 NR99 rRNA genes database[49] with an e-value threshold of 10^{-3}. An rRNA gene was considered detected in a scaffold if the gene and the scaffold produced a hit covering >50% of the gene length. Additionally, vOTU representatives were clustered with 1,489 dereplicated *Crassvirales* sequences larger than 50 kbp and the genomes of the reference database "ProkaryoticViralRefSeq211-Merged" using vConTACT2 v0.11.3 with default parameters[50]. Sequences assigned the status 'Overlap', 'Singleton' and 'Outlier' by vConTACT2 were treated as genus viral clusters consisting of a single scaffold in all subsequent analyses. The resulting viral clusters (VCs) of putative viral scaffolds were subjected to a second decontamination procedure based on the following criteria. VCs were excluded if any of the cluster members: (1) contained an rRNA gene, (2) contained ≥1 ribosomal protein gene and < 3 pVOGs per 10 kb, was VirSorter-negative and non-circular and (3) contained >3 ribosomal protein genes. Specific members of VCs were retained if they satisfied any of the following criteria: (1) circular and had ≥1 pVOGs, (2) circular and VirSorter-positive or (3) VirSorter-positive and had no ribosomal protein genes. The final curated database of virus sequences generated from our dataset included 102,280 vOTU representatives ranging in size from 1 kbp to 476 kbp.

To align quality-filtered metavirome reads to the final curated vOTU representatives, we used Bowtie2 v2.4.5 in 'end-to-end' mode. A count table was then generated using SAMTools v1.14.[51] The sequence coverage breadth per scaffold was calculated per sample using the SAMTools v1.14 'mpileup' command. To remove spurious Bowtie2 alignments, read counts with a breadth of sequence coverage less than 1 × 75% of a scaffold length were set to zero[44]. Consequently, 102,210 virus sequences were utilised for the construction of the final count table, recruiting on average 89.3% of quality-trimmed reads per metavirome sample. A reads per kilobase per million reads mapped (RPKM) value transformation was applied to the final count table, which was then used for downstream analyses.

Temperate bacteriophages were identified using either the presence of integrase genes or the co-presence of recombinase genes with CI-repressor-like protein genes from the pVOGs annotation within a vOTU representative. The complete list of pVOGs used for temperate phage assignment is available at https://github.com/GRONINGEN-MICROBIOME-

CENTRE/Lifelines_NEXT/blob/main/NEXT_pilot_microbial_transmission_mother_infant/Virome_discovery/list_VOGs_repressor_halmarks.txt.

Virome annotation in total metagenomes

Quality-trimmed reads from 322 total metagenomes were aligned to the curated virus database on a per sample basis using Bowtie2 v2.4.5 in 'end-to-end' mode[30]. A count table was generated and transformed as for metaviromes.

Prediction of viral hosts

Virus host assignment was performed using the iPHoP (v1.2.0) framework with the default settings (FDR<10%) and the database "Sept_2021_pub"[28]. In total, the microbial host was predicted for 68,299 of 102,210 viral scaffolds (67.3%) at the genus level and for 85,135 (82.9%) of all vOTUs at the species level. In cases where multiple hosts were predicted for a virus sequence at the genus or species level, we selected the host taxonomy with the highest Confidence.score from iPHoP. To ensure consistency with the bacterial taxonomic annotation of MetaPhlAn4, the predicted host taxonomy from iPHoP was manually curated. For associations with phenotypes, the RPKM counts of vOTUs were aggregated based on the genus and species levels of the predicted host taxonomy provided by iPHoP.

Ecological measurements and statistical analyses

To assess bacterial and viral alpha diversity, no filters were applied to the relative abundance (bacteria) and RPKM counts (viruses). The alpha diversity for both the bacteriome and virome was calculated using the Shannon diversity index using the *diversity()* function in R package '*vegan*' v.2.6-4. [52]

Beta diversity analysis of the virome and microbiome communities was performed at the vOTU and bacterial species levels using Bray-Curtis dissimilarity. The Bray-Curtis dissimilarity between samples was calculated using the function *vegdist()* from the R package '*vegan*'. We used NMDS to visualise the similarity of bacteriome and virome samples. For that, the function *metaMDS()* from the R package '*vegan*' was employed with 2 dimensions for visualisation purposes and 1 dimension for the analyses related to biome composition

changes. Additionally, *envfit()* with 999 permutations was used to determine the correlation between NMDS and timepoint along with the vector coordinates for Fig, 1b-e.

To test the difference in overall composition of virome and total microbiome (between mothers and infants and between different timepoints), we used a linear mixed model using lmerTest (3.1-3)[53]. The outcome was NMDS1 (dimension). The predictor variable was timepoint (expressed as exact ages in years or days after birth), and we corrected for the number of quality-trimmed reads and DNA concentration as fixed effects. Individual ID was incorporated as a random effect.

To test the difference in the Shannon diversity index between mothers and infants for bacterial and viral abundances, we used a linear mixed model. Here the variable tested was sample type (mother or infant), and we corrected for the number of quality-trimmed reads and DNA concentration as fixed effects and considered individual ID as a random effect. A similar linear mixed model was employed to analyse the effect of timepoint on Shannon diversity in mothers and infants separately, with timepoint (expressed as exact ages in years or days after birth) being the predictor variable, and we corrected for the number of quality-trimmed reads and DNA concentration as fixed effects and individual ID as a random effect. Similar linear mixed models were used for the vOTUs richness (number of detected vOTUs per sample) comparison between mothers and infants.

To compare virome Shannon indices between mothers and infants at 1 year of age, we performed a Wilcoxon rank sum test. To analyse changes in the abundance of vOTUs aggregated at the level of host genus and microbial genus over the first year of an infant's life, a centred log-ratio (clr) transformation was applied using the function *decostand()* from the R package '*vegan*'. The pseudo count specific to the biome, expressed as half of the minimal abundance in community data, was utilised. Only microbial genera and host genera vOTU aggregates present in more than n (10%) of infant samples were considered. Subsequently, a linear mixed model was used with timepoint (expressed as exact ages in days after birth) as the predictor variable and correction for the number of quality-trimmed reads, DNA concentration and mode of delivery as fixed effects and individual ID as a random effect.

We employed a bootstrap resampling approach with replacement to calculate the 95% CIs for the metrics of interest. The goal was to estimate the range within which the true population values for the metrics were likely to fall. We calculated the mean value from each bootstrap sample and repeated this process multiple times (n=1000). A 95% CI was

699 determined by computing the quantiles corresponding to the lower and upper bounds of the
700 distribution (0.025 and 0.975 quantiles, respectively).

701 **Association of vOTUs aggregated by predicted host and bacterial species with phenotypes**
702 The association analysis with phenotypes was conducted on infant samples using a linear
703 mixed model, focusing exclusively on bacterial species and vOTU aggregates by bacterial hosts
704 present in at least 10% of the infant samples. In each model, we tested the predictor
705 (maternal age, infant sex, feeding mode, birthweight, place of birth and gestational age) as a
706 fixed effect. We further corrected for timepoints (expressed as exact ages in days after birth),
707 the number of quality-trimmed reads, DNA concentration and mode of delivery as fixed
708 effects. Individual ID was included as a random effect.
709 For all analyses, an FDR correction was applied to correct for multiple testing, with
710 changes considered statistically significant at FDR<0.05 using the Benjamini-Hochberg
711 method. All statistical tests are two-sided unless explicitly stated.
712
713 **vOTU and bacterial strain-specific analysis**
714 To study viral strain-sharing within mother-infant pairs, we focused on a subset of vOTU
715 representatives that were covered by reads at over 95% of the genome length and shared
716 between maternal metaviromes and/or metagenomes and infant metaviromes. This subset
717 consisted of 4,965 vOTUs. We then selected vOTU representatives for further analysis based
718 on the following criteria: 1) a high-quality or complete genome predicted by CheckV or
719 circularised genomes[43], 2) sequence length ≥3 kbp and 3) present in samples from at least five
720 different families. There were 51 vOTU representatives fulfilling these criteria. For each
721 selected vOTU, we reconstructed consensus sequences for all samples where the vOTU of
722 interest was covered at over 95% of the genome length. We employed the function
723 `consensus` with flags `-m simple -r` from SAMTools on the read alignments from Bowtie2
724 output that were used for the RPKM table construction.
725 We next performed global alignments of consensus sequences per vOTU using *kalign*
726 v1.04[54]. To improve alignment quality, we trimmed 100 bp from both ends of the global
727 alignment, which were enriched in gaps. Pairwise genetic distances were then calculated
728 using the *dist.dna()* function from the R package *ape* v.5.7-1[55] with default parameters

resulting in Kimura 2-parameter (pairwise nucleotide substitution rate between strains) pairwise distances. To compare the pairwise Kimura distances for virus strains between samples of related individuals, we used a one-sided Wilcoxon rank sum test with an alternative hypothesis that the distances between strains identified in samples of related individuals are smaller than the distances between unrelated samples. Significance of the distance comparison was derived in a permutation test with 1,000 iterations, designed to account for the highly unequal number of distances between strains of related and unrelated individuals. FDR correction for multiple testing was applied as described above.

To investigate strain-sharing between mothers and infants, we selected those viruses with lower distances between strains of related individuals compared to unrelated individuals. To define strain-sharing events within mother-infant pairs, we used an approach similar to that used in Valles-Colomer et al.[56]. In short, we compared the median-normalised distances within individual strains over the entire study period (maximum 9 months for maternal samples and maximum 12 months for infants) to the normalised distances between strains of unrelated individuals, per vOTU. The strain identity cut-off was calculated using the *cutpointr()* function from the R package *cutpointr* v.1.1.2[57]. For the identification of the optimal cutpoint, we used the *oc_youden_kernel* parameter along with the *youden* metric. Additionally, empirical FDR was defined as the 5th percentile of the unrelated individual comparisons when Youden's index was above 5%. We then compared the percentage of shared versus different dominant strains in related and unrelated mother-infant pairs (all timepoint pairs) using the defined strain identity cut-off. If normalised distances between strains were greater than the cut-off, the strains were deemed different. If they were smaller, this was considered a strain-sharing event. This allowed us to calculate a percentage of dominant strain-sharing between related and unrelated mother-infant pairs, which we then tested for significance using the one-sided Fisher's exact test with subsequent FDR correction for multiple testing using Benjamini-Hochberg.

We found 26 viruses to be shared between mothers and infants. Bacterial hosts for 25 of the 26 transmitted viruses were predicted at the species level using iPHoP. All predicted hosts of viruses were used for co-transmission analysis.

Bacterial-species-specific strain analysis

We reconstructed bacterial strain SNP haplotypes for the predicted hosts of 25 transmitted viruses using StrainPhlAn4[35], resulting in 37 bacterial strain SNP haplotypes. This method is based on reconstructing consensus sequence variants within species-specific marker genes and using them to estimate strain-level phylogenies[50]. We then performed multiple sequence alignment and used the Kimura 2-parameter method from the '*EMBOSS'* package[51] to create phylogenetic distance matrices that contain the pairwise nucleotide substitution rate between strains. We next employed the same methods for the identification of strain-sharing events as described above.

Virus–host co-transmission from mothers to infants

To determine if the shared viruses were co-transmitted with their predicted bacterial hosts, we employed the Mantel partial test on modified genetic distance matrices for bacterial and virus strains. This test assessed the correlation between the strain-sharing events for bacteria and phages while controlling for longitudinally collected samples.

First, constructed Kimura genetic distance matrices were normalised by the median genetic distance per bacterial strain and vOTU, respectively. Next, the normalised values of genetic distances were replaced with 0 if the distance did not exceed the vOTU- or bacterial-strain-specific cut-off for individual strain variation (Youden index or 5% FDR), otherwise, it was replaced with 1. This modification allowed us to focus on strain-sharing events rather than on the correlation between genetic distances themselves.

Next, for each vOTU and bacterial strain, we selected subsets of concurrent samples where both the vOTU and bacterial strain were reconstructed. For bacterial strains, only total metagenomes were used. For the viral strains, strains reconstructed from both total metagenomes and metaviromes were included. If both types were available for the same individual and timepoint, strains from metavirome samples were prioritised.

To account for repeated measurements, we created control matrices for the selected subsets of concurrent samples. In this matrix, we assigned a value of 0 when the strain was reconstructed in samples from the same individual and a value of 1 when the strain was reconstructed in samples from different individuals. In this analysis, mothers and their infants were treated as different individuals.

The Mantel partial test, using *mantel.partial()* function from the R package *vegan* was performed on the modified genetic distance matrices for bacteria and viruses. We used the

Pearson correlation method and 999 permutations to assess the significance of the correlation while controlling for longitudinal samples using the control matrix. In cases when virus and bacterium distance matrices were collinear, we used *mantel()* function from the R package *vegan*. In cases when one of the distance matrices was collinear with the control distance matrix (e.g., in the subset of concurrent samples, the strain-sharing pattern repeated the structure of the longitudinal samples belonging to the same/different individuals), it was not possible to calculate the Mantel partial test while accounting for the repeated measurements. These pairs of virus-bacterium were not used for the testing of co-transmission frequency. The p-values obtained from the Mantel partial test were adjusted for multiple testing using FDR correction using the Benjamini-Hochberg method.

Additionally, we conducted a one-sided Fisher's exact test to compare the frequency of co-transmission events between phage-bacteria pairs connected by a virus–host relationship and random phage-bacteria pairs.

To further test the non-randomness of co-sharing patterns of viruses and their bacterial hosts, we calculated the difference between the observed frequency of co-transmission of virus and bacterial strains (i.e. the proportion of concurrent samples where both virus and bacterial strain had a modified genetic distance equal to 0) and the expected frequency based on independent strain sharing events (the product of the fractions of samples with a modified genetic distance of 0 for viruses and bacteria independently). The difference between the observed and expected frequencies is computed and termed “linkage disequilibrium”. The significance of linkage was calculated using a chi-squared test.

Bacterial genome binning and read alignment to patched genomes

Scaffolds were assembled from total metagenomes in the same way as for metaviromes. All scaffolds were then used for genome binning using metaWRAP[58]. Taxonomy was assigned to these bins using GTDB-Tk v2[59]. Bacterial genomic bins were then patched and scaffolded using RagTag [60] using ‘scaffold’ option, and the genomes of isolates from MGnify, if available, or metagenome-assembled genomes[61] (MGYG000132487 for *Bifidobacterium bifidum* and MGYG000003383 for *Bifidobacterium scardovii*). vOTU representatives of interest were then mapped to the patched bacterial genomes using minimap2 [62] using `-X -N 50 -p 0.1 -c` flags to locate the prophage region. Next, quality-trimmed reads from total metagenomes and metaviromes were aligned to the patched genomes using Bowtie2 as described above. The

sequence coverage breadth per scaffold was calculated per sample using the SAMTools 'mpileup' command. A 1001 nt window with 1001 nt step was used to calculate mean coverage depth; the most 3'-terminal window was extended to include up to 1001 3'-terminal nucleotides.

Gene annotation of vOTU representatives of interest was performed using VIBRANT in `-virome` mode[63].

Data visualisation

Results were visualised in graphical form using a set of custom R scripts, including calls to functions from the following packages: *ggplot2* v.3.4.2[64], *ggrepel* v.0.9.3[65], *ggforce* v.0.4.1[66], *patchwork* v.1.1.2[67], *tidyverse* v.2.0.0[68], *EnhancedVolcano* v.1.16.0[69], *ggforestplot* v.0.1.0[70], *corrplot* v.0.92[71] and *ggtree* v.3.6.2[72]. All boxplots were prepared using *ggplot2* and represent standard Tukey type with interquartile range (IQR, box), median (bar) and Q1 – 1.5 × IQR/Q3 + 1.5 × IQR (whiskers). Phylogenetic trees were built based on the Kimura 2-parameter distance matrices. First, hierarchical clustering was applied to the distance matrices using the function *hclust()* from the R package *'stats'* v.4.2.1[73]. Clustering trees were then converted into phylogenetic trees with function *as.phylo()* from R package *`ape`* v.5.7.-1. Visualisation of the phylogenetic trees was done using the function *ggtree()* from the R package *'ggtree'*[72].

Data and code availability

Sample information, basic phenotypes, family structure and quality-trimmed sequencing reads can be found in the EGA archive (study ID: EGAS00001005969). The datasets generated during this study including virus scaffolds and their metadata are available at the Figshare repository under https://doi.org/10.6084/m9.figshare.23926593.v1. All codes used in this study can be found at: https://github.com/GRONINGEN-MICROBIOME-CENTRE/Lifelines_NEXT/tree/main/NEXT_pilot_microbial_transmission_mother_infant.

References

1. Fan, Y. & Pedersen, O. Gut microbiota in human metabolic health and disease. *Nat. Rev. Microbiol.* **19**, 55–71 (2021).
2. Robertson, R. C., Manges, A. R., Finlay, B. B. & Prendergast, A. J. The Human Microbiome and Child Growth – First 1000 Days and Beyond. *Trends Microbiol.* **27**, 131–147 (2019).
3. Tamburini, S., Shen, N., Wu, H. C. & Clemente, J. C. The microbiome in early life: implications for health outcomes. *Nat. Med.* **22**, 713–722 (2016).
4. Stewart, C. J. *et al.* Temporal development of the gut microbiome in early childhood from the TEDDY study. *Nature* **562**, 583–588 (2018).
5. Yatsunenko, T. *et al.* Human gut microbiome viewed across age and geography. *Nature* **486**, 222–227 (2012).
6. Bäckhed, F. *et al.* Dynamics and Stabilization of the Human Gut Microbiome during the First Year of Life. *Cell Host Microbe* **17**, 690–703 (2015).
7. Duranti, S. *et al.* Maternal inheritance of bifidobacterial communities and bifidophages in infants through vertical transmission. *Microbiome* **5**, (2017).
8. Van Daele, E., Knol, J. & Belzer, C. Microbial transmission from mother to child: improving infant intestinal microbiota development by identifying the obstacles. *Crit. Rev. Microbiol.* **45**, 613–648 (2019).
9. Wang, S. *et al.* Maternal Vertical Transmission Affecting Early-life Microbiota Development. *Trends Microbiol.* **28**, 28–45 (2020).
10. Lou, Y. C. *et al.* Infant gut strain persistence is associated with maternal origin, phylogeny, and traits including surface adhesion and iron acquisition. *Cell Rep. Med.* **2**, (2021).

11. Ferretti, P. *et al.* Mother-to-Infant Microbial Transmission from Different Body Sites Shapes the Developing Infant Gut Microbiome. *Cell Host Microbe* **24**, 133-145.e5 (2018).
12. Yassour, M. *et al.* Strain-Level Analysis of Mother-to-Child Bacterial Transmission during the First Few Months of Life. *Cell Host Microbe* **24**, 146-154.e4 (2018).
13. Podlesny, D. & Fricke, W. F. Strain inheritance and neonatal gut microbiota development: A meta-analysis. *Int. J. Med. Microbiol.* **311**, 151483 (2021).
14. Korpela, K. *et al.* Selective maternal seeding and environment shape the human gut microbiome. *Genome Res.* **28**, 561–568 (2018).
15. Asnicar, F. *et al.* Studying Vertical Microbiome Transmission from Mothers to Infants by Strain-Level Metagenomic Profiling. *mSystems* **2**, e00164-16 (2017).
16. Megli, C. J. & Coyne, C. B. Infections at the maternal-fetal interface: an overview of pathogenesis and defence. *Nat. Rev. Microbiol.* 1–16 (2021) doi:10.1038/s41579-021-00610-y.
17. Garmaeva, S. *et al.* Stability of the human gut virome and effect of gluten-free diet. *Cell Rep.* **35**, 109132 (2021).
18. Brum, J. R., Hurwitz, B. L., Schofield, O., Ducklow, H. W. & Sullivan, M. B. Seasonal time bombs: dominant temperate viruses affect Southern Ocean microbial dynamics. *ISME J.* **10**, 437–449 (2016).
19. Emerson, J. B. *et al.* Host-linked soil viral ecology along a permafrost thaw gradient. *Nat. Microbiol.* **3**, 870–880 (2018).
20. Liang, G. *et al.* The stepwise assembly of the neonatal virome is modulated by breastfeeding. *Nature* **581**, 470–474 (2020).
21. Maqsood, R. *et al.* Discordant transmission of bacteria and viruses from mothers to babies at birth. *Microbiome* **7**, 156 (2019).

22. 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).
23. Kennedy, K. M. *et al.* Fetal meconium does not have a detectable microbiota before birth. *Nat. Microbiol.* **6**, 865–873 (2021).
24. Kennedy, K. M. *et al.* Questioning the fetal microbiome illustrates pitfalls of low-biomass microbial studies. *Nature* **613**, 639–649 (2023).
25. Sinha, T., Brushett, S., Prins, J. & Zhernakova, A. The maternal gut microbiome during pregnancy and its role in maternal and infant health. *Curr. Opin. Microbiol.* **74**, 102309 (2023).
26. Koren, O. *et al.* Host Remodeling of the Gut Microbiome and Metabolic Changes during Pregnancy. *Cell* **150**, 470–480 (2012).
27. Shkoporov, A. N. *et al.* The Human Gut Virome Is Highly Diverse, Stable, and Individual Specific. *Cell Host Microbe* **26**, 527-541.e5 (2019).
28. 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).
29. McCann, A. *et al.* Viromes of one year old infants reveal the impact of birth mode on microbiome diversity. *PeerJ* **6**, e4694 (2018).
30. Warmink-Perdijk, W. D. B. *et al.* Lifelines NEXT: a prospective birth cohort adding the next generation to the three-generation Lifelines cohort study. *Eur. J. Epidemiol.* **35**, 157–168 (2020).
31. Bushnell, B. BBMap: A Fast, Accurate, Splice-Aware Aligner. in (2014).
32. Beghini, F. *et al.* Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. *eLife* **10**, e65088 (2021).

923 33. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods*
924 **9**, 357–359 (2012).
925 34. Andrews, S. FastQC: A Quality Control Tool for High Throughput Sequence Data.
926 (2010).
927 35. Blanco-Míguez, A. *et al.* Extending and improving metagenomic taxonomic profiling
928 with uncharacterized species using MetaPhlAn 4. *Nat. Biotechnol.* (2023)
929 doi:10.1038/s41587-023-01688-w.
930 36. Hill, J. E., Penny, S. L., Crowell, K. G., Goh, S. H. & Hemmingsen, S. M. cpnDB: a
931 chaperonin sequence database. *Genome Res* **14**, 1669–1675 (2004).
932 37. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. metaSPAdes: a new versatile
933 metagenomic assembler. *Genome Res.* **27**, 824–834 (2017).
934 38. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site
935 identification. *BMC Bioinformatics* **11**, 119 (2010).
936 39. McGinnis, S. & Madden, T. L. BLAST: at the core of a powerful and diverse set of
937 sequence analysis tools. *Nucleic Acids Res.* **32**, W20–W25 (2004).
938 40. Eddy, S. R. Accelerated Profile HMM Searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).
939 41. Grazziotin, A. L., Koonin, E. V. & Kristensen, D. M. Prokaryotic Virus Orthologous
940 Groups (pVOGs): a resource for comparative genomics and protein family annotation.
941 *Nucleic Acids Res.* **45**, D491–D498 (2017).
942 42. Roux, S., Enault, F., Hurwitz, B. L. & Sullivan, M. B. VirSorter: mining viral signal from
943 microbial genomic data. *PeerJ* **3**, e985 (2015).
944 43. Crits-Christoph, A. *et al.* Functional interactions of archaea, bacteria and viruses in a
945 hypersaline endolithic community: Halite metagenome. *Environ. Microbiol.* **18**, 2064–
946 2077 (2016).

947 44. Yutin, N. *et al.* Analysis of metagenome-assembled viral genomes from the human gut
948 reveals diverse putative CrAss-like phages with unique genomic features. *Nat. Commun.*
949 **12**, 1044 (2021).
950 45. Guerin, E. *et al.* Biology and Taxonomy of crAss-like Bacteriophages, the Most
951 Abundant Virus in the Human Gut. *Cell Host Microbe* **24**, 653-664.e6 (2018).
952 46. Gulyaeva, A. *et al.* Discovery, diversity, and functional associations of crAss-like phages
953 in human gut metagenomes from four Dutch cohorts. *Cell Rep.* **38**, (2022).
954 47. Schoch, C. L. *et al.* NCBI Taxonomy: a comprehensive update on curation, resources and
955 tools. *Database* **2020**, baaa062 (2020).
956 48. Roux, S. *et al.* Minimum Information about an Uncultivated Virus Genome (MIUViG).
957 *Nat. Biotechnol.* **37**, 29–37 (2019).
958 49. Quast, C. *et al.* The SILVA ribosomal RNA gene database project: improved data
959 processing and web-based tools. *Nucleic Acids Res.* **41**, D590–D596 (2013).
960 50. Bin Jang, H. *et al.* Taxonomic assignment of uncultivated prokaryotic virus genomes is
961 enabled by gene-sharing networks. *Nat. Biotechnol.* **37**, 632–639 (2019).
962 51. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**,
963 2078–2079 (2009).
964 52. Jari Oksanen, M. F., Roeland Kindt, Pierre Legendre, Dan McGlinn, Peter R. Minchin, R.
965 B. O'Hara, Gavin L. Simpson, Peter Solymos, M. Henry H. Stevens, Eduard Szoecs,
966 Helene Wagner. vegan: Community Ecology Package. (2020).
967 53. Kuznetsova, A., Brockhoff, P. B. & Christensen, R. H. B. lmerTest Package: Tests in
968 Linear Mixed Effects Models. *J. Stat. Softw.* **82**, 1–26 (2017).
969 54. Lassmann, T. Kalign 3: multiple sequence alignment of large datasets. *Bioinformatics* **36**,
970 1928–1929 (2020).

971 55. Paradis, E. & Schliep, K. ape 5.0: an environment for modern phylogenetics and
972 evolutionary analyses in R. *Bioinformatics* **35**, 526–528 (2019).
973 56. Valles-Colomer, M. *et al.* The person-to-person transmission landscape of the gut and
974 oral microbiomes. *Nature* **614**, 125–135 (2023).
975 57. Thiele, C. & Hirschfeld, G. cutpointr: Improved Estimation and Validation of Optimal
976 Cutpoints in R. *J. Stat. Softw.* **98**, 1–27 (2021).
977 58. Uritskiy, G. V., DiRuggiero, J. & Taylor, J. MetaWRAP—a flexible pipeline for genome-
978 resolved metagenomic data analysis. *Microbiome* **6**, 158 (2018).
979 59. Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk v2: memory
980 friendly classification with the genome taxonomy database. *Bioinformatics* **38**, 5315–
981 5316 (2022).
982 60. Alonge, M. *et al.* Automated assembly scaffolding using RagTag elevates a new tomato
983 system for high-throughput genome editing. *Genome Biol.* **23**, 258 (2022).
984 61. Richardson, L. *et al.* MGnify: the microbiome sequence data analysis resource in 2023.
985 *Nucleic Acids Res.* **51**, D753–D759 (2023).
986 62. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–
987 3100 (2018).
988 63. Kieft, K., Zhou, Z. & Anantharaman, K. VIBRANT: automated recovery, annotation and
989 curation of microbial viruses, and evaluation of viral community function from genomic
990 sequences. *Microbiome* **8**, 90 (2020).
991 64. Wickham, H. ggplot2: Elegant Graphics for Data Analysis. (2016).
992 65. Slowikowski, K. ggrepel: Automatically Position Non-Overlapping Text Labels with
993 'ggplot2'. (2023).
994 66. Pedersen, T. ggforce: Accelerating 'ggplot2'. (2022).
995 67. Pedersen, T. patchwork: The Composer of Plots. (2022).

68. Wickham, H. Welcome to the Tidyverse. *J. Open Source Softw.* (2019).

69. Blighe, K. EnhancedVolcano: Publication-ready volcano plots with enhanced colouring and labeling.". (2018).

70. NightingaleHealth. ggforestplot. (202AD).

71. Wei, T. R package 'corrplot': Visualization of a Correlation Matrix. (2021).

72. Xu, S. *et al.* Ggtree: A serialized data object for visualization of a phylogenetic tree and annotation data. *iMeta* **1**, e56 (2022).

73. R Core Team. R: A language and environment for statistical computing. (2013).

Acknowledgements

We thank the participants of the Lifelines NEXT cohort for their participation. We thank doctors' assistants and the service bureau team of Lifelines NEXT and acknowledge significant contributions of the persons involved in the collection, processing and storage of samples: Brenda Hijmans, Annet Jansen, Gea Lamberts, Rianne de Roos, Daphne Teuben and Ettje Tigchelaar. We thank Kate Mc Intyre for editing the manuscript. We also thank the Genomics Coordination Center and the Center for Information Technology of the University of Groningen for their support and for providing access to the Gearshift and Peregrine high-performance computing clusters. We thank Daoming Wang for his assistance in the figure designs and Dianne Jansen for her assistance in DNA isolations.

SG and TS hold scholarships from the Graduate School of Medical Sciences, University of Groningen and the Junior Scientific Masterclass, University of Groningen, respectively. SG was awarded a de Cock-Hadders Stitching grant (grant number: 2021-08). TS was awarded a de Cock-Hadders Stichting grant (Winston Bakker Fonds 2021-WB-08). JES was supported by grants from the de Cock-Hadders Stichting (grant number: 2021-57) and the International Society for Research in Human Milk and Lactation (ISRHML). SB was supported by EUCAN-connect, a federated FAIR platform enabling large-scale analysis of high-value cohort data connecting Europe and Canada in personalized health. Furthermore, this project was funded by the Netherlands Heart Foundation (IN-CONTROL CVON grant 2018-27 to AZ and JF), the Netherlands Organization for Scientific Research (NWO) (NWO Gravitation Exposome-NL (024.004.017) to JF, AK and AZ; NWO-VIDI 864.13.013 and NWO-VICI VI.C.202.022 to JF; NWO-VIDI 016.178.056 to AZ and NWO Spinoza Prize SPI 92-266 to CW), the European Research Council (ERC) (ERC Advanced Grant 2012-322698 to CW, ERC Consolidator Grant 101001678 to JF and ERC Starting Grant 715772 to AZ) and the RuG Investment Agenda Grant Personalized Health to CW. JF and CW are also supported by the Netherlands Organ-on-Chip Initiative, an NWO Gravitation project (024.003.001) funded by the Ministry of Education, Culture and Science of the government of the Netherlands. ANS was supported by a Wellcome Trust Research Career Development Fellowship (220646/Z/20/Z) and the ERC under the European Union's Horizon 2020 research and innovation programme (ERC Consolidator, grant agreement no. 101001684). This research was funded in whole, or in part, by the Wellcome Trust [220646/Z/20/Z]. For the purpose of open access, the authors have applied a CC BY

1036 public copyright licence to any Author Accepted Manuscript version arising from this
1037 submission. The study was also financially supported with a public-private partnership
1038 allowance of Health Holland Topsector Life Sciences & Health to stimulate public-private
1039 partnerships. The funders had no role in the study design, data collection and analysis,
1040 decision to publish, or preparation of the manuscript.

Figure legends

Figure 1: Overall composition of the maternal and infant gut virome and bacteriome during pregnancy and over the first year of life.

a, Numbers of successfully sequenced maternal and infant faecal samples collected at gestational weeks 12 and 28, at birth, and months 1, 2, 3, 6, 9 and 12 after birth (numbers in grey circles). Timepoints are displayed in grey circles with a white background. Metavirome is the VLP-enriched viral fraction of the microbial population. Total metagenome is the primarily bacterial fraction. Numbers in green and yellow circles indicate the number of samples per timepoint. **b–e**, Non-metric multidimensional scaling (NMDS) analysis based on Bray-Curtis dissimilarity calculated at the level of b, infant vOTUs, **c**, infant bacterial species, **d**, mother vOTUs and **e**, mother bacterial species. In all graphs, each point represents one sample, and different colours indicate samples from various timepoints. In **b** and **c**, all maternal samples from all timepoints are denoted in yellow. The distance between points reflects the level of dissimilarity. The centroids for the clusters of samples from the same timepoint are represented by diamonds of the respective colours. Boxed labels with arrows indicate significant vectors determined by Envfit analysis. The boxplots on top of NMDS plots depict the distribution of NMDS1 per timepoint.

Figure 2: Longitudinal dynamics of the infant and maternal gut virome and their bacterial hosts.

a, **b**, Alpha diversity, as measured by the Shannon Diversity Index in **a,** virome at the vOTUs level and **b,** bacteriome communities at the species level. Colours of boxplots represent the sample type: infant (coral) and maternal (blue). **c**, Number and **d**, relative abundance of vOTUs retained from month 1 in infants at months 2, 3, 6, and 12 after birth. **e**, Number and **f**, the relative abundance of vOTUs retained from the 7th month of pregnancy in mothers at birth and months 1 to 3 after delivery. In **c**-**f**, the 95% confidence intervals (CI) for the means are depicted in grey and brown whiskers for retained and not present categories, respectively. **g**, Percentage of vOTUs classified into fractions and **h**, the cumulative relative abundance of vOTUs classified into biome fractions based on vOTUs prevalence in the metaviromes of 14 infants with at least three longitudinal samples. Personal persistent viromes (PPVs) are those present in >= 75% of individual samples. Transiently detected viruses (TDVs) are those present

in less than 75% of the samples of an individual). **i**, Stacked bar plot showing the top-10 most abundant vOTU aggregated by their bacterial hosts at the genus level in infants and mothers. The relative abundance of vOTU aggregates was rescaled to fit to 1 after removing Unassigned and less abundant aggregates. Translucent stacked bars indicate the virome annotation in whole metaviromes for the timepoints that were unavailable for the VLP-enriched metavirome profiling (timepoints P3 and M9). Opaque stacked bars indicate the composition of the metaviromes. **j**, Stacked bar plot of top-10 abundant bacterial genera in infants and mothers. The relative abundance of these bacterial genera was rescaled as described above after removing Unknown fraction.

Figure 3: Infant gut virome in relation to diet and maternal factors over time.

a, Relative abundances of active temperate phages in the infant gut at months 1, 2, 3, 6 and 12 after birth and maternal gut at gestational week 28, infant birth and months 1, 2, and 3 after delivery. **b**, Alpha diversity of metaviromes, as measured by the Shannon Diversity Index for vOTUs, and **c**, Richness of active temperate phages in response to feeding mode. Boxplot colours indicate the infant feeding type during sample collection: breastfed in red and exclusively formula fed in blue. **d**, Volcano plot showing the results of the linear mixed model testing for differential prevalence (present in more than 10% infant samples) vOTU aggregates at the host species level based on the infant feeding mode. P-values were corrected for multiple testing using the Benjamini-Hochberg method. The horizontal line indicates an FDR cut-off of 0.05, and vertical lines indicate a log-fold change cut-off of -1 or 1. Each dot represents a vOTU aggregate, and its color represents the FDR significance. **e**, Percentage of infant vOTUs shared with pre-birth (orange) and post-birth (green) active maternal virome. **f**, Percentage of sharedness of infant vOTUs taking into account the prophages (based on the presence of temperate phages in whole metaviromes) with pre-birth (orange) and post-birth (green) maternal samples.

Figure 4: Sharing of gut bacterial and viral strains between mothers and infants.

Log scaled Kimura 2-parameter distances of various **a**, viruses and **c**, bacterial strains with significantly lower genetic distances in related mother-infant pairs (green) compared to

unrelated mother-infant pairs (grey) mother-infant pairs. Asterisks denote FDR significance. The panels on the right of a and c indicate the total number of pairwise distances for the related and unrelated mother-infant pairs for the virus/bacterium. Percentage of mother-infant sample pairs sharing the same strain of (**b**) virus and (**d**) bacterial strains as defined by the within-individual strain variation cut-off. Y-axis of **a** and **b** contain the labels of viruses composed of scaffold length preceded by 'L' and predicted lifestyle (0 for virulent and 1 for temperate preceded by 'LS').

Figure 5: Co-transmission of viral strains and their bacterial host strains from maternal to infant guts.

a, Correlogram showing the Pearson correlation coefficient between virus and their bacterial host transmission events in concurrent samples. Squares marked in red indicate the assigned host of the virus. Circle size and colour intensity indicate the correlation's strength. A square with an X indicates cases in which either no concurrent samples with both bacteria and virus being present were available for the partial mantel test or when one of the distance matrices was collinear with the repeated measurements control matrix. Blank squares and white circles in red squares indicate the FDR-insignificant correlations (> 0.1). **b**, Mosaic plot of the proportions of viruses co-transmitted (in green) and not co-transmitted (yellow) with their predicted bacterial hosts. **c**, Phylogenetic trees of a virus (L85266_LS0) and its predicted host (*Bacteroides uniformis*). Colours indicate the mother-infant pair they belong to. Shape indicates the type of the sample (infant or mother). Symbol border indicates whether the strain was reconstructed in a metagenomic (MGS, light grey) or metaviromic (VLP, black) sample.

Figure 6: Example of virus-host co-transmission: a temperate phage of *Bifidobacterium bifidum*

a, Phylogenetic trees of a virus (L34922_LS1) and its predicted host (*Bifidobacterium bifidum*). Colours indicate the mother-infant pairs (family) affiliation, and shapes indicate the infants and mothers. Symbol borders show whether the strain was reconstructed from metagenomic (MGS, light grey) or metaviromic (VLP, black) samples. **b**, Presence patterns of L34922_LS1

and *B. bifidum* in the samples of the mother-infant pair (FAM0234) where the transmission of L34922_LS1 was detected. X-axis indicates timepoints, Y-axis shows the individuals within the family (mother or infant). Polygon shape indicates the sample biome: total metagenome (square), metavirome (triangle), whole metavirome (rotated triangle). Color of the shape indicates the presence (red) or absence (white) of the virus or bacterial strain in the biome sample, and the missing samples are colored in grey. **c**, Synteny plot of L34922_LS1 genome sequence mapping to the fragment of the *B. bifidum* patched genome fragment. X-axis indicates the genome coordinates (in megabases). Lines connecting the L34922_LS1 and *B. bifidum* genome fragment indicate the prophage insertion. **d**, Genome organization of L34922_LS1. X-axis depicts the genome coordinates in base pairs. A polygon represents every predicted protein, and its orientation indicates the location of the predicted protein at the positive (right-orientation) or negative (left-orientation) strands. Colors indicate the source of protein annotation: Pfam (dark green), VOG (white), and proteins with no functional annotations are shown in orange. Hypothetical proteins and annotated proteins of unknown function are shown in respective colors without text annotation. **e**, Depth of the *B. bifidum* patched genome fragment coverage by sequencing reads from total metagenomes that were positive for *B. bifidum*, and **f**, metaviromes that were positive for L34922_LS1. Each transparent grey line corresponds to a sample and represents mean depth in a 1,001-nt sliding window. X-axis depicts the genome coordinate. Colored lines represent samples from the mother-infant pair (FAM0234), and colors are associated with timepoints. Dashed lines indicate the prophage insertion region.

[§]Lifelines NEXT cohort study

Jackie Dekens[1,2], Aafje Dotinga[3], Sanne Gordijn[4], Soesma Jankipersadsing[1], Ank de Jonge[5,9], Marlou L.A. de Kroon[6], Gerard H. Koppelman[7], Folkert Kuipers[8], Lilian L. Peters[5,9], Jelmer R. Prins[4], Sijmen A. Reijneveld[6], Sicco Scherjon[4], Jan Sikkema[2], Morris A. Swertz[1], Henkjan J. Verkade[8], Cisca Wijmenga[1], Alexandra Zhernakova[1]

[1]Department of Genetics, University of Groningen and University Medical Center Groningen, Groningen, the Netherlands

[2]University Medical Center Groningen, Center for Development and Innovation

[3]Lifelines cohort study, Groningen, the Netherlands

[4]Department of Obstetrics and Gynecology, University of Groningen and University Medical Center Groningen, Groningen, the Netherlands

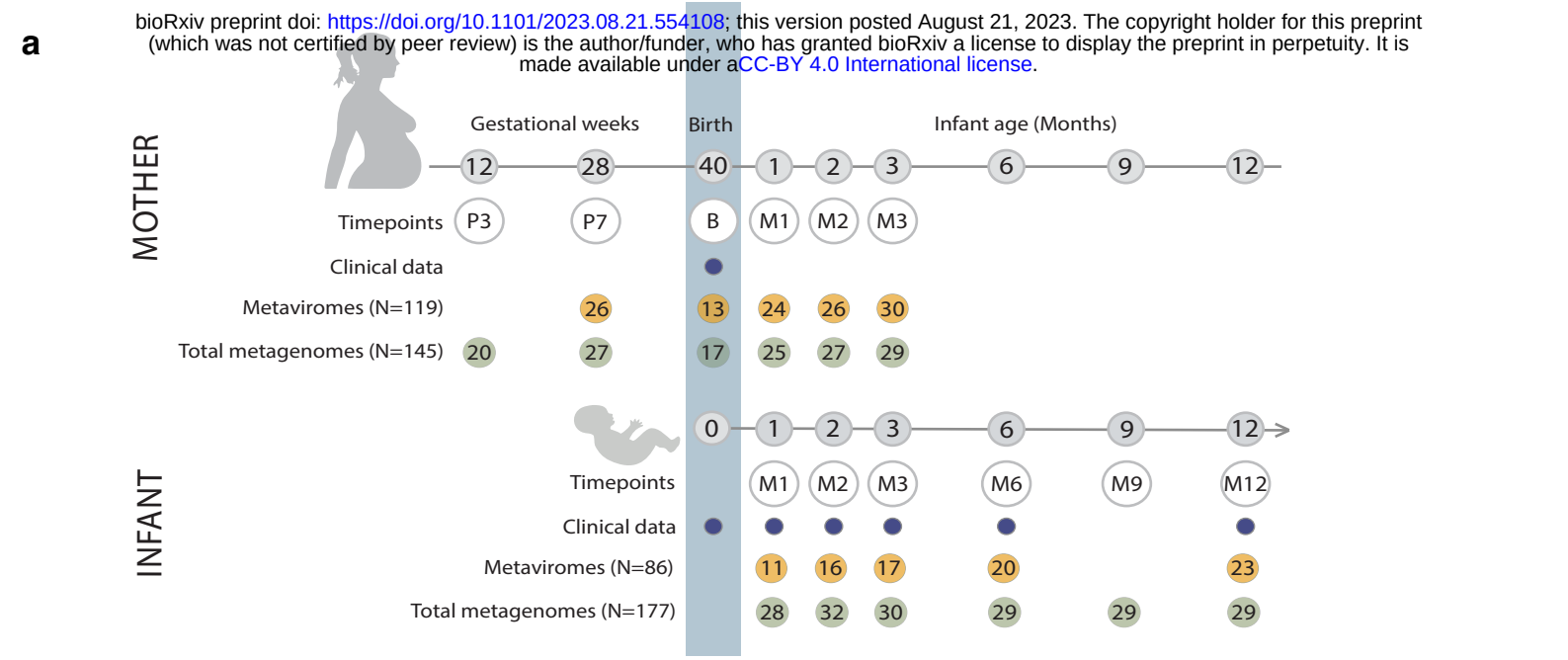
[5]Department of Midwifery Science, Amsterdam University Medical Centre, Vrije Universiteit Amsterdam, AVAG/Amsterdam Reproduction and Development, Amsterdam, the Netherlands

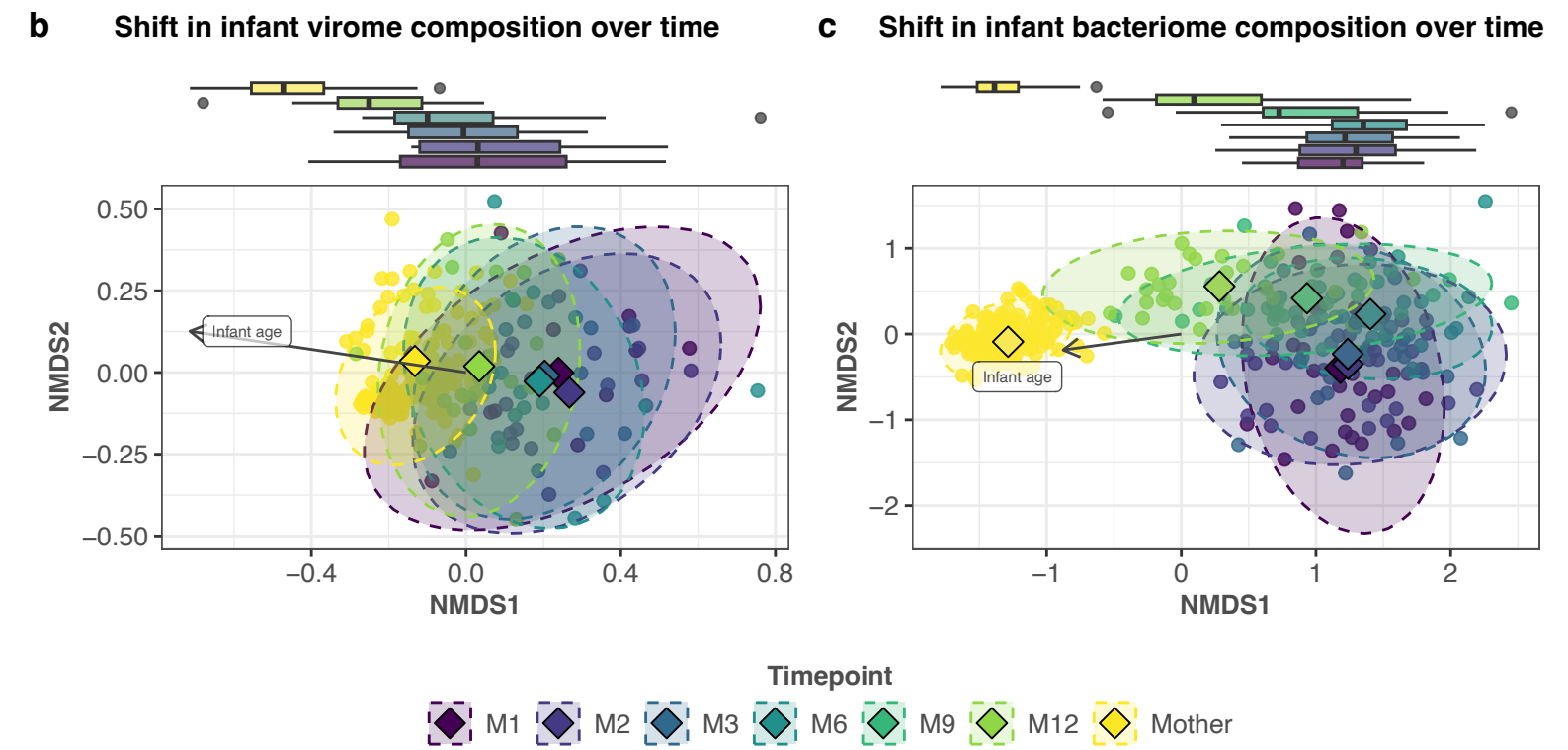
[6]Department of Health Sciences, University of Groningen and University Medical Center Groningen, University of Groningen, Groningen, the Netherlands

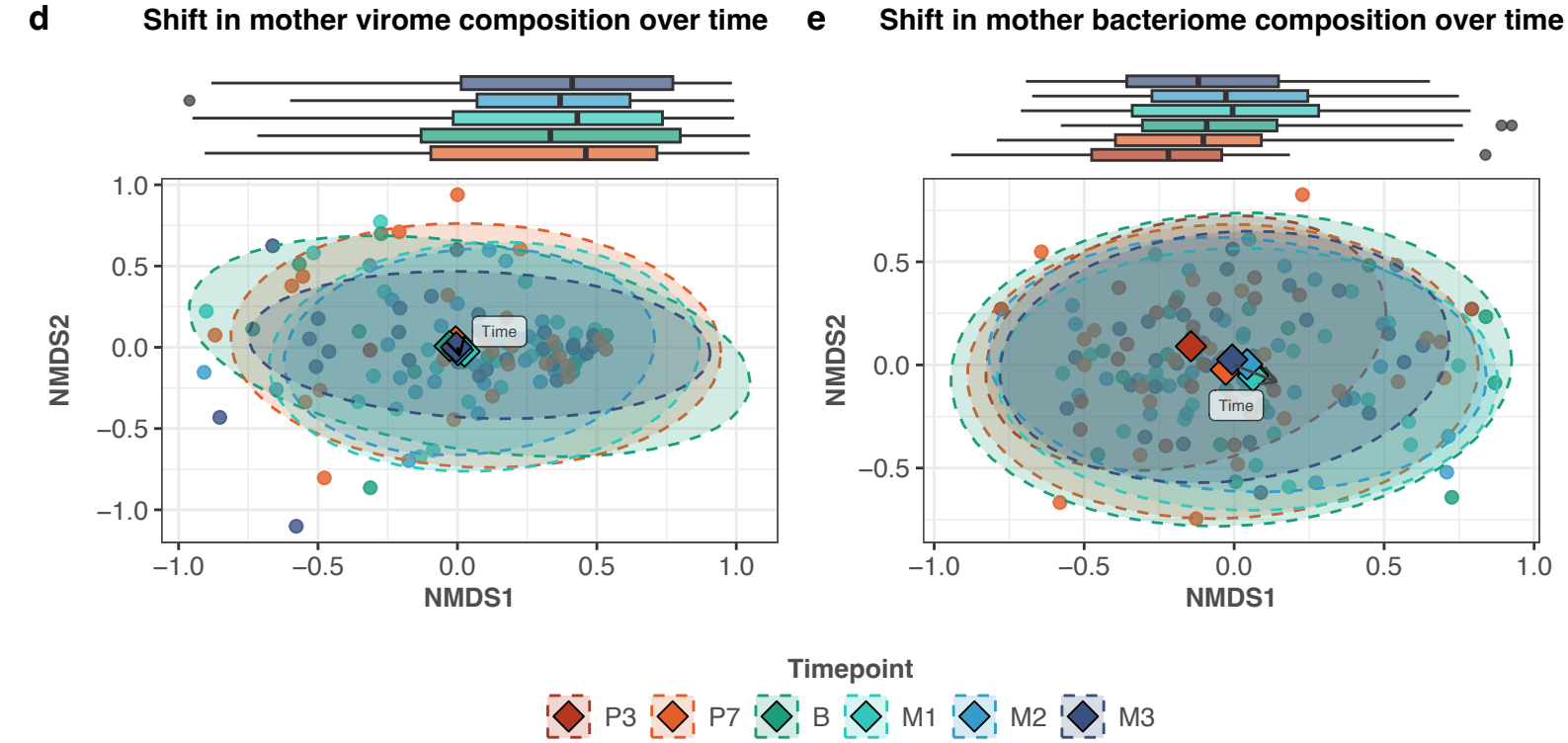
[7]Department of Paediatric Pulmonology and Paediatric Allergology, Beatrix Children's Hospital, University of Groningen, University Medical Center Groningen, Groningen, the Netherlands.

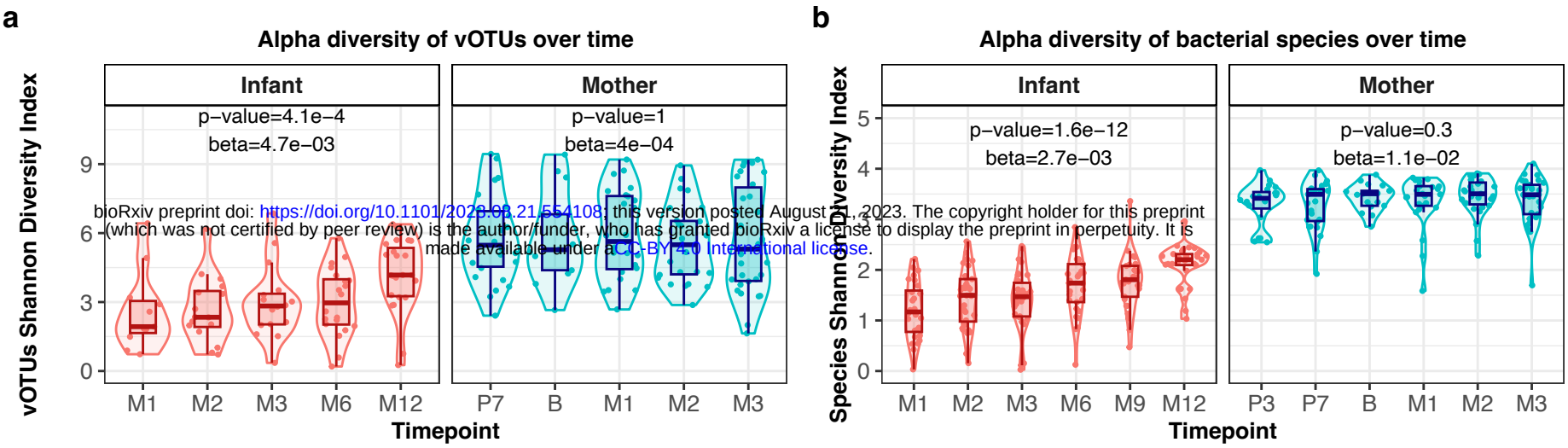
[8]Department of Pediatrics, University of Groningen and University Medical Center Groningen, Groningen, the Netherlands

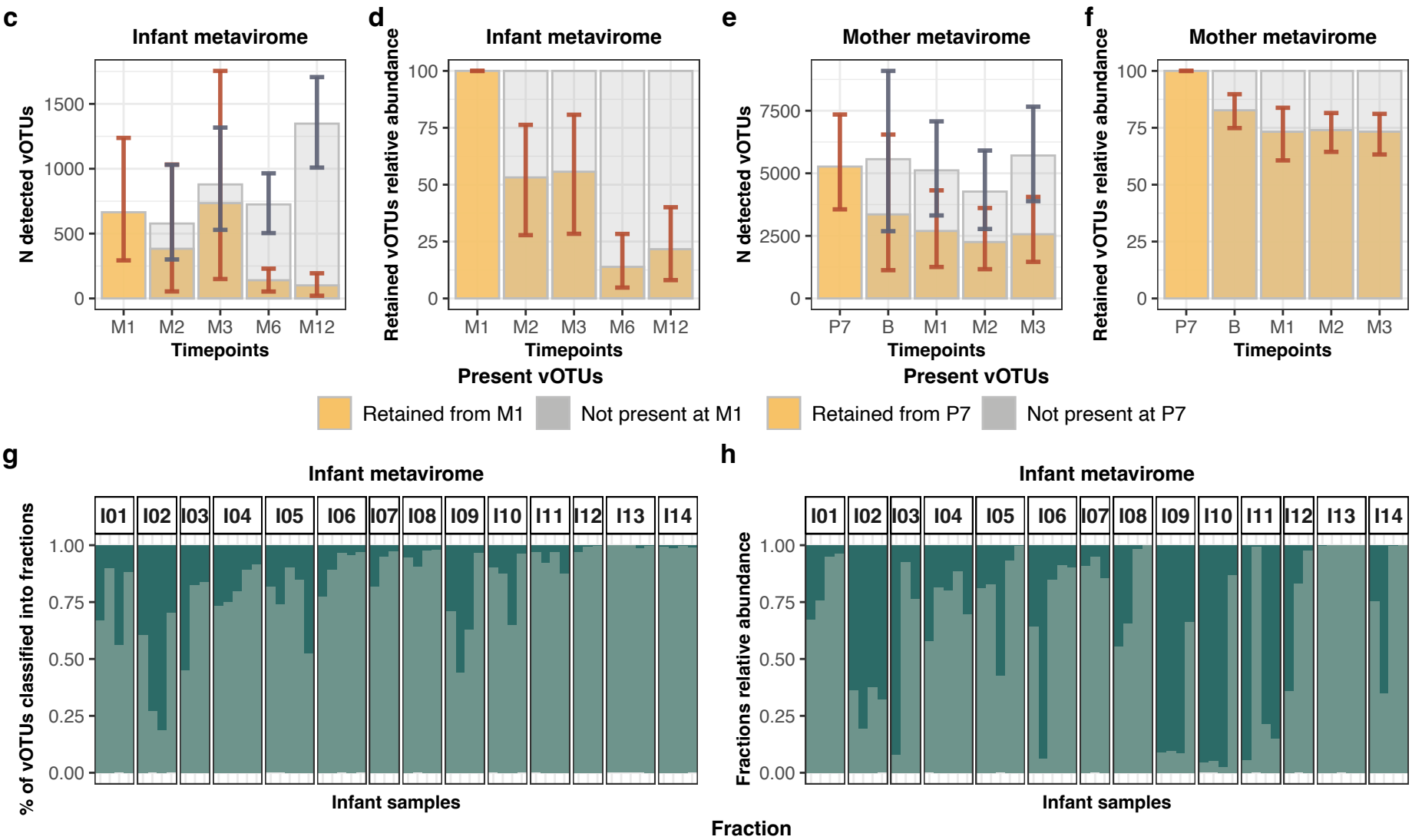
[9]Department of Midwifery Science, Amsterdam University Medical Center, Vrije Universiteit Amsterdam AVAG/Amsterdam Public Health and Department of General Practice & Elderly Care Medicine, University Medical Center Groningen, section Midwifery Science AVAG, University of Groningen, 9700 RB, Groningen, the Netherlands.

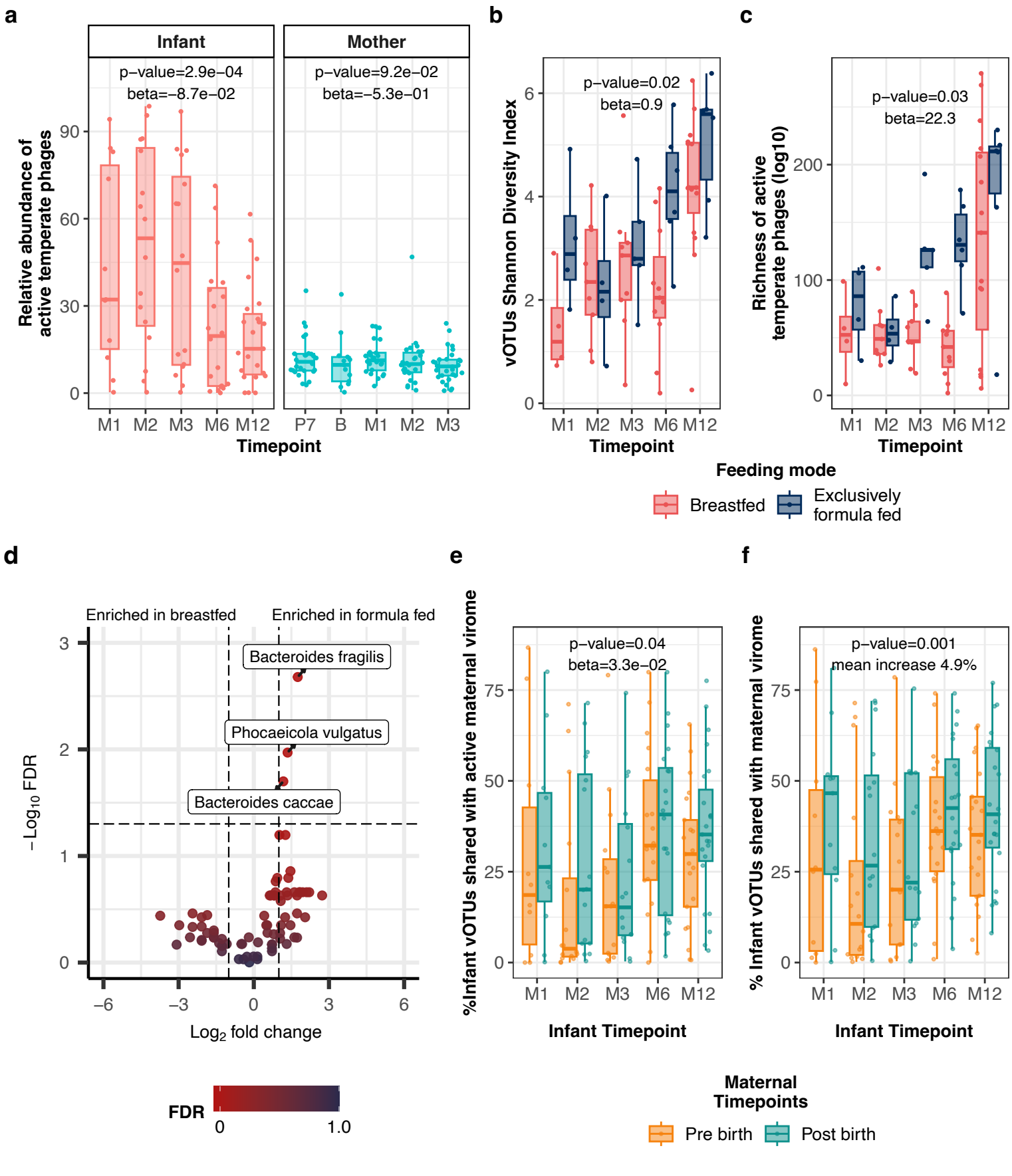











a
Infant
Mother
p−value=2.9e−04
beta=−8.7e−02
p−value=9.2e−02
beta=−5.3e−01
Relative abundance of active temperate phages
0
30
60
90
M1 M2 M3 M6 M12
P7 B M1 M2 M3
Timepoint
b
p−value=0.02
beta=0.9
vOTUs Shannon Diversity Index
0
2
4
6
M1 M2 M3 M6 M12
Timepoint
c
p−value=0.03
beta=22.3
Richness of active temperate phages (log10)
0
100
200
M1 M2 M3 M6 M12
Timepoint
Feeding mode
Breastfed
Exclusively formula fed
d
Enriched in breastfed
Enriched in formula fed
Bacteroides fragilis
Phocaeicola vulgatus
Bacteroides caccae
−Log10 FDR
0
1
2
3
−6 −3 0 3 6
Log2 fold change
FDR
0
1.0
e
p−value=0.04
beta=3.3e−02
%Infant vOTUs shared with active maternal virome
0
25
50
75
M1 M2 M3 M6 M12
Infant Timepoint
f
p−value=0.001
mean increase 4.9%
% Infant vOTUs shared with maternal virome
0
25
50
75
M1 M2 M3 M6 M12
Infant Timepoint
Maternal Timepoints
Pre birth
Post birth