**SUPPLEMENTARY INFORMATION**

**This PDF contains:**

**Supplementary Note 1: Role of breastfeeding in the neonatal period and infancy**

Breastfeeding has recently been reported as the most important factor shaping the gut microbiota during infancy (as measured from 3 months of life)[1,2]. In the BBS cohort, breastfeeding status became statistically significant only from day 7 and exhibited smaller effect (day 7, R2=0.742%, P=0.0043; exclusive breastfeeding R2=1.691%, p=0.0035; day 21, R2=0.777%, P=0.0234; exclusive breastfeeding R2=1.868%, p= 0.0070) than mode of delivery (day 7, R2=3.972%, P=0.0043; day 21, R2=2.408%, P=0.0234) during the neonatal period. After progressing from neonatal to infancy period (around 8 months of age in this cohort), the impact of breastfeeding became comparable to the impact of mode of delivery in the infancy period (mode of delivery R2=1.004%, P=0.0240; breastfeeding, R2=0.894%, P=0.0390).

Whilst this finding does not directly validate the central claim of breastfeeding status as the predominant factor of the gut microbiota from months 3 to 14 of life observed in the TEDDY 16S cohort [1], the infancy period $R^2$ values in this study (up to 2.338% in vaginal-born babies) are comparable to the TEDDY metagenomics cohort[2] that used the same method (PERMANOVA) to calculate effect size (months 3-6, R2=1.69%, p=0.0015; months 7-10, R2=1.84%, p=0.0015), in comparison to the TEDDY 16S cohort[1] which determined R2=~10% based on a different method *EnvFit.*

In addition to differences that might originate from using different statistical analysis methods, the prevalence of breastfeeding also differs between the Baby Biome Study and the TEDDY study. Over 80% of the BBS babies were exposed to breastfeeding during the neonatal period and remained at this high proportion (86.64%) in the infancy period (8.75 ± 1.98 months), in contrast to 53% of the TEDDY babies during the same period (months 7-10). UK born BBS babies seemed to enter the weaning phase much later than the published large cohorts that highlighted the importance of breastfeeding cessation on microbiota

maturation. Only 13.36% of the BBS babies reached to the cessation of breast milk by the point of infancy sampling, in contrast to 46% of the TEDDY babies (27% in months 3-6, 72% in months 11-14) and 86% in the Swedish cohort[3] (12-month-old infant). To fully assess the effect of breastfeeding during later developmental stages in the BBS cohort, future continuous longitudinal samplings of the infancy gut microbiota and detailed dietary information (including the proportional/quantitative weighting of breast/non-breasting) are warranted.

We did not observe substantially greater effect from breastfeeding after stratification by vaginal and C-section babies separately, with no statistical power to detect significant microbiota compositional difference (by feeding mode) across time points. We found that the effect sizes of breastfeeding (exclusive and partial) were comparable in vaginal and C-section babies during the neonatal period (Supplementary Table 2), suggesting that breastfeeding did not impact the gut microbiota differently according to modes of delivery, during the first few weeks after birth.

Interestingly, as the baby progressed from neonatal to infancy period, breastfeeding status began to exhibit greater effect which was only detected in vaginally born babies ($R^2$=1.45% on day 21, $R^2$=2.34% in infancy), with no significant effect observed in C-section born babies in these late sampling time points (Supplementary Table 2). The vaginally delivered babies who stopped breastfeeding at the time of infancy sampling carried significantly lower abundance of commensal-associated *Bifidobacterium* and *Lactobacillus* (relative abundance 48.7% vs 26.0%, p=0.014 and 3.48% vs 0.69%, p=0.004, respectively), whilst high abundance of *Bifidobacterium* was significantly associated with breastfeeding as observed in the TEDDY cohort[1,2].

In summary, our results agreed with previous findings that continuous exposure to breastfeeding during infancy might have a greater lasting effect[1,2], in comparison to the immediate impact of a one-time event such as birth.

**Supplementary Note 2: The under-reported carriage of opportunistic pathogens in the previous term and preterm infant cohorts**

We found that in previous neonatal and infancy metagenomic studies, C-section babies did carry a higher level of opportunistic pathogens, but these were not reported due to small effect size (infancy) or inadequate statistical power in cohorts with much smaller sample sizes (neonatal period sampling). Among the few neonatal gut microbiome datasets in public repositories, we focused on two most relevant studies in terms of the sampling period (<30 days of age) and sample size (≥100 cross-sectional samples), namely a Swedish WGS-based cohort[3], and a US 16S amplicon-sequencing-based cohort[4].

In the Swedish cohort[3], the authors presented the data of 95 neonatal (referred as 'newborn' in the paper) samples from 80 vaginal and 15 C-section babies. Among the species/OTUs with the greatest difference in relative abundance between vaginal and C-section babies, four of the top five were opportunistic pathogen species, although they were not necessarily supported by statistical significance (Wilcoxon rank-sum test). These opportunistic pathogens, some overlapped with the ones identified in this study, included *Enterococcus faecalis* (vaginal 2.64%, C-section 7.68%, p = 0.0980), *Enterobacter aerogenes (*vaginal 0.00%, C-section 5.25%, p = 0.5362*), Enterobacter hormaechei/Enterobacter cancerogenus (*vaginal 0.01%, C-section 12.12%, p = 0.0277) and *Haemophilus parainfluenzae (*vaginal 1.28%, C-section 7.95%, p = 1.46E-07). Furthermore, *C. perfringens* was also found to be more abundant in C-section newborns (vaginal 0.55%, C-section 1.67%, p = 0.17).

In the 16S cohort[4] of 116 neonatal samples (meconium) from 83 vaginal and 33 C-section babies, we found that three out of the five most differently abundant genera in C-section

(versus vaginal) babies were associated with known opportunistic pathogen species. These genera included *Neisseria* (incl. *N. meningitidis/gonorrhoeae*; vaginal 0.17%, C-section 3.21%, p = 0.0107), *Klebsiella* (vaginal 5.77%, C-section 8.01%, p = 0.613) and *Serratia* (incl. *S. marcescens*; vaginal 0.001%, C-section 1.16%, p = 0.9075), which could not be identified at the species/strain level with the detection limit of 16S amplicon sequencing. Previous cohorts of comparable (or larger) sample sizes such as the DIABIMMUNE study[5] (sampled from 2 months of age) and the TEDDY study[1], sampled from 3 months of age), were not directly relevant as they exclusively sampled a different sampling window period (post-neonatal infancy). Although these studies did not sample the neonatal period (first month of life), the C-section babies in these cohorts were also enriched in potential opportunistic pathogens such as *Enterococcus*[1] and *Clostridium/Clostridium perfringens*[1,5] during the first year of life. Overall, in these cohorts, the differences in the levels of opportunistic pathogens were very small or not statistically significant in babies sampled in post-neonatal infancy period, which is consistent with our observations with the UK babies when sampled in infancy.

Prior to this study, the colonisation of opportunistic pathogens in the infant's gut has also been observed in hospitalised preterm and low-birthweight newborns who were densely sampled during the very early life period[6-10]. Furthermore, asymptomatic colonisation of hospital-associated *C. difficile* was known to be common among neonates[11], in particular in those delivered via C-section[12]. By sampling the full-term healthy babies in the very same early window, we have extended the very similar opportunistic pathogen colonisation pattern to full-term, hospital-born neonates with those both by C-section being predisposed to the very prevalent carriage. This result highlights the underappreciated important role of healthcare environment-associated opportunistic pathogens in seeding the gut microbiota of hospital-born babies.

**Supplementary Note 3: AMR and virulence genes in WGS isolates**

Focusing on the most prevalent opportunistic pathogen in C-section born babies, we analysed the genomes of a diverse population of BBS *E. faecalis* strains in the context of publicly available genomes of human and environmental strains (Fig. 4c). We found that 53.9% of the BBS strains were represented by five major lineages, each of which was distributed across vaginal and C-section babies and mothers in the three BBS hospitals (Extended Data Fig. 5a) and UK hospital patients, but did not include high-risk UK epidemic lineages enriched in multi-drug resistance (MDR) and virulence[13] (Fig. 4c). In congruence with the phylogenetic placement of the BBS strains with the human gastrointestinal and environmental strains, these non-epidemic *E. faecalis* exhibited comparable levels of carriage of AMR genes, although the BBS strains encoded a higher number of virulence factors such as aggregation substance, surface adhesins, hyaluronidase, and the toxin cytolysin (Extended Data Fig. 5b-c). While all *E. faecalis* are intrinsically resistant to cephalosporins[14] routinely used in IAP and C-section, 21.6% (77/356) of the BBS strains also carried genes encoding AMR to aminoglycosides commonly used in enterococcal infections (Extended Data Fig. 5d-e). Similar to *E. faecalis*, the BBS *Enterobacter* and *Klebsiella* strains also exhibited high-level population diversities even within individual hospitals (Extended Data Fig. 6a-c). The strains cultured from vaginal and C-section babies and mothers are spread across phylogenetically distant lineages (Fig. 4d, Extended Data Fig. 6d-f), and under-represented in the dominant UK epidemic lineages (*E. cloacae* VIII, $p=0.0043$; KoII, $p<0.0001$; KpI, $p=0.0059$; Fisher's exact test), in line with the epidemiological pattern of these environmental opportunistic pathogens[15-17]. While these Gram-negative opportunistic pathogens are also known to exhibit intrinsic resistance to penicillins and most cephalosporins used in IAP and C-section, genes encoding for AmpC, class A and extended-spectrum beta-lactamases (ESBL) were detected in most of the BBS *E. cloacae* (*bla*$_{ACT}$, 88.5%), *K. oxytoca* (*bla*$_{OXY}$, 100%) and *K.*

*pneumoniae* (*bla*$_{SHV}$, 82.1%) strains, respectively (Extended Data Fig. 7a-b). With the exception of tetracycline resistance enriched in the BBS *K. oxytoca*, the prevalence of AMR and virulence gene in the BBS strains are mostly representative of the non-epidemic lineages circulating in hospital environments and healthy populations (Extended Data Fig. 7), rather than selective colonisation of the hypervirulent and ESBL-enriched (*bla*$_{SHV}$, *bla*$_{CTX-M}$, *bla*$_{TEM}$) epidemic lineages. These results are consistent with the phylogenetic under-representation of epidemic lineages in the BBS collection, and also in line with our observations with *E. faecalis*.

**Supplementary Note 4: Intra-individual persistence and maternal transmission in WGS cultivated strains**

For opportunistic pathogen species, the proportion of maternally transmitted strains (57.14%) is significantly lower in comparison to the known typical maternally-transmitted *Bacteroides/Parabacteroides* species (93.35%, Fisher's exact test, $p<0.0001$). Furthermore, the total number of transmissions of opportunistic pathogens detected in this study represented only a tiny proportion (~2%) of the total neonatal transmissions (n=657) and typable sample-species pairs (n=995) across all analysed species (Supplementary Table 4), indicating that maternal transmission of the opportunistic pathogen species was very rare overall. However, this analysis was limited to <10% of the total mother-neonate pairs (n=178) due to rare and very low-level carriage in healthy adults with sufficient sequencing coverage required by metagenomic strain transmission analysis.

In the majority of the babies (78.53%, 95% CI 66.51-90.54%) carrying one of these opportunistic pathogens, identical strains were cultured across multiple longitudinal samples indicating stable colonisation of a single strain. Importantly, 11 out of 11 neonatal *Enterobacter* and *Klebsiella* strains were phylogenetically distant to the equivalent species isolated from their mothers (Extended Data Fig. 7a-c).

This result corroborates the metagenomic strain analysis, in which four opportunistic pathogen species (*E. faecalis/faecium*, *K. oxytoca/pneumoniae*, 9 transmissions, 12 non-transmissions) accounted for only ~2% of the total maternal transmission events with a significantly lower transmission rate (57.14%) compared to known maternally-transmitted *Bacteroides/Parabacteroides* strains (93.35%, Fisher's exact test, $p<0.0001$, Supplementary Table 4). Furthermore, the total number of transmissions of opportunistic pathogens detected in this study represented only a tiny proportion (~2%) of the total neonatal transmissions (n=657) and typable sample-species pairs (n=995) across all analysed species (Supplementary Table 4), indicating that maternal transmission of the opportunistic pathogen species was very rare overall. However, this analysis was limited to <10% of the total mother-neonate pairs (n=178) due to rare and very low-level carriage in healthy adults with sufficient sequencing coverage required by metagenomic strain transmission analysis. Overall, there is very limited evidence from metagenomics and cultured isolate WGS data to suggest that the opportunistic pathogens were maternally derived.

**Supplementary Note 5: Assessment of the impact of short-term storage on the faecal microbiota**

In order to assess the feasibility of multi-centre faecal sample collection, as part of the pilot study protocol design, we performed benchmarking experiments to assess whether faecal samples can be sent by post without significant loss of sensitivity in microbiota analysis. In practice, we sought to determine the impact of short-term storage on the faecal microbiota diversity and composition by varying the time between sample collection and sample initial processing (DNA extraction), and by varying the storage temperature prior to extraction (room temperature vs 4°C).

In this pilot study, six participants (mothers, n=3; babies, n=3) were recruited with informed written consent at the University College London Hospital (UCLH) maternity unit between

February and July 2014. The stool samples were collected at the hospital and then divided into aliquots. DNA was extracted from one aliquot (fresh sample) immediately, and the other samples were stored at ambient temperature and 4°C prior to DNA extraction for 2, 4 and 7 days, respectively. Please refer to the main study SOP for detailed sampling and processing protocols[18]. 16S rRNA sequencing (MiSeq) was used to assess the bacterial population composition of the faecal samples. The V1-V2 hypervariable region of the bacterial 16S rRNA gene was amplified from faecal DNA extracts using the barcoded fusion primers: MiSeq-27F and MiSeq-338R. Illumina-based partial 16S rRNA sequences were processed and sequenced using mothur v1.34.1 and the MiSeq SOP (http://www.mothur.org/wiki/MiSeq_SOP). Low-quality sequences with less than 500 bp sequence length and sequences that had homopolymers longer than 7 bases were removed. Chimeric sequences were removed in mothur using Perseus algorithm. The operational taxonomic unit (OTU) clusters were defined based on the 97% sequence identity cut-off. Results of mothur sequencing analysis were analysed in R v3.2.1 and statistical tests performed in GraphPad Prism v6.

In order to examine the potential loss in species richness and diversity during storage, we stratified by sample storage time and performed a pairwise comparison of alpha diversities between freshly processed faecal samples with those stored at ambient temperature and 4°C, respectively. We observed no statistical difference in Shannon indexes of the faecal samples stored at ambient temperature and 4°C for 2 or 4 days, in comparison to the same fresh samples processed immediately (paired Wilcoxon tests for 2-day storage at ambient temperature versus fresh, p = 0.0840; at 4°C versus fresh, p = 0.2324; 4-day storage at ambient temperature versus fresh, $p = 0.1309$, at 4°C versus fresh, p = 0.3750). Storing at ambient temperature for 7 days did result in a statistically significant decrease in alpha diversity, but not for 7-day storage at 4°C (paired Wilcoxon tests for ambient temperature

versus fresh, p = 0.0059; at 4°C versus fresh, p = 0.0840). These data suggest that the faecal microbiota richness would not be significantly affected by short term transport/storage for 2-4 days at ambient temperature or at 4°C for up to 7 days.

To establish whether the composition of maternal faeces was affected by storage temperature prior to DNA extraction, we compared the taxonomic relative abundances at the genus level for freshly processed samples and those split and stored at either 4°C or ambient temperature for up to 7 days. We observed no significant difference in the relative abundances of the 20 most prevalent genera (prevalence >50% across all samples, including *Bacteroides*) between different storage temperatures for samples stored for either 2, 4 or 7 days, in comparison to freshly processed samples (Wilcoxon matched-pairs signed-rank tests, $p < 0.05$).

**Supplementary Note 6: Microbiota measurements are robust to sequencing depth variation**

We first assessed the impact of sequencing depth on the microbiota species richness and microbial composition. We did not observe any strong impact on the Shannon diversity (Spearman correlation coefficient 0.1220, 95% CI 0.07322-0.1703, $p < 0.05$) and sequencing depth explained only 1.29% variation (PERMANOVA, $p < 0.05$) on the microbial composition.

We also did not observe any statistical difference in the sequencing depth (before or after the quality and contaminant trimming) between vaginal and C-section samples across all sampling age groups except on day 4. Where microbiota species richness measurements were directly relevant in mother-baby species and strain sharing analysis, we found that observed species richness differences between vaginal and C-section born babies were not affected by sequencing depth variation. For the day 4 gut metagenomes, we observed no correlation between the sequencing depth and species richness; the Spearman correlation coefficient between sequencing depth and Shannon diversity index was -0.002028 (P=0.9716), and

-0.06792 (P=0.2331) with the species count (above 0.01% relative abundance used in species sharing measurement). Despite having slightly lower sequencing depth on day 4, the vaginal-born neonatal gut microbiotas still exhibited higher species richness on day 4, as measured by both the species count and alpha (Shannon) diversity (data not shown). We further performed ANOVA tests to confirm that sequencing depth (post trimming) did not interact with the delivery mode in explaining the variance of alpha diversity (P = 0.59308) nor species count (P = 0.9872) of day 4 gut metagenomes. These data suggest that the species richness measurements of this study are robust to sequencing depth variation, which was only observed in day 4 samples, and did not impact on the observed microbiota species and strain differences between vaginal and C-section born babies.

Lastly, we have also considered sequencing depth as a potential confounding covariate in multivariate analysis of clinical covariates with individual bacterial taxa (MaAsLin results, **Supplementary Table 3**), which did not preclude our main findings regarding the enrichment of opportunistic pathogen and the depletion of *Bacteroides* species.

## References


1. Stewart, C. J. *et al.* Temporal development of the gut microbiome in early childhood from the TEDDY study. *Nature* **562,** 583–588 (2018).
2. Vatanen, T. *et al.* The human gut microbiome in early-onset type 1 diabetes from the TEDDY study. *Nature* **562,** 589–594 (2018).
3. 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).
4. Chu, D. M. *et al.* Maturation of the infant microbiome community structure and function across multiple body sites and in relation to mode of delivery. *Nat. Med.* (2017). doi:10.1038/nm.4272
5. Vatanen, T. *et al.* Variation in Microbiome LPS Immunogenicity Contributes to Autoimmunity in Humans. *Cell* **165,** 1551 (2016).
6. Gibson, M. K. *et al.* Developmental dynamics of the preterm infant gut microbiota and antibiotic resistome. *Nature Microbiology* **1,** –10 (2016).
7. Raveh-Sadka, T. *et al.* Gut bacteria are rarely shared by co-hospitalized premature infants, regardless of necrotising enterocolitis development. *eLife Sciences* **4,** e05477 (2015).
8. Raveh-Sadka, T. *et al.* Evidence for persistent and shared bacterial strains against a background of largely unique gut colonisation in hospitalised premature infants. *The ISME Journal* **10,** 2817–2830 (2016).
9. Olm, M. R. *et al.* Identical bacterial populations colonise premature infant gut, skin, and oral microbiomes and exhibit different in situ growth rates. *Genome Res.* **27,** gr.213256.116–612 (2017).
10. Rose, G. *et al.* Antibiotic resistance potential of the healthy preterm infant gut microbiome. *PeerJ* **5,** e2928 (2017).
11. Rousseau, C. *et al.* Clostridium difficile carriage in healthy infants in the community: a potential reservoir for pathogenic strains. *Clin. Infect. Dis.* **55,** 1209–1215 (2012).
12. Penders, J. *et al.* Factors Influencing the Composition of the Intestinal Microbiota in Early Infancy. *Pediatrics* **118,** 511–521 (2006).
13. Raven, K. E. *et al.* Genome-based characterisation of hospital-adapted Enterococcus faecalis lineages. *Nature Microbiology* **1,** 15033 (2016).
14. Gilmore, M. S. *et al.* Enterococcal Infection—Treatment and Antibiotic Resistance. *Enterococci: From Commensals to Leading Causes of Drug Resistant Infection [Internet]* (2014).
15. Moradigaravand, D., Reuter, S., Martin, V., Peacock, S. J. & Parkhill, J. The dissemination of multidrug-resistant Enterobacter cloacae throughout the UK and Ireland. *Nature Microbiology* **1,** 16173 (2016).
16. Moradigaravand, D., Martin, V., Peacock, S. J. & Parkhill, J. Population Structure of Multidrug-Resistant Klebsiella oxytoca within Hospitals across the United Kingdom and Ireland Identifies Sharing of Virulence and Resistance Genes with K. pneumoniae. *Genome Biology and Evolution* **9,** 574–584 (2017).
17. Moradigaravand, D., Martin, V., Peacock, S. J., Parkhill, J. & Chiller, T. Evolution and Epidemiology of Multidrug-Resistant Klebsiella pneumoniae in the United Kingdom and Ireland. *MBio* **8,** e01976–16 (2017).
18. Bailey, S. R. *et al.* A pilot study to understand feasibility and acceptability of stool and cord blood sample collection for a large-scale longitudinal birth cohort. *BMC Pregnancy Childbirth* **17,** 439 (2017).