nature microbiology

Article

https://doi.org/10.1038/s41564-024-01804-9

Primary succession of Bifidobacteria drives pathogen resistance in neonatal microbiota assembly

Received: 22 April 2024

Accepted: 5 August 2024

Published online: 06 September 2024

Check for updates

Yan Shao[1]✉, Cristina Garcia-Mauriño[2], Simon Clare[1], Nicholas J. R. Dawson[1], Andre Mu[1], Anne Adoum[1], Katherine Harcourt[1], Junyan Liu[1], Hilary P. Browne[1], Mark D. Stares[1], Alison Rodger[2], Peter Brocklehurst[3], Nigel Field[2] & Trevor D. Lawley[1]✉

Human microbiota assembly commences at birth, seeded by both maternal and environmental microorganisms. Ecological theory postulates that primary colonizers dictate microbial community assembly outcomes, yet such microbial priority effects in the human gut remain underexplored. Here using longitudinal faecal metagenomics, we characterized neonatal microbiota assembly for a cohort of 1,288 neonates from the UK. We show that the pioneering neonatal gut microbiota can be stratified into one of three distinct community states, each dominated by a single microbial species and influenced by clinical and host factors, such as maternal age, ethnicity and parity. A community state dominated by *Enterococcus faecalis* displayed stochastic microbiota assembly with persistent high pathogen loads into infancy. In contrast, community states dominated by *Bifidobacterium*, specifically *B. longum* and particularly *B. breve*, exhibited a stable assembly trajectory and long-term pathogen colonization resistance, probably due to strain-specific functional adaptions to a breast milk-rich neonatal diet. Consistent with our human cohort observation, *B. breve* demonstrated priority effects and conferred pathogen colonization resistance in a germ-free mouse model. Our findings solidify the crucial role of Bifidobacteria as primary colonizers in shaping the microbiota assembly and functions in early life.

Human gut microbiota colonization commences immediately at birth when neonates are exposed to microorganisms from the surrounding environment and maternal sources (for example, gut[1–5], vagina[2–4], skin[3,4], breast milk[3,6]). We recently reported in the UK Baby Biome Study (BBS) that maternal transmission of primary colonizers, such as commensal *Bifidobacterium* and *Bacteroides* species, is disrupted in caesarean-section (CS) and antibiotic-exposed births, instead predisposing the neonatal gut microbiota (NGM) to colonization by antibiotic resistant healthcare-associated pathogens[1]. This observation suggests the possibility of 'priority effects' in human gut microbiota assembly, which posits the arrival order of primary colonizer species determines the outcome of the microbiota assembly during a primary ecological succession (from sterile to complex communities)[7,8]. The NGM represents the earliest window of opportunity for intervention with probiotics or prebiotics to prevent or restore impaired microbiota development. However, little is known about the ecological priority

[1]Host–Microbiota Interactions Laboratory, Wellcome Sanger Institute, Hinxton, UK. [2]Institute for Global Health, University College London, London, UK. [3]Birmingham Clinical Trials Unit, University of Birmingham, Birmingham, UK. ✉e-mail: ys4@sanger.ac.uk; tl2@sanger.ac.uk

effects in the NGM assembly due to a lack of high-resolution, longitudinal human microbiome data from the neonatal period (that is, the first month of life).

Results

To comprehensively examine NGM assembly dynamics, we expanded on phase 1 of our BBS cohort (BBS1)[1,9] with an additional 688 neonatal participants (primarily day 7) in phase 2 (BBS2), effectively doubling our sampling effort. A large-scale, longitudinal metagenomic characterization of the combined BBS dataset, comprising 2,387 gut microbiota samples from 1,288 healthy UK neonates (≤1 month), enabled us to study neonatal microbiota assembly with unparalleled scale and resolution (Extended Data Fig. 1a,b and Supplementary Tables 1–3). To investigate the origin and both short-term and long-term stability of the NGM primary colonizers, we utilized three subgroups from the expanded BBS2 cohort. These included (1) 183 neonate–mother pairs (representing 14% of participants), referred to as investigating 'maternal transmission'; (2) 359 participants with longitudinal sampling within the neonatal period (median = 3 samples per participant on days 4, 7 and 21; representing 28% of participants), referred to as investigating 'neonatal longitudinal colonization'; and (3) 302 participants with paired samples taken both in the neonatal period and later in infancy (at 8.75 ± 1.98 months; representing 23% of participants), referred to as investigating 'infancy persistence' (Extended Data Fig. 1c).

Complementing the increased sample size, we have also updated extensive, high-quality clinical and sociodemographic metadata harmonized from BBS clinical record forms and hospital electronic records (Methods), thereby facilitating robust statistical and epidemiological assessment of primary succession patterns. Most neonates in this cohort (84.5%, $N = 836$) were at least partially breastfed by their mothers, with 44.1% being exclusively breastfed ($N = 436$). A large majority of participants at the time of infancy sampling were still being breastfed (86.2%; $N = 199$), with very few fully weaned (0.87%; $N = 2$). Only 11.3% ($N = 123$) received postnatal antibiotics during the first week of life (Supplementary Table 4).

Three community states in the neonatal gut microbiota

To delineate the primary succession patterns of the NGM, we sought to identify the primary colonizers driving gut microbial community structure during the neonatal period. Applying partitioning around medoids (PAM) clustering to 1,904 BBS neonatal gut metagenomes at the species level revealed an optimal clustering of three within the NGM, hereafter referred to as 'NGM community states'[10] (Fig. 1a and Extended Data Fig. 2a,b). These three community states were further validated by another widely used microbial community typing method: the Dirichlet multinomial mixture (DMM) modelling framework (Extended Data Fig. 2c,d). Both the PAM and DMM-based approaches showed strong concordance in community state assignments (Cramér's V correlation of 0.726; Extended Data Fig. 2e) and core species compositions (Extended Data Fig. 2f). Notably, these three community states were consistently observed across the three main sampling points in the BBS cohort (days 4, 7 and 21), underscoring their representativeness of the neonatal period, irrespective of the timing of sample collection (Extended Data Fig. 3).

Three bacterial species, *Bifidobacterium longum* subsp. *longum* (BL), *Bifidobacterium breve* (BB) and *Enterococcus faecalis* (EF) acted as the taxonomic drivers for each community state (Fig. 1b and Extended Data Figs. 2g and 4). Each species dominated their respective NGM community states with a relative mean abundance of 56.5% for BB, 21.7% for EF and 27.2% for BL (Fig. 1c). Henceforth, they are referred to as NGM driver species with acronyms indicating each respective community state.

The observed single-species dominance of either *B. breve*, *B. longum* or *E. faecalis* in very early life can also be consistently observed in other cohorts, albeit underreported owing to the previous undersampling during the neonatal period (the largest sample size being <100). Evidence for this comes from diverse populations and methodologies, including 16S gene or quantitative PCR (qPCR)-based observations in Norway[11] ($N = 87$) and Denmark[12] ($N = 16$), as well as shotgun metagenomic surveys of neonates across industrialized urban populations similar to the UK BBS cohort in Europe (Sweden[13]), Asia (Israel[14]) and North America (the TEDDY cohort[15,16]) (Extended Data Fig. 5). Importantly, the NGM community states observed across industrialized cohorts are paralleled in non-industrialized populations. In a peri-urban cohort in South Asia (Bangladesh[17]), although *B. breve* continues to be a primary NGM driver species, the community states typically driven by *B. longum* and *E. faecalis* in industrialized settings are instead represented by closely similar species: *B. infantis* (closely related to *B. longum*) and *Escherichia coli* (sharing facultative anaerobic and opportunistic pathogenic traits with *E. faecalis*). Collectively, these cross-study validations strengthen the generalizability of our results in neonatal populations from different geographical regions and lifestyles beyond the UK, and using different methodologies.

Of note, *B. longum* subsp. *infantis* (*B. infantis*), which is closely related to BL and often used as an infant probiotic, was not identified as a driver species. It was rarely detected (~2% prevalence based on 0.5% relative abundance) in the BBS neonates[14]. The near absence of *B. infantis* in our UK neonatal cohort aligns with findings from other Western industrialized countries, including a recent meta-analysis[14] of cohorts from Israel, Sweden, Finland, Estonia, Italy and the USA[18], where there is little evidence of *B. infantis* naturally colonizing the gut microbiota of healthy, full-term infants. This underscores the importance of distinguishing between closely related species that exhibit very different host colonization patterns.

Applying metagenomic strain tracking analysis on the 'maternal transmission' subset, only *B. longum* exhibited evidence of maternal transmission, with all evaluable BL neonates (15 out of 15) harbouring the exact same *B. longum* strain found in their mothers' gut microbiota. This result, consistent with a recent global meta-analysis[19], strongly indicates the maternal gut microbiota as the main source of the BL community state (Extended Data Fig. 6). While we could have overlooked maternal transmission of very low-abundance *B. breve* and *E. faecalis* below the metagenomic strain detection limit, we consider it more likely that they originate from unsampled maternal (for example, *B. breve* in breast milk[20,21]) or environmental sources (for example, *E. faecalis* in the hospital birth environment[22,23]) previously implicated as potential sources of these species in the NGM.

The abundant dominance of single driver species was particularly pronounced in community state BB, in which *B. breve* constituted over half of the NGM by mean relative abundance, and exhibited the lowest microbial richness and evenness, as reflected by the alpha (Shannon) diversity (Fig. 1d). In comparison, the other two NGM community states, BL and EF, had higher microbial diversity, and other moderately abundant species frequently co-occurred with the driver species (Extended Data Fig. 2f,g); *B. longum* with commensal *E. coli*, *Bacteroides* and other *Bifidobacterium* species; *E. faecalis* with environment and skin-associated *Streptococcus*, *Staphylococcus* spp., as well as healthcare-associated opportunistic pathogens *Enterococcus*, *Klebsiella*, *Enterobacter* spp. and *C. perfringens*. Notably, these less-dominant species in EF were also known signatures of hospital CS birth not only in this UK cohort[1] but also in cohorts from North America[24,25], Latin America[24] and Europe[13,24,26].

Factors influencing the acquisition of the NGM community states

To determine the perinatal factors influencing the acquisition of each NGM community state, we performed epidemiological analyses using 20 high-quality clinical and sociodemographic metadata variables ($N = 1{,}108$ eligible participants; Fig. 2 and Supplementary Table 5). After adjusting for potential confounders in multivariate fixed-effect logistic

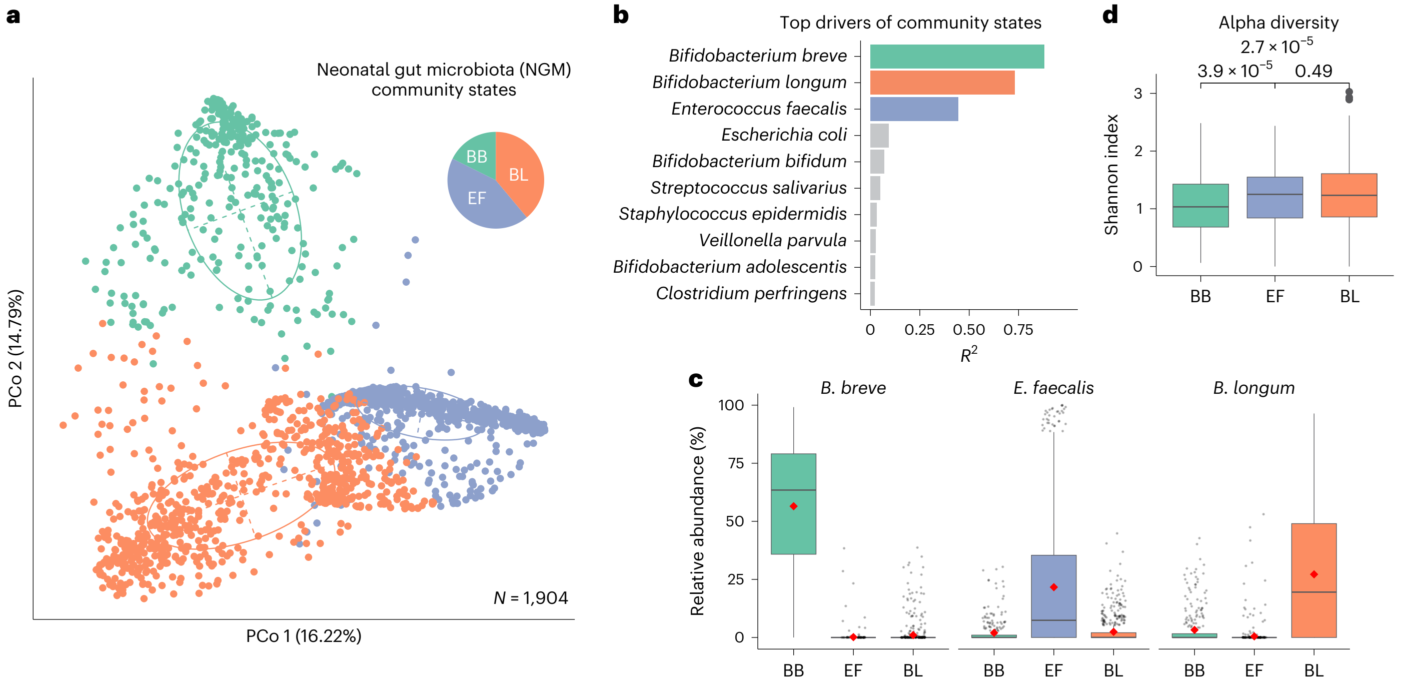


Fig. 1 | Dominant species driving three NGM community states. a, Principal coordinates analysis (PCoA) plots of 1,904 neonatal gut metagenomes sampled within the first 30 days of life and clustered using the PAM algorithm on the basis of species-level JSD. Three distinct NGM community states (optimal number clusters $k = 3$) were identified via PAM clustering. The inset pie chart displays the proportion of the three NGM community states, each labelled according to its primary driver species, namely *B. breve* (BB, green; $N = 336$, 17.6% of the samples), *E. faecalis* (EF, purple; $N = 827$, 43.4% of the samples) and *B. longum* (BL, orange; $N = 741$, 38.9% of the samples). Ellipses encapsulate 67% of the samples within each respective cluster. **b**, Top 10 driver species contributing to variation observed in the ordination space, as ranked by effect size ('envfit' R^2, false discovery rate (FDR)-corrected two-sided test, $P < 0.05$). **c**,**d**, Each NGM community state is dominated by a single driver species, as measured by the high relative abundance of the driver species (**c**) and the low alpha diversity (**d**) across the three NGM community states (FDR-corrected, two-sided Wilcoxon test). Boxplot centre line and red point indicate the median and mean, respectively; box limits indicate the upper and lower quartiles; and whiskers indicate 1.5× the interquartile range (BB $n = 336$, EF $n = 827$, BL $n = 741$).

regression models, we found that the acquisition of an EF community state was independently associated with being born via CS birth (compared to vaginal delivery (VD); adjusted odds ratio (AOR) = 2.30 [95% CI 1.34–3.95], $P = 0.003$; 70.5/23.6/40.0% among EF/BL/BB, respectively) and with the mother receiving intrapartum antibiotics during labour (AOR = 3.69 [95% CI 2.11–6.42], $P < 0.001$; 80.8/32.7/46.3% among EF/BL/BB, respectively). Conversely, being born via CS birth and labour antibiotics exposure were negatively associated with BL acquisition (AOR for CS vs VD = 0.36 [95% CI 0.21–0.64], $P < 0.001$; AOR for receiving antibiotics during labour = 0.46 [95% CI 0.26–0.79], $P = 0.005$, respectively).

Interestingly, several intrinsic host factors including sex (male with BB), maternal ethnicity (Asian with EF and BB), age (<30 and ≥40 with EF and BL, respectively) and parity (first time giving birth with EF) were also independently associated with specific community states. For example, mothers identifying as Asian (compared with white participants) were more likely to acquire BB (AOR = 2.11 [95% CI 1.32–3.38], $P = 0.006$) but less likely to acquire EF (AOR = 0.63 [95% CI 0.41–0.95], $P = 0.04$; 9.0/12.1/19.5% among EF/BL/BB, respectively). It is noteworthy that BB is the only community state that was exclusively influenced by host factors and independent of any clinical factors including mode of birth and antibiotics, which may suggest a distinct route of BB acquisition that remains unaffected by the perturbations associated with hospital births. These observations align with the hypotheses that maternal factors, such as genetic determinants of breast milk composition (for example, secretor status of the mothers)[27], a history of previous pregnancies or cohabitation with children[28], as well as cross-cultural differences in infant-care-associated behaviours[7] may influence the vertical transmission of maternal microbiota.

Neither postnatal antibiotics nor breastfeeding exposure, whether immediately after birth or within the first week of life, appeared to predispose neonates to any specific community state. This lack of association is probably attributed to the uniformly high-levels of antibiotic-free status (84.7/90.8/89.5% among EF/BL/BB, respectively) and breastfeeding rates (79.1/81.8/88.6% among EF/BL/BB, respectively) during the earliest postnatal window sampled in this cohort. The absence of an association between breastfeeding and EF also aligns with previous reports that, despite its antimicrobial properties, breast milk alone does not inhibit *E. faecalis* growth in vitro[29,30].

Priority effects in NGM community state stability

We reasoned that the three primary colonizers as NGM drivers could benefit from priority effects, which would be evident through the exclusion of, or replacement by, later-arriving species in the NGM. To search for evidence of such priority effects, we sought to examine the stability and temporal signals of both the NGM community states and their driver species in the 'neonatal longitudinal' subset, stratified by birth modes. Most VD neonates who initially acquired a *Bifidobacterium*-dominated community state (either 92% for BB or 89% for BL, 79% or 72% by considering transient switches between day 4 and 7) during week 1 retained their community state when resampled in week 3 (Fig. 3a and Extended Data Fig. 7a). By contrast, EF was the most unstable community state, with less than half of the neonates (29% in VD and 39% in CS) remaining in their early EF community state during the neonatal period (EF vs BB AOR 16.2 [95%CI 3.84–68.10], EF vs BL AOR 13.89 [4.02–48.02]; $P < 0.001$; Supplementary Table 6). Irrespective of birth mode, BB proved more stable than EF (pairwise chi-square test, corrected $P < 0.001$), while the sample size was insufficient to be confident about the relative stability of BL in CS neonates (65% versus 48% for EF; pairwise chi-square test, corrected $P = 0.52$).

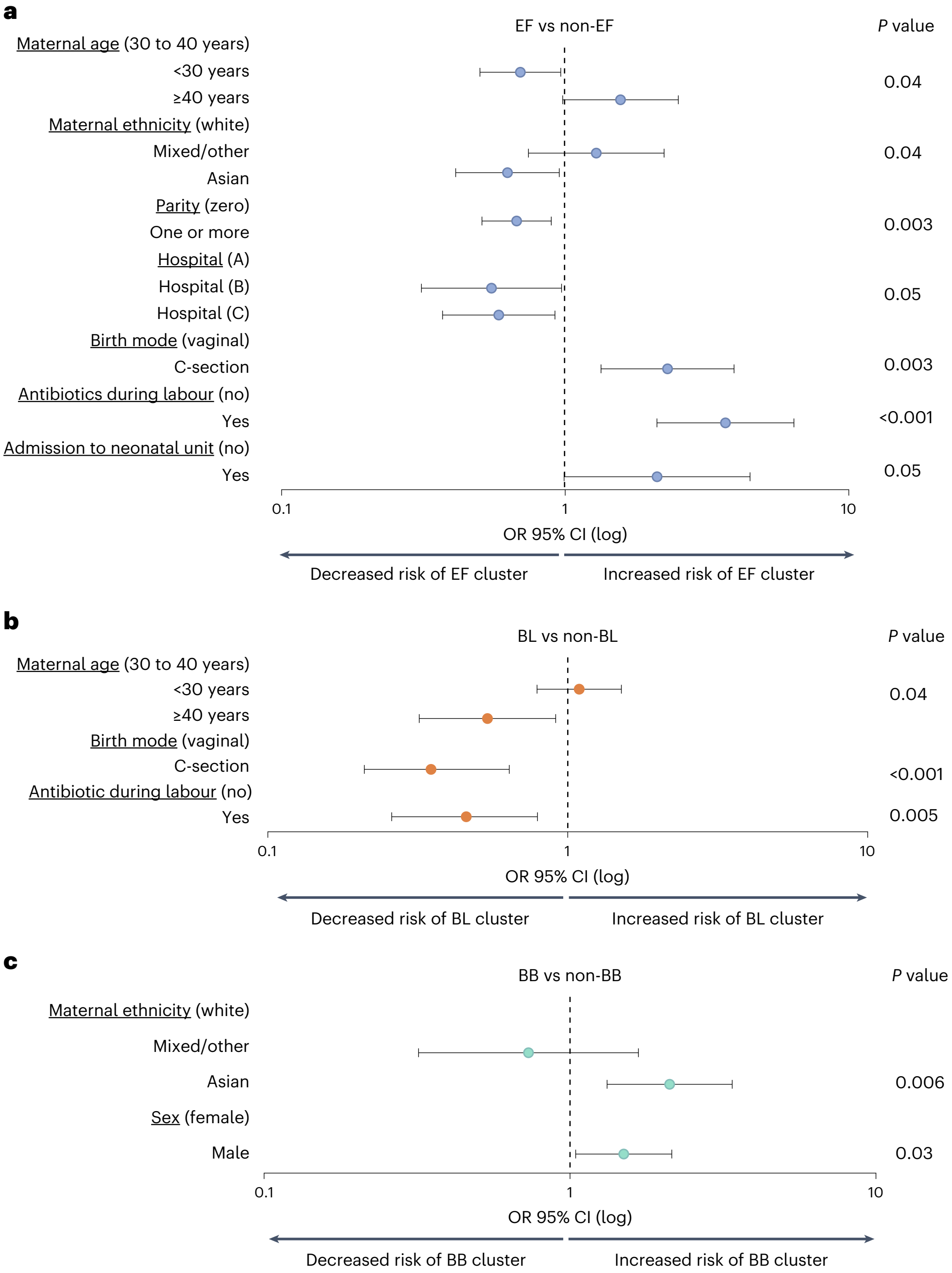


Fig. 2 | Clinical and sociodemographic variables associated with NGM community state in the first week of life (N = 1,108). a–c, Multivariate associations between clinical and sociodemographic variables and each week-1 NGM community state. Three different models were built: EF vs non-EF (**a**), BL vs non-BL (**b**) and BB vs non-BB (**c**). Likelihood ratio tests (two-sided) were used to calculate P values (without FDR correction), with $P \leq 0.05$ in the multivariate models displayed. Odds ratios (OR) are plotted on a $\log_{10}$ scale. For details of univariate and multivariate analyses, refer to Supplementary Tables 5 and 6. The week-1 NGM community state was identified for each eligible participant using the earliest available sample from week 1, either on day 4 (N = 64) or day 7 (N = 1,044).

The stability of the underlying driver species closely mirrored the observed community state dynamics. In contrast to *E. faecalis*, which rapidly declined throughout the stochastic assembly trajectory of the early community state EF, both *B. breve* and *B. longum* retained their high abundance within their respective community states throughout the 3-week neonatal sampling window (Fig. 3b and Extended Data Fig. 7b). Notably, both species, as late-arriving secondary colonizers (that is, colonized NGM only in week 3), exhibited signs of competitively excluding *E. faecalis* in CS neonates who initially acquired the EF community state (Fig. 3b). This competitive exclusion effect seemed most pronounced for *B. breve*; in contrast to *B. longum*, it was able to colonize VD neonates at increasing levels as a late-arriving species (Extended Data Fig. 7b). Among the primary colonizers that dominated the NGM in the first week, *B. breve* is the only species conferring durable colonization dominance (relative to the other driver species), which persisted as far as the final neonatal period sampling point at week 3 (P < 0.001 in VD and CS; Fig. 3c).

The stability of the two *Bifidobacterium* species is also reflected at the strain level (Extended Data Fig. 6); most of the neonates retained the same *B. longum* (79.5%, N = 35/44 BL neonates) or *B. breve* (75%, N = 24/32 BB neonates) strain they initially acquired throughout the neonatal period, in contrast to 62.3% for *E. faecalis* (N = 43/69 EF neonates;

the denominators represent longitudinally sampled individuals with detectable strain sharing events).

Together, as primary colonizers, both *Bifidobacterium* species benefit from priority effects, maintaining a stable NGM assembly trajectory owing to their ability to confer durable species dominance and inhibit the later arrival of opportunistic pathogens such as *E. faecalis*. In particular, *B. breve* exhibits stronger priority effects between the two species (that is, only as a primary colonizer), as well as strong deterministic exclusion of *E. faecalis* (that is, as either a primary or a secondary colonizer).

Stability of NGM driver species into infancy

We also assessed the longer-term engraftment of the NGM driver species in participants resampled 6–12 months beyond the neonatal period using the 'infancy persistence' subset. Remarkably, the relative dominance of *B. breve* (over the other driver species, in VD, $P < 0.05$; Fig. 3d) also extended into infancy when there was still no significant difference in breastfeeding rates between early NGM community states (BB/EF/BL: 88.4%/89.6%/80.5%, chi-square test, $P = 0.18$). In addition, the long-term competitive exclusion effect of *B. breve* was evident in CS neonates who either retained or transitioned into BB (primarily from EF) by week 3. These long-term stability patterns were exclusively observed for *B. breve*, with its abundance in infancy being almost double in neonates who previously had a BB community state compared with those with other community states (Fig. 3e).

Although NGM driver species rarely retained their differential abundance later in infancy (except *B. breve*), the frequency of carriage for all three driver species was consistently higher in infants stratified by their corresponding NGM community states (Fig. 3f). As many as 93% of VD (or 77% of CS) neonates with week-1 community state of BB still carried *B. breve*, compared with 58% and 66% (or 65% of CS) of VD neonates with week-1 community states EF and BL, respectively (pairwise chi-square tests, $P < 0.001$). While levels of *E. faecalis* in community state EF waned over time to non-differential levels later in infancy, neonatal acquisition of EF remains a predisposing factor for longer-term carriage of *E. faecalis*. This opportunistic pathogen species was still detected in higher proportions (44%) in neonates from the EF community state during their first week (relative to 37–41% in BB and 35–38% in BL) when resampled later in infancy, regardless of their birth mode (pairwise chi-square tests, $P < 0.001$; Fig. 3f).

EF state enriched with virulence and antibiotic resistance genes

To determine the functional differences among NGM community states, we leveraged their driver species as proxies for functional analyses, using 1,249 high-quality isolate ($N = 133$) and metagenome-assembled genomes ($N = 1,116$) generated from the corresponding community state samples (BB $N = 297$, EF $N = 561$, BL $N = 391$; Supplementary Table 7). We found a striking difference between *Bifidobacterium* spp. and *E. faecalis* functional profiles in antimicrobial resistance (AMR) and virulence potential. Importantly, all *E. faecalis* strain genomes recovered from neonates with EF community states encoded known virulence factors including 70% predicted to produce the toxin cytolysin[31]. By contrast, both *Bifidobacterium* driver species genomes displayed markedly reduced levels of AMR and virulence-associated genes, with a burden 10- to a 100-fold less than in EF (median 17 versus 0; Fig. 4a). Further AMR gene screening of the entire gut resistome within each community state revealed a higher carriage of high-risk AMR genes, such as CTX-M-15 linked to extended-spectrum beta-lactamase (ESBL), in both BL and EF community states (Fig. 4b). This underscores the notable pathogenic potential of ESBL-carrying Enterobacteriaceae pathogens co-occurring in non-BB community states. These findings align with our risk factor analyses (Fig. 2), which identified maternal antibiotics exposure during labour (to some VD and all CS neonates) as a strong risk for the acquisition of an EF community state that bears increased risk of AMR and virulence.

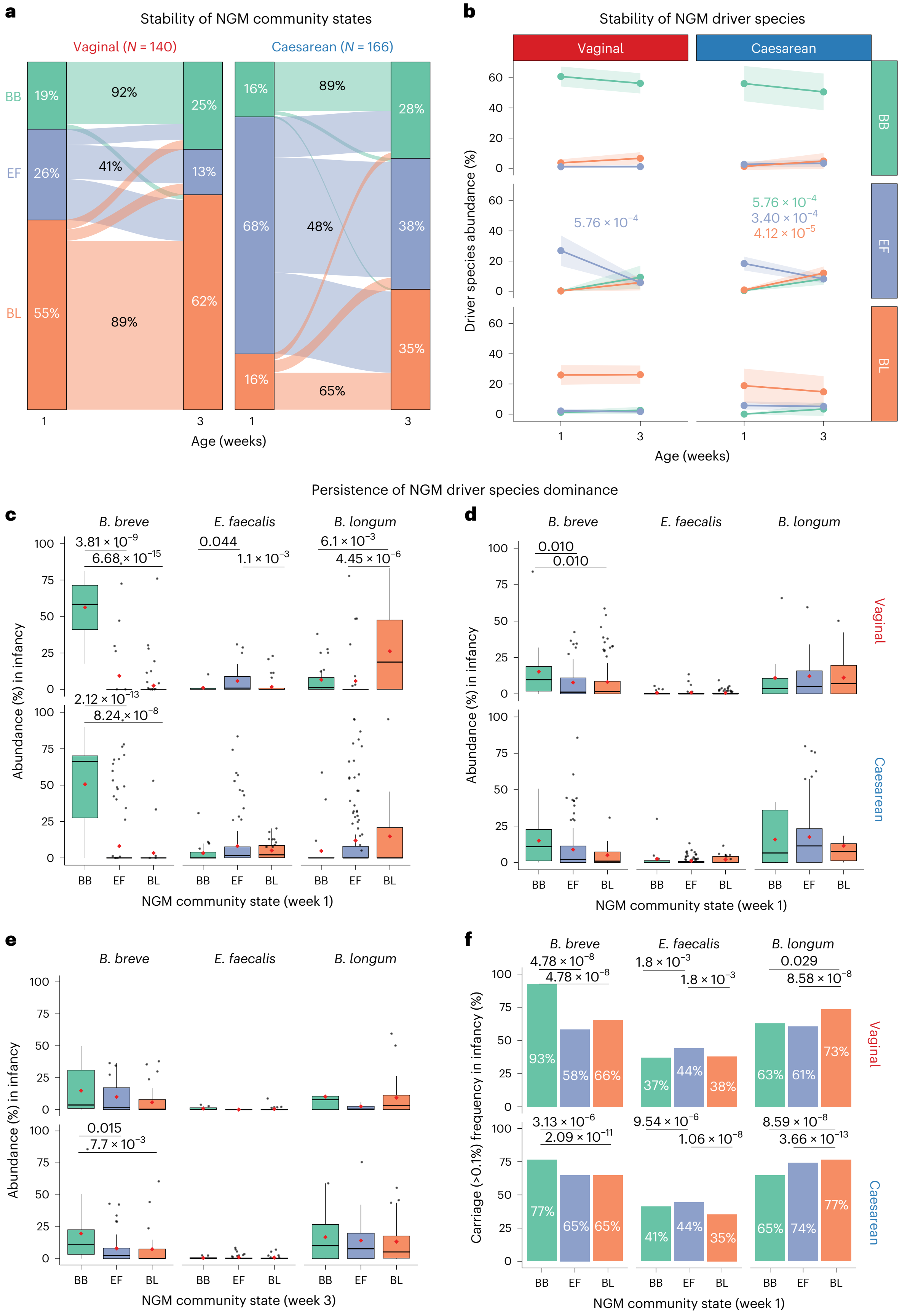
Pathogen resistance of *B. breve* via metabolic adaptation to HMOs

At the genome-wide functional level, we observed distinct metabolic landscapes of NGM community states based on KEGG orthologues (Extended Data Fig. 8a), particularly in metabolic repertoire of carbohydrate-active enzymes (Extended Data Fig. 8b). Both *Bifidobacterium* community states, in contrast to EF, exhibited an enrichment in carbohydrate-active enzymes associated with metabolizing human milk oligosaccharides (HMOs) abundant and exclusively found in human breast milk. By contrast, EF predominantly possesses genes tailored for utilizing complex dietary glycans such as mannan and chitin, as well as those like starch and cellulose that are commonly found in a plant-based diet usually consumed later in life (Extended Data Figs. 8b and 9).

Compared with the limited HMO metabolic capability of the BL community state, BB is capable of utilizing the all the major HMO substrates including lacto-*N*-tetraose (LNT), lacto-*N*-neotetraose (LNnT) and lacto-*N*-biose (LNB), as well as the primary end-products of HMO metabolism L-fucose and D-lactose, which are naturally present in human breast milk (Fig. 4c). Interestingly, among the three community states, only BB—comprising nearly all *B. breve* genomes (97.6%, $N = 290/297$)—encode the enzyme (α-L-fucosidase, GH95 or GH29) required for metabolizing the most abundant HMO component 2′-fucosyllactose (2′-FL). Although these *B. breve* strains lack known transporters for importing 2′-FL for intracellular metabolism, previous in vitro experiments have shown that similar strains are capable of growing on 2′-FL[32,33]. Therefore, *B. breve* might be able to metabolize 2′-FL via a previously uncharacterized pathway. In contrast, such capability is extremely rare among BL (5.0%, $N = 19/391$) and completely absent in EF (Fig. 4c). Notably, the species-level variations in HMO utilization observed in the study strains are representative of BB/BL/EF species, exhibiting patterns consistent with those previously reported[34]. These patterns are not influenced by breastfeeding rates in this neonatal cohort, which are uniformly high and statistically indistinguishable among the community states (79.1%, 81.8% and 88.6% for EF, BL and BB, respectively).

Fig. 3 | Dynamics and stability of NGM community states and their driver species. a,**b**, Stability of NGM community states (**a**) and levels of three species driving NGM community states (**b**) (week 1, based on the earlier sample of day 4 or 7) in neonates longitudinally sampled from weeks 1 to 3 (day 21, total $N = 306$; VD $N = 140$; CS $N = 166$). The proportion of community states that remained consistent from weeks 1 to 3 is depicted as a percentage of their initial sample size in week 1 (labelled in black). Participants starting with BB or BL on week 1 were significantly more likely to retain their community state in week 3 compared with those with EF (pairwise chi-square tests with FDR correction, $P < 0.001$). **c**–**e**, Persistence of the dominant abundance of driver species of NGM community states in week 1 (**c**,**d**) or week 3 (**e**) in the paired longitudinal samples obtained later at week 3 (**c**) and in infancy (**d**,**e**). **f**, Persistent carriage of week-1 driver species in paired longitudinal samples obtained later in infancy. Species carriage is defined using a threshold of 0.1% relative abundance. Sample sizes of participants longitudinally sampled for weeks 1 and 3 shown in **a**–**c** are: total $N = 306$; VD $N = 26/39/75$ among BB/EF/BL, respectively; CS $N = 26/114/26$ among BB/EF/BL, respectively; for week 1 and infancy (also referred to as the 'infancy persistence' group) shown in **d** and **f**: total $N = 302$; VD $N = 27/43/90$ among BB/EF/BL, respectively; CS $N = 17/108/17$ among BB/EF/BL, respectively; and for week 3 and infancy shown in **e**: total $N = 146$; VD $N = 12/11/43$ among BB/EF/BL, respectively; CS $N = 17/37/26$ among BB/EF/BL, respectively. Colour represents NGM community states or driver species: BB and *B. breve* in green; EF and *E. faecalis* in purple; BL and *B. longum* in orange. Boxplots as in Fig. 1. Statistical differences in abundance between time points (**a**), species (**c**–**e**) and carriage frequency (**f**) were determined using paired *t*-tests, Wilcoxon tests and chi-square tests (all two-sided) with FDR correction, respectively.

Given that opportunistic pathogens including *E. faecalis*, *E. faecium*, *Klebsiella oxytoca*, *K. pneumoniae*, *Enterobacter cloacae* and *Clostridium perfringens*, which are enriched in the EF community state, lack the capability to metabolize HMOs and their by-products, we hypothesize that *B. breve*'s versatility in utilizing these predominant neonatal dietary components substantially enhances its fitness against opportunistic pathogens in vivo. Considering that all neonates in the study would have been exposed to the same level of HMOs through a



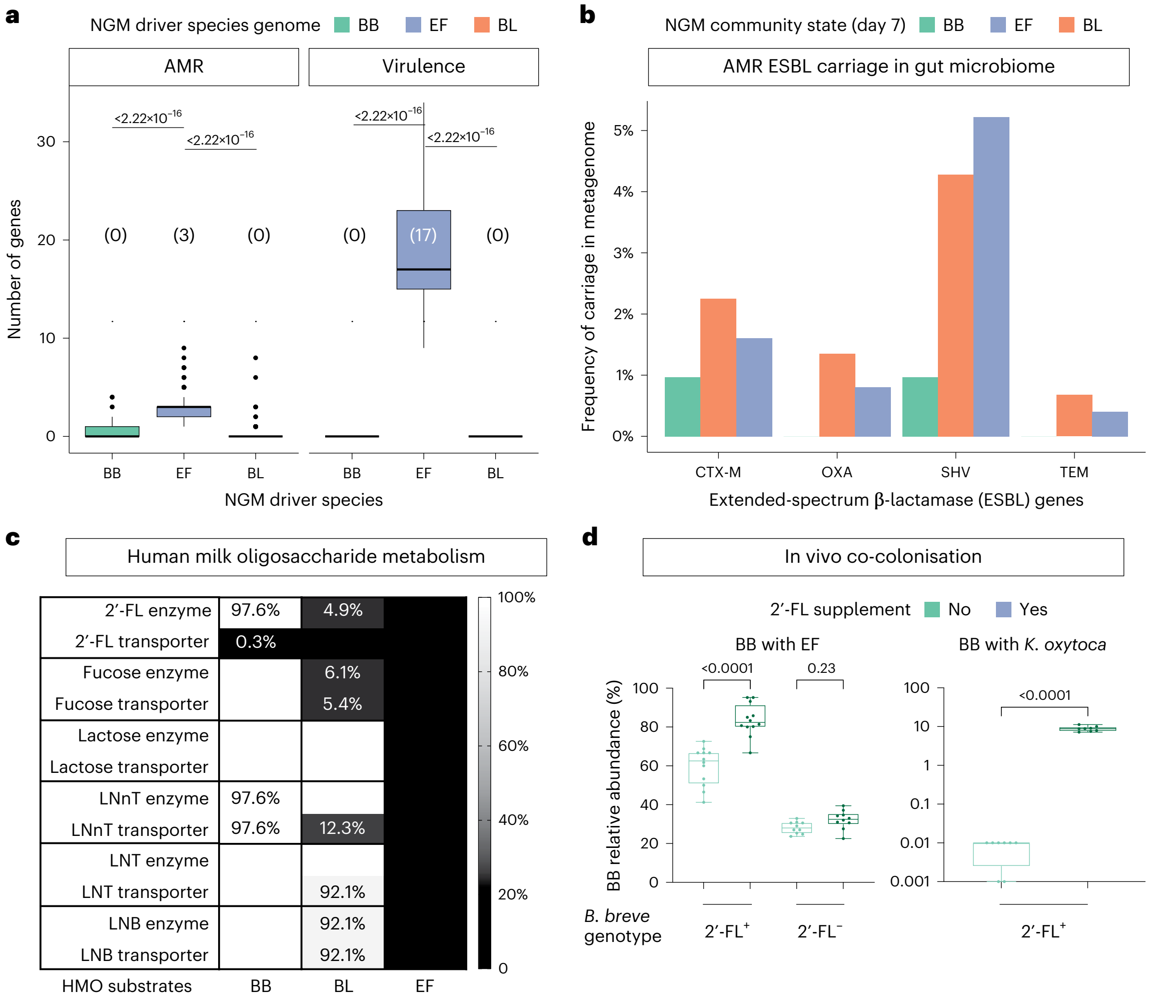


Fig. 4 | *Bifidobacterium* species drive resistance to antimicrobial resistance and pathogen colonization. a, Counts of detected AMR and virulence genes in driver species genomes, with median values enclosed in brackets. Wilcoxon test (two-sided) with FDR correction; number of genomes (isolates in brackets): BB $N = 297$ (30), EF $N = 561$ (54) and BL $N = 391$ (49). **b**, Carriage of high-risk AMR genes associated with ESBL in the day-7 NGM community state samples based on raw metagenomic assemblies (BB $N = 207$, EF $N = 498$, BL $N = 444$). The *x* axis shows the most clinically prevalent ESBL genes belonging to CTX-M, OXA, SHV and TEM families. **c**, Proportion of species genomes, indicated by a colour gradient, predicted to utilize HMOs or their primary downstream products, lactose and fucose. The actual proportions are labelled for genotypes that are not completely present. The predictions are based on the presence of both the gene and its encoded transporters required for utilization of each substrate. 2′-fucosyllactose (2′-FL) liberates lactose and fucose which are also present in breast milk. Utilizations of LNnT, LNT and LNB will all liberate lactose. **d**, NGM driver species BB confers pathogen colonization resistance in vivo. The boxplot depicts the relative abundance of BB compared to the opportunistic pathogen species EF or *K. oxytoca* (KO). The *x* axis represents three experimental groups co-colonized as follows: (1) BB type strain DSM 20213 (2′-FL$^+$) with EF; (2) BB natural variant D19 (2′-FL$^-$, isolated from a BBS neonate) with EF; and (3) BB type strain (2′-FL$^+$) with KO (D63). The BB genotype (2′-FL$^{+/-}$) indicates whether the strain encodes the α-L-fucosidase (GH95) enzyme encoding for 2′-FL metabolism. In each co-colonization group, one group of mice also received a 2′-FL supplement (50 mg ml^{-1} per day) in their daily drinking water. The *y* axis for BB co-colonization with KO is shown on a log scale. Each experimental condition included 3–5 mice per cage and 3 technical replicate cages. Statistical differences between treatment groups were determined using a *t*-test with Welch's correction (two-sided). Boxplot centre line indicates the median, box limits indicate upper and lower quartiles, and whiskers indicate 1.5× the interquar tile range.

predominantly breast milk-based diet, regardless of their community state, we reason that the metabolic capability to utilize HMOs, including but not limited to 2′-FL, not only contributes to the dominance and stability of the BB community state but also enables *B. breve* to outcompete pathogenic species that cannot utilize HMOs. Supporting our hypothesis, we demonstrate in a gnotobiotic mouse model, co-colonized with *B. breve* and the opportunistic pathogen driver species *E. faecalis*, that *B. breve* dominates, and this dominance is amplified by dietary 2′-FL supplementation (Fig. 4d). The 2′-FL-mediated pathogen resistance in vivo phenotype of *B. breve* also extends to the Gram-negative enteropathogen *K. oxytoca*, albeit to a lesser extent. Importantly, the anti-pathogen effect was absent in mice colonized with a natural *B. breve* variant isolated from a BBS neonate lacking the α-L-fucosidase (GH95) enzyme necessary for 2′-FL metabolism. These findings suggest that *B. breve*'s strain-specific and gene-dependent utilization of HMOs could have a crucial role in enhancing resistance to pathogen colonization by inhibiting pathogen growth.

Discussion

In presumably the largest neonatal gut metagenome study ever undertaken, we discovered three distinct NGM community states in over 1,000 healthy, full-term neonates drawn from the general UK population, representing diverse ethnicities and sociodemographic backgrounds. Factors that may influence the maternal gut microbiota, such as maternal age, ethnicity and parity, as well as events that influence its vertical transmission to the neonatal gut during the perinatal period (for example, CS and maternal antibiotics), serve as independent determinants of the acquisition of primary colonizers. The presence of a highly unstable community state (EF) with AMR-enriched opportunistic pathogens underscores the hospital environments and practices, such as maternal antibiotics during labour and elective CS births, as important risk factors[1,35–38]. Although antibiotics after birth and breastfeeding are known important factors shaping the later infant-stage microbiome development[13,15,39,40], these postnatal factors had no observable effect on very early NGM dynamics on either the acquisition or the switching of the NGM community states. Together, our findings highlight that the NGM assembly outcome is highly dependent on the succession of primary colonizer species, with prenatal and perinatal factors associated with birth exerting profound influences.

Although the early-life microbiota is thought to be highly dynamic as reflected by high inter-individual variation[1], here we describe an undisturbed, native primary succession pattern in microbiota assembly driven by a single *Bifidobacterium* species. *B. longum* is strongly linked to factors that promote maternal gut microbiota transmission at birth, such as vaginal delivery and absence of antibiotics. While *B. breve* seems unaffected by these factors, its independent association with maternal ethnicity (Asian) could be linked to the mother's FUT2 secretor status, which determines the presence of 2′-FL and other HMOs in breast milk and is reportedly more common in Asian participants than in white participants[41]. The pattern of exclusive dominance by either *B. breve* or *B. longum* during very early life could also be observed in other cohorts across geographically diverse populations[11–16]. Earlier neonatal cohorts, limited by their smaller sample sizes ($N < 100$ compared with $N > 1{,}000$ in this study) and lack of longitudinal samplings, were unable to report such patterns as distinctly and conclusively as we have in this study. Given that de novo identification of optimal community state clusters is sample size dependent[10], our expanded BBS dataset—nearly 10 times larger than the previously largest neonatal dataset[13]—provided us with the statistical power to report a distinct tripartite NGM community structure. This includes a previously undescribed at-risk community state (EF) harbouring AMR-carrying opportunistic pathogens, and presumably for the first time, the epidemiological and longitudinal dynamics signatures of each NGM community state. Our findings provide crucial evidence that can guide the rational selection of species and strains for infant interventional trials, as well as the development of next-generation microbiota-based therapeutics. Future studies can stratify infants by their earliest gut community states to examine potential associations with longer-term health outcomes.

Both *Bifidobacterium* community states can drive deterministic and stable assembly trajectories in vivo through optimized utilization of HMOs exclusively present in human breast milk, the predominant diet during the neonatal period. Our human and in vivo data are in agreement with recent observations based on in vitro experiments[42,43], showing that *B. breve* is functionally better adapted to an HMO-rich diet in very early life and dominate NGM through priority effects. Here we further demonstrated, in human and mouse, the functional impact of *B. breve* priority effects, resulting in stronger colonization resistance against AMR-enriched pathogens, including *E. faecalis* and *K. oxytoca*.

While the exact origins of opportunistic pathogens such as *E. faecalis* contributing to EF remain to be confirmed, their strong association with disruptions of natural birth (for example, CS and antibiotics) and their ubiquitous presence in the hospital birth environment[22,23] strongly indicate the hospital operating room as the most likely source, with exposure further exacerbated by the lack of maternal microbiota transmission that frequently occurs during natural birth. Although the EF perturbation patterns appear to be largely transient, with the neonatal microbiota naturally recovering from a delayed colonization trajectory[1,44], inadequate pathogen clearance could persist into infancy. Along with the short-term exposure to high AMR and virulence, early acquisition of pathogens represents increased risk for infection susceptibility due to the immature immune system in very early life[45]. Also, the delayed or lack of exposure to commensal *B. breve* or/and *B. longum* as a primary colonizer in the critical neonatal window of immunity[45] and neurological[46] development could potentially result in neurodevelopment and immune-mediated disorders later in childhood[47]. Epidemiological evidence from other independent birth cohorts indicates that a non-*Bifidobacterium* (for example, EF) community state may predispose neonates to an increased risk of neurological disorders[48] and respiratory diseases (for example, asthma and atopy[49,50]), including respiratory infections[26,51], later in childhood.

Bifidobacterium spp. are known to achieve bifidogenic effects through the provision of HMOs, with a notable focus on *B. infantis* and its probiotic application as a specialized HMO-utilizing species. Despite its prevalence and dominance in infants from low- to middle-income and non-industrialized settings[17,52], *B. infantis* is notably absent in this UK cohort and other Western cohorts, suggesting that it may no longer be naturally colonizing newborns in Western, industrialized populations. Its notable absence indicates a potential lack of a reservoir for *B. infantis* to establish itself as a primary colonizer, despite the considerable selective advantage that extensive exposure to HMOs during the neonatal period would presumably provide. Our results demonstrate that an HMO functional niche could be filled by other species (*B. breve* or *B. longum*) capable of metabolizing HMO if they are prevalent in the perinatal microbial species pool. The findings of strain-dependent utilization of HMOs, including but not limited to 2′-FL, and colonization resistance phenotypes of *B. breve* further highlight that the success of primary succession is probably dependent on both the species prevalence and strain-level functional variation.

Maternal seeding of microbial metabolizers of the specialized bioactives in breast milk probably represents an evolutionally conserved strategy to prime human gut microbiota assembly with primary colonizers with the highest likelihood for priority effects, such as *B. breve* and, to a lesser extent, *B. longum*. While both species have been associated with maternal origins[53], strain transmission analyses from both our work as well as that of others[19] have identified only *B. longum* as the most frequently transmitted species from the mother's gut. Although *B. breve* did not appear to originate from the maternal gut microbiota, we cannot rule out the possibility of vertical transmission of very low-abundance *B. breve* strains. Recent cultivation-based evidence has confirmed that such transmission can occur below the limits of metagenomic strain detection[54]. Other unsampled maternal or environmental sources could also be involved in seeding *B. breve*. One likely source is breast milk microbiota, where *B. breve* has been detected and implicated in the entero-mammary pathway—a retrograde mechanism for milk inoculation[21]. Future research should investigate the global strain reservoir and transmission patterns of *Bifidobacterium* species, especially for the poorly understood *B. breve*. Considering the limited success of probiotic-derived *B. infantis* strains in natural engraftment of neonatal gut microbiota in both industrialized and non-industrialized populations[18,55], comprehensive strain-level functional characterizations of naturally prevalent and stable primary colonizers, such as *B. breve*, are vital. This effort will expedite the discovery of infant probiotics that are better optimized for local populations.

Methods

Study population

The Baby Biome Study (BBS) participants were recruited at the Barking, Havering and Redbridge University Hospitals NHS Trust, the University Hospitals Leicester NHS Trust and the University College London Hospitals NHS Foundation Trust from May 2014 to December 2017. The study was approved by the NHS London City and East Research Ethics Committee (REC reference 12/LO/1492). Mothers provided written informed consent for their participation and the participation of their children in the study. The study was performed in compliance with all relevant ethics regulations.

Whole-genome sequencing and analysis

The study participants, drawn from a general population of women giving birth in hospitals in the UK without any clinical inclusion or exclusion criteria as per the BBS study protocol[56], are predominantly healthy, full-term neonates. The study dataset comprised 2,387 metagenomes, with 1,679 from the previously published[1] BBS phase 1 (BBS1) and 708 new neonatal gut metagenomes in BBS phase 2 (BBS2), totalling 1,288 participants. The aim of BBS2 was to sequence all the remaining neonatal samples collected from the original BBS study. The study sample size was predicated on detecting differences by mode of birth rather than providing statistical power to discern differences in microbial community states. The sampling and data processing protocols, ranging from sample collection to sequence data generation, quality control (low-quality trimming and human decontamination) and processing, remained unchanged from those previously described[1] for BBS1. In brief, faecal samples were collected at home by parents from neonates in the first 3 weeks of life (primarily on days 4, 7 and 21) and later in infancy. Paired maternal faecal samples were taken at the hospital around the time of birth. Most new samples in BBS2 were collected on day 7 of life. The only change was an institute-wide upgrade in the Illumina sequencing platform, transitioning from HiSeq 2500-v4 (2 ×125 bp) to HiSeq 4000 (2 ×151 bp). A multiplexing strategy was employed to ensure that the target depth remained consistent with BBS1. While the upgraded sequencing platform has resulted in a marginal increase in sequencing depth for BBS2 (from 19.3 to 20.4 million reads per sample post-quality control, calculated with seqkit (v.2.4.0)[57], $P < 0.001$, two-sided t-test), it did not impact either the community state assignment ($P = 0.4731$, likelihood ratio test via multinomial logistic regression) or the recovery of high-quality genomes (proportion of the total genome bins) for NGM driver species ($P = 0.9716$, Mantel–Haenszel chi-squared test, stratified by species).

Read-based taxonomic classification was performed against the Genome Taxonomy Database (GTDB, RS207) representative bacterial and archaeal species genomes ($N = 65,703$) using bowtie2 (v.2.3.5)[58] and inStrain (v.1.3.0)[59] 'profile' with the recommended '–database mode' and 50% genome breadth (covered by ≥1 read) cut-off, as previously described[52,59]. The R package phyloseq (v.1.12.0)[60] was used for metagenomic data analysis, and results were processed and visualized using tidyverse (v.2.0.0) in RStudio (v.4.1.0).

Strain sharing analysis was performed using StrainPhlAn4 (ref. 61), following the workflow and species-specific strain identity thresholds previously described[19]. Where appropriate, multiple testing corrections were applied to all statistical tests using the Benjamini–Hochberg FDR method with a significance threshold of 5%, unless otherwise specified.

Cultivation and whole genome sequencing of the NGM species isolates were performed using the previously established workflow[1] for BBS1. In brief, the NGM species in driver NGM samples were cultured from corresponding frozen faecal samples using selective media: *Bifidobacterium* selective media (Sigma-Aldrich) for *B. longum* and *B. breve*, and *Enterococcus* selective agar (Sigma-Aldrich) for *E. faecalis*. Purified bacterial isolates were sequenced on the Illumina HiSeq X or NovaSeq 6000 system (2 ×151 bp), and assembled and quality-controlled using shovill (v.1.1.0; https://github.com/tseemann/shovill) and CheckM2 (ref. 62), respectively.

Clinical and sociodemographic metadata sources and management

Participant data were collected using a clinical record form at enrolment by the BBS research midwives or from available clinical records at birth. Hospital maternity electronic records with pregnancy and perinatal clinical information were obtained directly from the hospital trusts, and databases containing the variables of interest were merged. Variables were harmonized where possible across different databases. For discrepancies, data from the BBS clinical record forms were given priority, and hospital electronic data were used to complete missing data. At the time of stool sample collection, mothers completed a short form on feeding mode and antibiotic exposure. A total of 20 variables were included in the final analyses on the basis of clinical relevance, quality of data and completeness ($N = 6$ maternal, $N = 8$ perinatal or at time of delivery, $N = 5$ postnatal, $N = 1$ at the time of stool sample collection variables). Ten variables had no missing or <1% missing data. Four had between <1% and 5% missing data (index of multiple deprivation (IMD), maternal smoking, prolonged rupture of membranes (PROM) and neonatal labour antibiotics after birth), two had between 5% and 15% missing data (maternal ethnicity and feeding mode at the time of stool sample collection), and one had >30% missing data (skin to skin).

We used participant postcode to determine IMD[63], which provides a measure of socioeconomic status that is calculated as an area-level relative deprivation score that we organized into quintiles from 1 (least deprived) to 5 (most deprived). The score considers seven individually weighted domains (income, employment, education, health, crime, barriers to housing and services, and living environment). Prophylactic antibiotics were administered to all mothers undergoing caesarean section in this cohort, as well as to newborns displaying risk factors or clinical indicators of early-onset neonatal infection, in accordance with local trust policies and UK national guidelines at the time[64,65]. To our knowledge, no participants were given antibiotics for treating bloodstream infections of *E. faecalis*. Skin-to-skin contact is defined as contact of mother and baby immediately after birth at least for 1 h or until the next feeding[66]. Feeding mode at the time of stool sample collection was determined through a questionnaire that included three categories: exclusive breastfeeding, exclusive bottle feeding, or both (that is, mixed feeding). For comparisons involving (non)exclusive breastfeeding, the latter two categories were merged into a single 'non-exclusive breastfeeding' category.

Statistical analyses

No statistical methods were used to pre-determine sample sizes, but this study already represents the largest dataset of longitudinal faecal metagenomes ($n = 1,904$; $n = 2,387$ including infancy samples) of newborns ($n = 1,288$). No data were excluded unless they failed quality control steps. Microbiome data collection and analysis were not randomized or performed blind to the conditions of the experiments, as this is an observational study. Biological counting experiments were blinded by another person other than the experimenter before being counted to avoid experimental bias. For mouse experiments, treatments were randomized by cage by researchers blinded to treatment conditions. Unless otherwise stated, non-parametric statistical tests were performed unless tests for normality and equal variances showed that these assumptions were met.

For the epidemiological analyses of NGM community states in the first week of life, BBS participants with sufficient metadata were explored (90.4%, $N = 1,108$ of 1,225 participants with week-1 sampling). The week-1 NGM community state was determined for each eligible participant by using the earliest available sample from week 1, collected either on day 4 ($N = 64$) or day 7 ($N = 1,044$).

To ascertain risk factors for specific NGM community states: BB versus non-BB, EF versus non-EF, and BL versus non-BL, univariate analyses using fixed-effect logistic regression models were initially performed. Subsequent multivariate models were constructed, also using fixed-effect logistic regression, and included only participants with complete datasets while excluding variables with over 15% missing data. Likelihood ratio tests were employed to calculate all P values. A hierarchical framework was applied in building the multivariate models. Variables were organized in a sequential order into either distal (maternal) or more proximal categories (delivery, postnatal care and the first week of life). Variables were considered potential confounders if they occurred simultaneously with or before exposure variables[67]. Within each category, all variables from that category or previous categories were incorporated into the model to account for confounding.

Sensitivity analyses were conducted to identify factors associated with NGM community state, switching between weeks 1 and 3. This included a subset of 'neonatal longitudinal' participants with sufficient metadata (87.6%, $N = 268$ of 306, corresponding to Fig. 2a). Both univariate and subsequent multivariate analyses were conducted using fixed-effect logistic regression in the same manner as described above. Multivariate models were further adjusted for the week-1 community state (that is, EF, BB or BL) to discern whether any associations were driven by the baseline community state. There was no strong evidence of association, other than for the baseline community state itself. These analyses could not extend to independent community states switches due to insufficient sample size. All analyses were conducted using Stata (v.17.0).

Community state assignment

The NGM community state assignment was applied to all neonatal samples ($N = 1,904$) using two popular methods, namely, the original clustering-based PAM method described in ref. 68 and the probabilistic modelling-based Dirichlet multinomial mixtures (DMM) approach described previously[69]. In accordance with the original protocols, PAM clustering was applied to the species-level relative abundance distance measured by the Jensen–Shannon divergence (JSD) using the R packages 'cluster' (v.2.1.4) and 'vegan' (v.2.6.4), and DMM models were fitted on the species-level relative abundance matrix, modelled by the Dirichlet multinomial distribution, using the R package 'DirichletMultinomial' (v.1.4). For both methods, the optimal number of clusters of three was determined on the basis of the Calinski–Harabasz index for PAM clustering and the model fit score based on Laplace approximation for DMM. The community states were named according to the top taxonomic driver (species) that contributed the most to microbial community variation ('envfit' R^2, $P < 0.05$) in PAM and to each Dirichlet component (cluster) in DMM. The strength of association between the PAM and the DMM-based community states was 0.726 (Cramer's V correlation). For downstream analyses, the PAM-based community state assignment was selected because it maximized both the sample size of community states BB and BL (Extended Data Fig. 2e) and the mean relative abundance of the driver species in the respective community state (*B. breve* in BB, *E. faecalis* in EF; Extended Data Fig. 2f).

To validate the single-species dominance in external neonatal cohorts, the same workflow for community state type assignment was independently applied to four public gut metagenomic datasets with a comparable sampling window (<6 months) to the BBS cohort, including partial or exclusive sampling of the neonatal period (0–1 month). The earliest sampling windows were from cohorts derived from diverse geographical populations and lifestyles, including Sweden[13,42] (PRJEB6456, days 4–12, $N = 37$), Israel[14] (PRJNA994433, weeks 1–24, $N = 60$), the USA (TEDDY cohort[15,16], PRJNA400115, months 2–6, $N = 69$) and Bangladesh[17] (PRJNA806984, months 0–2, $N = 234$).

Metagenome assembly and functional analyses

Quality-controlled, raw paired-end reads were first assembled with SPAdes (v.3.13.5)[70] with the option –meta. Unassembled reads were then filtered out by mapping raw reads back to metaSPAdes[71]-assembled contigs using bwa-mem (v.0.7.17)[72], followed by re-assembly with MEGAHIT (v.1.1.3)[73] using default parameters. Subsequently, the metaSPAdes and MEGAHIT assemblies were combined, sorted and short contigs (<1,500 bp) removed. The resulting assemblies were then independently binned with MetaBAT 2 (v.2.13)[74], MaxBin2 (v.2.2.4)[75] and CONCOCT (v.0.4)[76] using default parameters and a minimum contig length threshold of 1,500 bp (option –minContig 1500). The depth of contig coverage required for the binning was inferred by mapping the raw reads back to their assemblies with bwa-mem (v.0.7.17) and then calculating the corresponding read depths of each individual contig with samtools[77] ('samtools view -Sbu' followed by 'samtools sort') together with the 'jgi_summarize_bam_contig_depths' function from MetaBAT 2.

Thereafter, individual genome bin sets produced by three binning programs were consolidated into a refined bin set consisting of the best version of each bin based on the most optimal genome completion and contamination metrics among all seven versions of hybridized bin sets (MetaBAT 2, MaxBin2, CONCOCT, MetaBAT 2 + MaxBin2, MetaBAT 2 + CONCOCT, MaxBin2 + CONCOCT, MetaBAT 2 + MaxBin2 + CONCOCT) as estimated by CheckM (v.1.0.7)[78] using the metaWRAP (v.1.2)[79] 'bin_refinement' pipeline[79]. In total, 22,668 prokaryotic metagenome-assembled genomes (MAGs) met the criteria of having >50% completeness and <5% contamination, as determined by CheckM2 (ref. 62). These MAGs were subsequently taxonomically assigned using the GTDB[80] (R214) taxonomy with GTDB-Tk (v.2.3.0)[81].

For genome analyses of the three NGM driver species, data were derived from samples as either cultivated isolate genomes or metagenome-assembled genomes (MAGs) when cultured strains were unavailable. Only near-complete, high-quality MAGs were used in the functional analyses ($N = 1,116$). All genomes met strict quality control criteria, which included ≥90% completeness, ≤5% contamination, an N50 value of ≥10 kb, passing the GUNC test, an average contig length of ≥5 kb and ≤500 contigs, as previously described[82]. Genome annotation of metabolic function was performed using DRAM (v.1.4.5)[83], which integrates annotations from multiple databases, including Pfam, KEGG (KOfam), UniProt, dbCAN (carbohydrate-active enzymes) and MEROPS (peptidases). The functional gene counts from KEGG and CAZy annotations were used to generate a PCA plot using the R package 'pcaMethods', employing conventional singular value decomposition with imputation. The genome-based prediction of HMO substrate utilization was based on KEGG and CAZy annotations mapped against a list of manually curated relevant genes and pathways as described recently[42,43]. The genes corresponding to HMO substrates (enzymes; transporters) were: 2′FL (GH95 and/or GH29, FL1_Blon0341-0343 and/or FL2_Blon2202-2204), lactose (GH2, LacS), fucose (FumC/D/E/F/G, FucP), LNT (GH42 or GH136, GltABC), LNnT (GH20, Bbr_1554) and LNB (GH112, GltABC). In silico screening of AMR and virulence factor genes was performed at the species level with species MAGs and at the sample level with raw metagenome assemblies as input for ABRicate against the NCBI AMRFinderPlus and VFDB databases as previously described[1]. The AMR genes encoding for the extended-spectrum β-lactamase (ESBL) phenotype were annotated using the curated antibiotic subclass of the NCBI Pathogen Detection Reference Gene Catalog (as of 1 October 2023).

Bacterial strains and reagents

The bacterial strains used in this study were either part of the in-house (HMIL) culture collection cultivated from the BBS faecal samples or requested from public collections (DSMZ). Specific strains were: *B. breve* strains (type strain DSM 20213 and D19 isolated from a BBS neonate), *E. faecalis* (D13 isolated from a BBS neonate) and *K. oxytoca* (D63 isolated from a BBS neonate). Purified HMO 2′FL (GlyCare 2FL 9000, batch 20156002) and LNnT were purchased from Glycom, DSM.

Mouse experiment

Wild-type C57BL/6N mice were maintained under germ-free conditions at the Wellcome Sanger Institute Home Office-approved facility, with all procedures carried out in accordance with the UK Animals (Scientific Procedures) Act of 1986 under Home Office approval (PPL no. 80/2643). Germ-free mice were housed under a 12 h light/12 h dark cycle, ambient temperature and humidity condition in positive-pressure isolators (Bell), with faeces tested by culture, microscopy and PCR to ensure sterility. Consumables were autoclaved at 121 °C for 15 min before introduction into the isolators. For experimentation, 6-week-old mice of both sexes were randomly assigned to treatment groups. Cages were opened in a vaporized hydrogen peroxide-sterilized, class II cabinet (Bioquell), with mono-colonized gnotobiotic lines generated by oral gavage on day 1 (*B. breve*) at the concentration of 10^9 colony-forming units (c.f.u.) per ml and day 4 (challenged by opportunistic pathogen species *E. faecalis* or *K. oxytoca* at the concentration of 10^4 c.f.u. per ml). Materials were prepared in Dulbecco's PBS at 100 mg ml^{-1} immediately before administration under anaerobic conditions (10% H, 10% CO_2, 80% N) in a Whitley DG250 workstation at 37 °C. Mice were maintained in sterile ISOcages (Tecniplast) and housed on ISOrack for the period of the experiment.

Control groups of mice colonized with BB, EF or KO without any treatment were also included to confirm mono-colonization. One of the two groups of the co-colonized mice (for both BB + EF and BB + KO experiments) were exposed to 2′-FL via daily drinking water (50 mg ml^{-1} per day) throughout the experiment. Faecal samples were collected on each oral gavage day and plated to test for contamination. Mice were killed on day 11 (7 days post inoculation on day 4), with faecal samples collected and plated for colony count on yeast extract casitone fatty acids (YCFA) aerobically (to select for *E. faecalis* or *K. oxytoca*) and YCFA with mupirocin (to select for *Bifidobacterium* spp.) media under anaerobic conditions. YCFA is a complex, broad-range medium[84]. Each experimental condition included 3–5 mice per cage and 3 technical replicate cages.

DNA was extracted from faeces using FastDNA Spin Kit for Soil (MPBio) according to manufacturer instructions, and DNA eluted into 100 µl of double-distilled H_2O. Eluted DNA was then diluted 1:50 and qPCR performed using SYBR Green chemistry (Thermo Fisher). The absolute bacterial load in each faecal sample was determined by qPCR using a calibration curve generated with genomic DNA and taxon-specific primer sequences (*E. faecalis*, F: 5′-CCCTTATTGTTAGTTGCCATCATT-3′, R: 5′-ACTCGTTGTACTTCCCATTGT-3′; *Bifidobacterium* spp., F: 5′-CTCCTGGAAACGGGTGG-3′, R: 5′-GGTGTTCTTCCCGATATCTACA-3′; *K. oxytoca*, F: 5′-GGACTACGCCGTCTATCGTCAAG-3′, R: 5′-TAGCCTTTATCAAGCGGATACTGG-3′). As previously described[85], the relative abundance of each target species was estimated by normalizing to those of a universal bacterial 16S primer (F: 5′-GTGSTGCAYGGYTGTCGTCA-3′, R: 5′-ACGTCRTCCMCACCTTCCTC-3′).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Shotgun metagenomic sequencing data (after quality trimming and human decontamination) of the entire Baby Biome Study cohort have been deposited to the European Nucleotide Archive under study accession number ERP115334. Bacterial genome assemblies for the three species analysed have been deposited in Zenodo at https://doi.org/10.5281/zenodo.12667210 (ref. 86). Sample metadata and participant-level clinical metadata of de-identified study participants are provided in the Supplementary Tables. The raw faecal samples and bacterial isolates are available from the corresponding authors upon request.

Code availability

All software used to perform these analyses is publicly available. Software tools used are listed in the main text and Methods.

References

1. Shao, Y. et al. Stunted microbiota and opportunistic pathogen colonization in caesarean-section birth. *Nature* **574**, 117–121 (2019).
2. Mitchell, C. M. et al. Delivery mode affects stability of early infant gut microbiota. *Cell Rep. Med.* **1**, 100156 (2020).
3. Bogaert, D. et al. Mother-to-infant microbiota transmission and infant microbiota development across multiple body sites. *Cell Host Microbe* **31**, 447–460 (2023).
4. 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 (2018).
5. 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 (2018).
6. Fehr, K. et al. Breastmilk feeding practices are associated with the co-occurrence of bacteria in mothers' milk and the infant gut: the CHILD cohort study. *Cell Host Microbe* **28**, 285–297 (2020).
7. Sprockett, D., Fukami, T. & Relman, D. A. Role of priority effects in the early-life assembly of the gut microbiota. *Nat. Rev. Gastroenterol. Hepatol.* **15**, 197–205 (2018).
8. Debray, R. et al. Priority effects in microbiome assembly. *Nat. Rev. Microbiol.* **20**, 109–121 (2022).
9. Mäklin, T. et al. Strong pathogen competition in neonatal gut colonisation. *Nat. Commun.* **13**, 7417 (2022).
10. Costea, P. I. et al. Enterotypes in the landscape of gut microbial community composition. *Nat. Microbiol.* **3**, 8–16 (2018).
11. Avershina, E. et al. Bifidobacterial succession and correlation networks in a large unselected cohort of mothers and their children. *Appl. Environ. Microbiol.* **79**, 497–507 (2013).
12. Laursen, M. F. & Roager, H. M. Human milk oligosaccharides modify the strength of priority effects in the *Bifidobacterium* community assembly during infancy. *ISME J.* 17, 2452–2457 (2023).
13. 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).
14. Ennis, D., Shmorak, S., Jantscher-Krenn, E. & Yassour, M. Longitudinal quantification of *Bifidobacterium longum* subsp. *infantis* reveals late colonization in the infant gut independent of maternal milk HMO composition. *Nat. Commun.* **15**, 894 (2024).
15. Stewart, C. J. et al. Temporal development of the gut microbiome in early childhood from the TEDDY study. *Nature* **562**, 583–588 (2018).
16. Vatanen, T. et al. The human gut microbiome in early-onset type 1 diabetes from the TEDDY study. *Nature* **562**, 589–594 (2018).
17. Vatanen, T. et al. A distinct clade of *Bifidobacterium longum* in the gut of Bangladeshi children thrives during weaning. *Cell* **185**, 4280–4297.e12 (2022).
18. Casaburi, G. et al. Metagenomic insights of the infant microbiome community structure and function across multiple sites in the United States. *Sci. Rep.* **11**, 1472 (2021).
19. Valles-Colomer, M. et al. The person-to-person transmission landscape of the gut and oral microbiomes. *Nature* **614**, 125–135 (2023).
20. Martín, R. et al. Isolation of bifidobacteria from breast milk and assessment of the bifidobacterial population by PCR-denaturing gradient gel electrophoresis and quantitative real-time PCR. *Appl. Environ. Microbiol.* **75**, 965–969 (2009).
21. Kordy, K. et al. Contributions to human breast milk microbiome and enteromammary transfer of *Bifidobacterium breve*. *PLoS ONE* **15**, e0219633 (2020).
22. Brooks, B. et al. Microbes in the neonatal intensive care unit resemble those found in the gut of premature infants. *Microbiome* **2**, 1 (2014).
23. Brooks, B. et al. Strain-resolved analysis of hospital rooms and infants reveals overlap between the human and room microbiome. *Nat. Commun.* **8**, 1814 (2017).
24. Song, S. J. et al. Naturalization of the microbiota developmental trajectory of Cesarean-born neonates after vaginal seeding. *Med* **2**, 951–964.e5 (2021).
25. Dos Santos, S. J. et al. Maternal vaginal microbiome composition does not affect development of the infant gut microbiome in early life. *Front. Cell. Infect. Microbiol.* **13**, 303 (2023).
26. Reyman, M. et al. Impact of delivery mode-associated gut microbiota dynamics on health in the first year of life. *Nat. Commun.* **10**, 4997 (2019).

27. Lewis, Z. T. et al. Maternal fucosyltransferase 2 status affects the gut bifidobacterial communities of breastfed infants. *Microbiome* **3**, 13 (2015).
28. Martin, R. et al. Early-life events, including mode of delivery and type of feeding, siblings and gender, shape the developing gut microbiota. *PLoS ONE* **11**, e0158498 (2016).
29. Schlievert, P. M., Kilgore, S. H., Seo, K. S. & Leung, D. Y. Glycerol monolaurate contributes to the antimicrobial and anti-inflammatory activity of human milk. *Sci. Rep.* **9**, 14550 (2019).
30. Sweeney, E. et al. The effect of breastmilk and saliva combinations on the in vitro growth of oral pathogenic and commensal microorganisms. *Sci. Rep.* **8**, 15112 (2018).
31. Coburn, P. S. & Gilmore, M. S. The *Enterococcus faecalis* cytolysin: a novel toxin active against eukaryotic and prokaryotic cells. *Cell. Microbiol.* **5**, 661–669 (2003).
32. Bunesova, V., Lacroix, C. & Schwab, C. Fucosyllactose and L-fucose utilization of infant *Bifidobacterium longum* and *Bifidobacterium kashiwanohense*. *BMC Microbiol.* **16**, 248 (2016).
33. Ruiz-Moyano, S. et al. Variation in consumption of human milk oligosaccharides by infant gut-associated strains of *Bifidobacterium breve*. *Appl. Environ. Microbiol.* **79**, 6040–6049 (2013).
34. Sakanaka, M. et al. Varied pathways of infant gut-associated *Bifidobacterium* to assimilate human milk oligosaccharides: prevalence of the gene set and its correlation with bifidobacteria-rich microbiota formation. *Nutrients* **12**, 71 (2019).
35. Azad, M. B. et al. Impact of maternal intrapartum antibiotics, method of birth and breastfeeding on gut microbiota during the first year of life: a prospective cohort study. *BJOG* **123**, 983–993 (2016).
36. Tapiainen, T. et al. Impact of intrapartum and postnatal antibiotics on the gut microbiome and emergence of antimicrobial resistance in infants. *Sci. Rep.* **9**, 10635 (2019).
37. Nogacka, A. et al. Impact of intrapartum antimicrobial prophylaxis upon the intestinal microbiota and the prevalence of antibiotic resistance genes in vaginally delivered full-term neonates. *Microbiome* **5**, 93 (2017).
38. Li, W. et al. Vertical transmission of gut microbiome and antimicrobial resistance genes in infants exposed to antibiotics at birth. *J. Infect. Dis.* **224**, 1236–1246 (2021).
39. Bokulich, N. A. et al. Antibiotics, birth mode, and diet shape microbiome maturation during early life. *Sci. Transl. Med.* **8**, 343ra82 (2016).
40. Yassour, M. et al. Natural history of the infant gut microbiome and impact of antibiotic treatment on bacterial strain diversity and stability. *Sci. Transl. Med.* **8**, 343ra81 (2016).
41. Azad, M. B. et al. Human milk oligosaccharide concentrations are associated with multiple fixed and modifiable maternal characteristics, environmental factors, and feeding practices. *J. Nutr.* **148**, 1733–1742 (2018).
42. Ojima, M. N. et al. Priority effects shape the structure of infant-type *Bifidobacterium* communities on human milk oligosaccharides. *ISME J.* **16**, 2265–2279 (2022).
43. Lou, Y. C. et al. Infant microbiome cultivation and metagenomic analysis reveal *Bifidobacterium* 2′-fucosyllactose utilization can be facilitated by coexisting species. *Nat. Commun.* **14**, 7417 (2023).
44. Podlesny, D. & Fricke, W. F. Strain inheritance and neonatal gut microbiota development: a meta-analysis. *Int. J. Med. Microbiol.* **311**, 151483 (2021).
45. Olin, A. et al. Stereotypic immune system development in newborn children. *Cell* **174**, 1277–1292.e14 (2018).
46. Bethlehem, Ra. I. et al. Brain charts for the human lifespan. *Nature* **604**, 525–533 (2022).
47. Torow, N. & Hornef, M. W. The neonatal window of opportunity: setting the stage for life-long host–microbial interaction and immune homeostasis. *J. Immunol.* **198**, 557–563 (2017).
48. Beghetti, I. et al. Early-life gut microbiota and neurodevelopment in preterm infants: any role for *Bifidobacterium*? *Eur. J. Pediatr.* **181**, 1773–1777 (2022).
49. Depner, M. et al. Maturation of the gut microbiome during the first year of life contributes to the protective farm effect on childhood asthma. *Nat. Med.* **26**, 1766–1775 (2020).
50. Fujimura, K. E. et al. Neonatal gut microbiota associates with childhood multisensitized atopy and T cell differentiation. *Nat. Med.* **22**, 1187–1191 (2016).
51. Alcazar, C. G.-M. et al. The association between early-life gut microbiota and childhood respiratory diseases: a systematic review. *Lancet Microbe* **3**, e867–e880 (2022).
52. Olm, M. R. et al. Robust variation in infant gut microbiome assembly across a spectrum of lifestyles. *Science* **376**, 1220–1223 (2022).
53. Browne, H. P., Shao, Y. & Lawley, T. D. Mother–infant transmission of human microbiota. *Curr. Opin. Microbiol.* **69**, 102173 (2022).
54. Feehily, C. et al. Detailed mapping of *Bifidobacterium* strain transmission from mother to infant via a dual culture-based and metagenomic approach. *Nat. Commun.* **14**, 3015 (2023).
55. Barratt, M. J. et al. *Bifidobacterium infantis* treatment promotes weight gain in Bangladeshi infants with severe acute malnutrition. *Sci. Transl. Med.* **14**, eabk1107 (2022).
56. 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).
57. Shen, W., Sipos, B. & Zhao, L. SeqKit2: a Swiss army knife for sequence and alignment processing. *iMeta* **3**, e191 (2024).
58. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
59. Olm, M. R. et al. inStrain profiles population microdiversity from metagenomic data and sensitively detects shared microbial strains. *Nat. Biotechnol.* **39**, 727–736 (2021).
60. McMurdie, P. J. & Holmes, S. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS ONE* **8**, e61217 (2013).
61. Blanco-Míguez, A. et al. Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. *Nat. Biotechnol.* **41**, 1633–1644 (2023).
62. Chklovski, A., Parks, D. H., Woodcroft, B. J. & Tyson, G. W. CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. *Nat. Methods* **20**, 1203–1212 (2023).
63. Ministry of Housing, Communities and Local Government. *English indices of deprivation 2019* (GOV.UK, 2019).
64. *Caesarean Birth* NICE guideline [NG192] (NICE, 30 January 2024); https://www.nice.org.uk/guidance/ng192
65. *Neonatal Infection: Antibiotics for Prevention and Treatment* NICE guideline [NG195] (NICE, 19 March 2024); https://www.nice.org.uk/guidance/ng195
66. Widström, A., Brimdyr, K., Svensson, K., Cadwell, K. & Nissen, E. Skin-to-skin contact the first hour after birth, underlying implications and clinical practice. *Acta Paediatr.* **108**, 1192–1204 (2019).
67. Victora, C. G., Huttly, S. R., Fuchs, S. C. & Olinto, M. T. The role of conceptual frameworks in epidemiological analysis: a hierarchical approach. *Int. J. Epidemiol.* **26**, 224–227 (1997).
68. Arumugam, M. et al. Enterotypes of the human gut microbiome. *Nature* **473**, 174–180 (2011).
69. Holmes, I., Harris, K. & Quince, C. Dirichlet multinomial mixtures: generative models for microbial metagenomics. *PLoS ONE* **7**, e30126 (2012).

70. Bankevich, A. et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).
71. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. metaSPAdes: a new versatile metagenomic assembler. *Genome Res.* **27**, 824–834 (2017).
72. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
73. Li, D., Liu, C.-M., Luo, R., Sadakane, K. & Lam, T.-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674–1676 (2015).
74. Kang, D. D. et al. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* **7**, e7359 (2019).
75. Wu, Y.-W., Tang, Y.-H., Tringe, S. G., Simmons, B. A. & Singer, S. W. MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. *Microbiome* **2**, 26 (2014).
76. Alneberg, J. et al. Binning metagenomic contigs by coverage and composition. *Nat. Methods* **11**, 1144–1146 (2014).
77. Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
78. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).
79. Uritskiy, G. V., DiRuggiero, J. & Taylor, J. MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis. *Microbiome* **6**, 158 (2018).
80. Parks, D. H. et al. A complete domain-to-species taxonomy for Bacteria and Archaea. *Nat. Biotechnol.* **38**, 1079–1086 (2020).
81. Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk v2: memory friendly classification with the genome taxonomy database. *Bioinformatics* **38**, 5315–5316 (2022).
82. Nayfach, S., Shi, Z. J., Seshadri, R., Pollard, K. S. & Kyrpides, N. C. New insights from uncultivated genomes of the global human gut microbiome. *Nature* **568**, 505–510 (2019).
83. Shaffer, M. et al. DRAM for distilling microbial metabolism to automate the curation of microbiome function. *Nucleic Acids Res.* **48**, 8883–8900 (2020).
84. Duncan, S. H., Hold, G. L., Harmsen, H. J., Stewart, C. S. & Flint, H. J. Growth requirements and fermentation products of *Fusobacterium prausnitzii*, and a proposal to reclassify it as *Faecalibacterium prausnitzii* gen. nov., comb. nov. *Int. J. Syst. Evol. Microbiol.* **52**, 2141–2146 (2002).
85. Forster, S. C. et al. Identification of gut microbial species linked with disease variability in a widely used mouse model of colitis. *Nat. Microbiol.* **7**, 590–599 (2022).
86. Shao, Y. Bacterial genomes of the Baby Biome Study. *Zenodo* https://doi.org/10.5281/zenodo.12667210 (2024).

Acknowledgements

This work was funded by the Wellcome Trust and the Wellcome Sanger Institute (WT101169MA, 206194 and 220540/Z/20/A). For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission. We thank the participating families for their time and contribution to Baby Biome Study; and the research midwives at recruiting hospitals for recruitment and clinical metadata collection.

Author contributions

Y.S. and T.D.L. conceived and designed the study. Y.S. coordinated the experiments and performed computational analyses with assistance from A.M. Y.S. and M.D.S. cultured bacteria strains and performed DNA extraction. S.C. performed germ-free mouse experiments with assistance from N.J.R.D., A.A., K.H., J.L. and H.P.B. A.R., P.B., N.F. and T.D.L. conceived and designed the Baby Biome Study and obtained funding. N.F., A.R. and P.B. managed participant recruitment and sample collection, and coordinated the clinical metadata collection. C.G.-M. curated the clinical metadata and undertook the clinical epidemiological analyses with N.F. Y.S. and T.D.L. wrote the manuscript with inputs from H.P.B., A.M., C.G.-M., A.R., P.B. and N.F.

Competing interests

T.D.L. is the co-founder and CSO of Microbiotica. The other authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41564-024-01804-9.

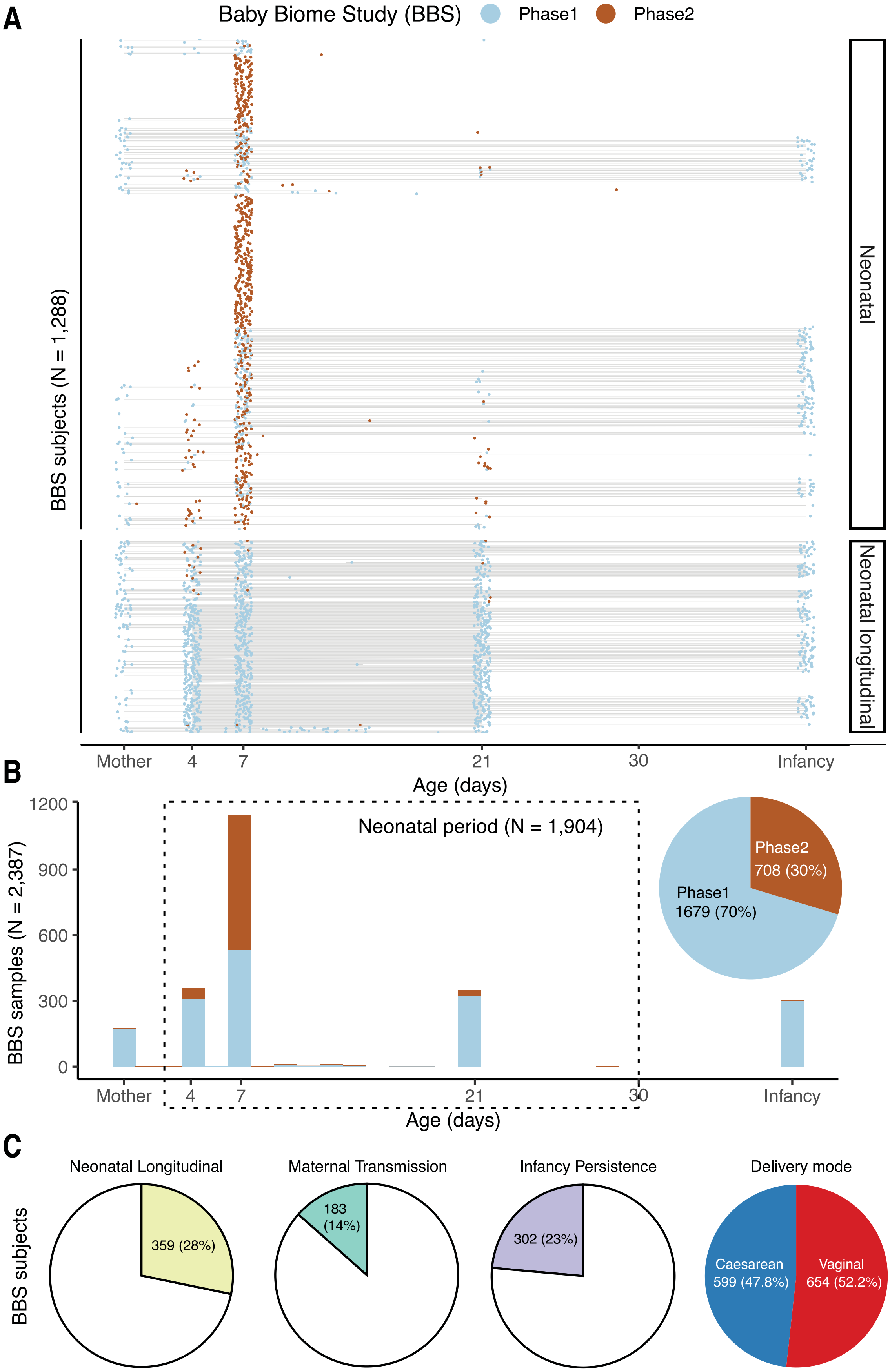
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41564-024-01804-9.

Correspondence and requests for materials should be addressed to Yan Shao or Trevor D. Lawley.

Peer review information *Nature Microbiology* thanks the anonymous reviewer(s) for their contribution to the peer review of this work.

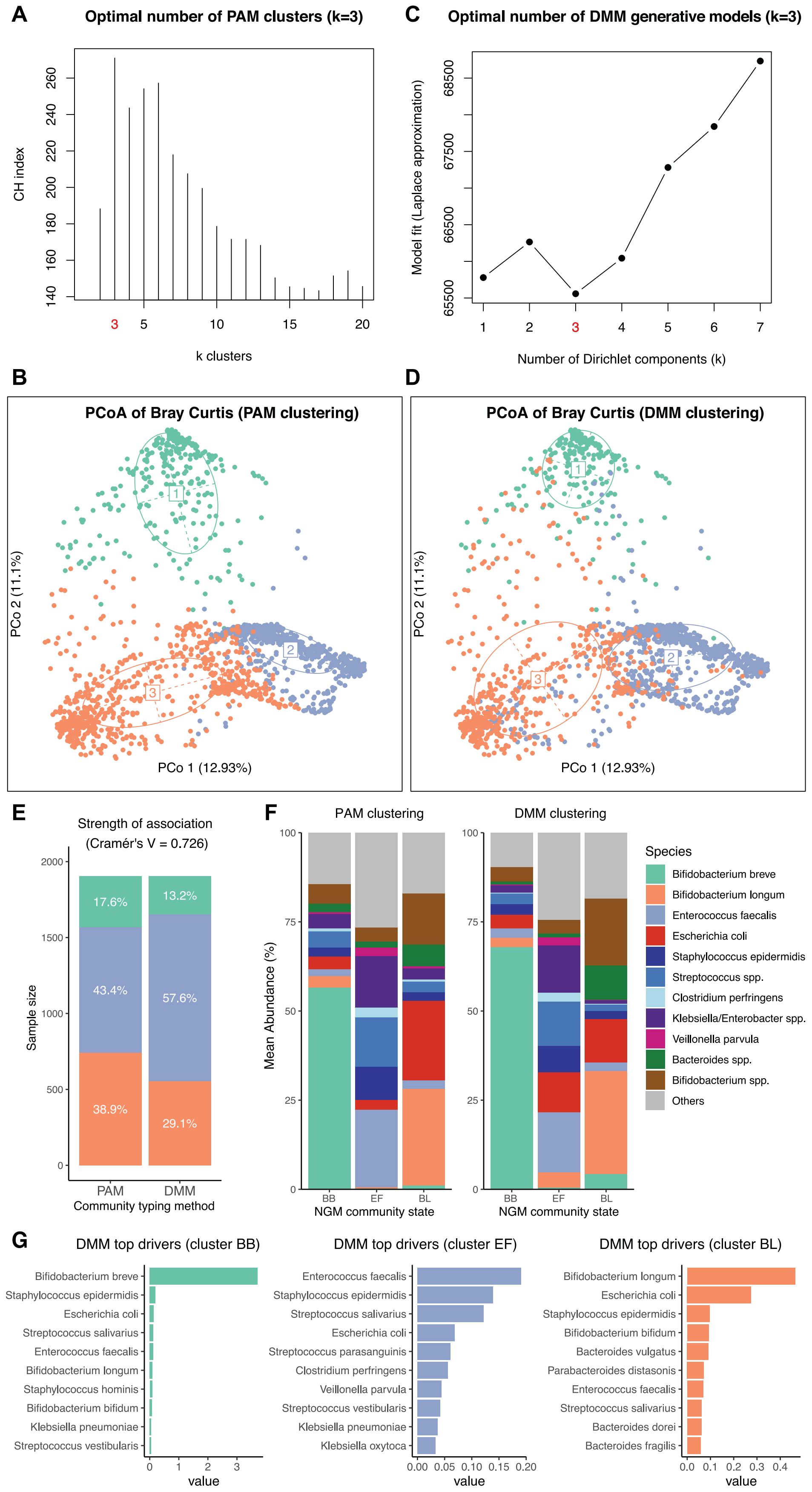
Reprints and permissions information is available at www.nature.com/reprints.

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



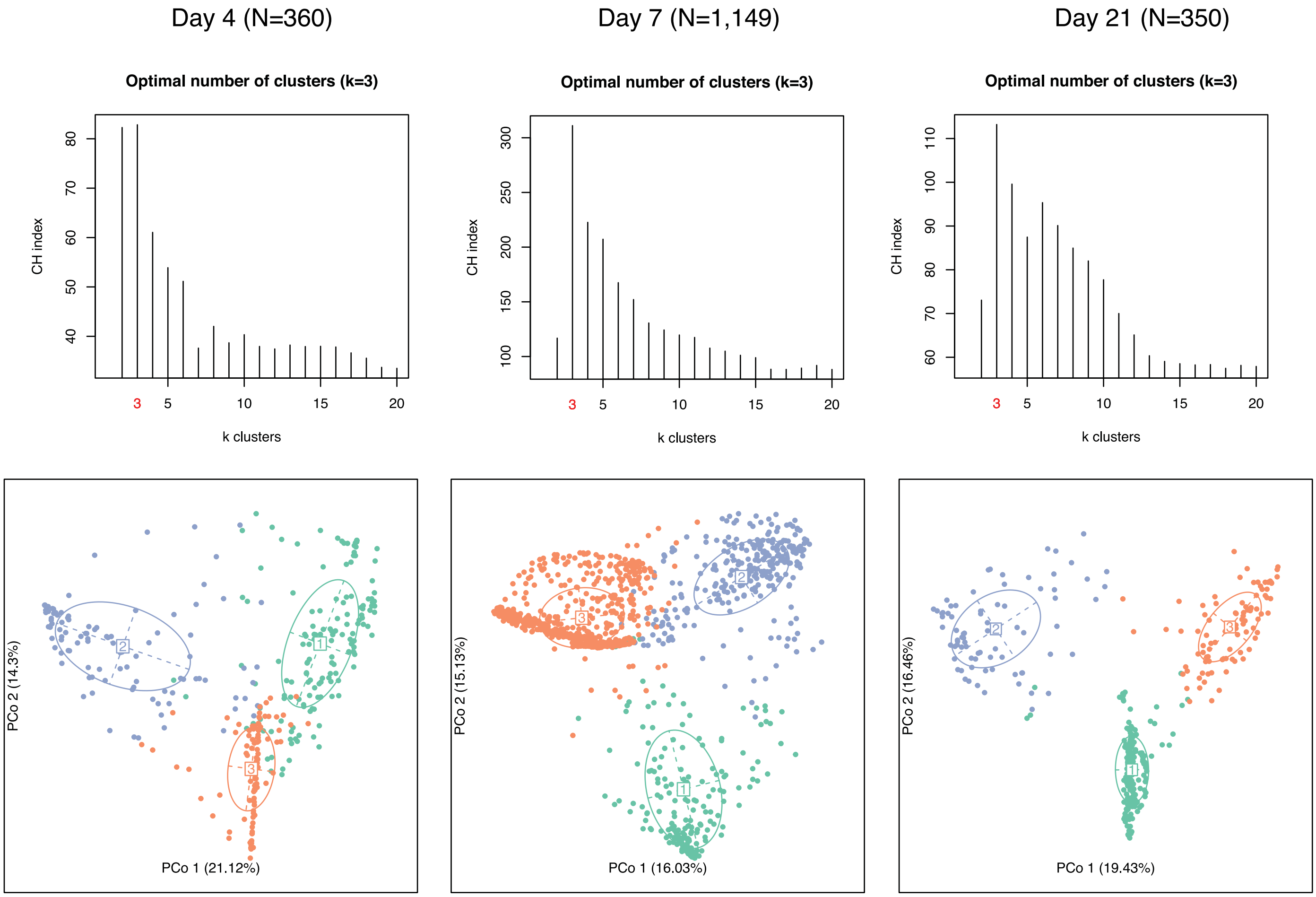
Extended Data Fig. 1 | See next page for caption.

Extended Data Fig. 1 | Overview of sampling in the Baby Biome Study. **(a-b)** Shotgun metagenomes of 2,387 faecal samples from 1,288 neonatal subjects across Phase 1 (Shao et al.[1]) and Phase 2 (this paper). The majority of samples (80% or 1,904) are from the neonatal period **(b)**, primarily taken on day 4 (N=360), day 7 (N=1,149), and day 21 (N=350). **(a)** Rows represent subjects with paired maternal samples (for 'maternal transmission' analysis), longitudinal samples taken during the neonatal period (for 'neonatal longitudinal' analysis), and samples from the infancy period (for 'infancy persistence analysis'). These relationships are indicated by lines linking the samples, with summarised proportions in **(c)**.

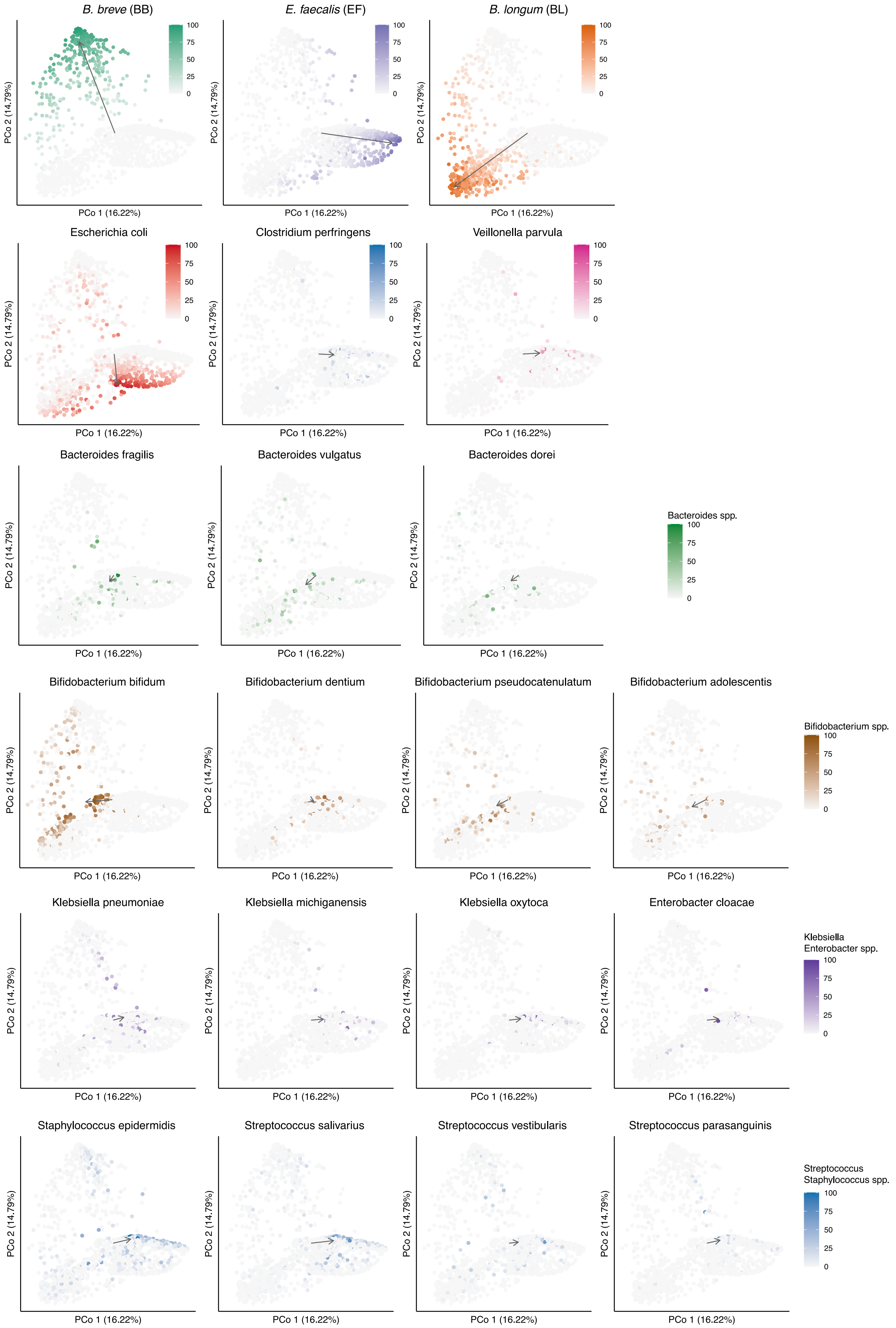


Extended Data Fig. 2 | See next page for caption.

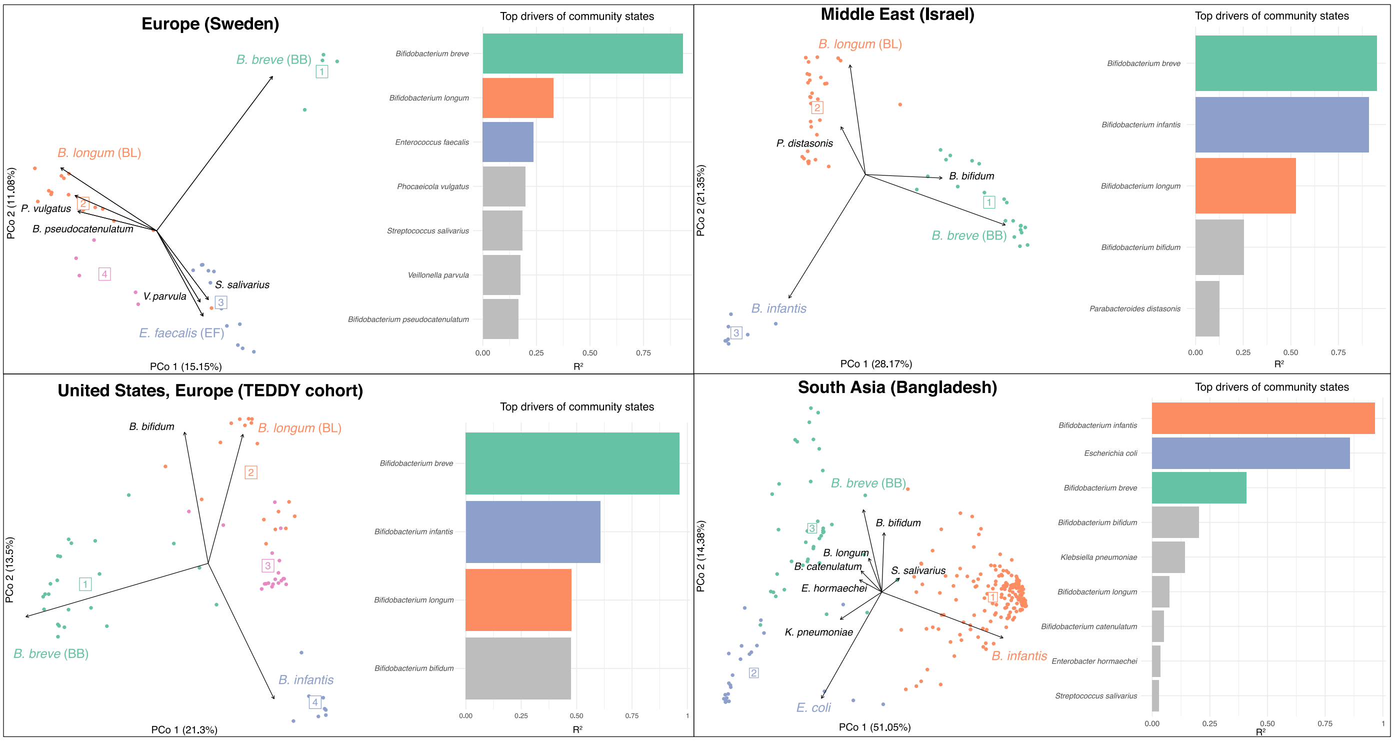
Extended Data Fig. 2 | Consistency of NGM community state assignment across typing methods. (a-d) Identification of three NGM community states using both **(a)** Partitioning Around Medoids (PAM) clustering of JSD, with statistical support from the Calinski-Harabasz (CH) index, and **(c)** Dirichlet Multinomial Mixture (DMM) modelling using the Laplace approximation. (**b, d**) PCoA plots, representing 1,904 neonatal gut metagenomes, are color-coded by community state assignments, and based on species-level Bray-Curtis distances. **(e-g)** PAM-based and DMM-based community state assignment concordance: (**e**) Correlation between community state assignments shown with a Cramér's V correlation of 0.726. The proportions of community states assigned by each method are labelled. The breakdown of community states BB/EF/BL in PAM is 336/827/741, and in DMM is 252/1097/555. (**f**) Overlap in the dominant core species (≥1% mean abundance) in each community state, grouped at the genus level, with exceptions for the driver species *B. breve*, *E. faecalis* and *B. longum*. PAM-based assignment was chosen in downstream analyses given the higher relative abundances of these driver species in their respective community states (versus DMM-based assignment): *B. breve* 67.9% vs 56.5% ($p<0.001$), *E. faecalis* 21.7% vs 16.8% ($p<0.001$), and *B. longum* 27.25% vs 29.90% ($p=0.24$). Wilcoxon-test (two-sided) with FDR correction. (**g**) The top 10 driver species for each DMM-based community state are displayed, ranked by their assignment strength, as indicated on the y-axis.



Extended Data Fig. 3 | Consistency of NGM community state assignment across neonatal time points. Identification of three NGM community states using PAM-based clustering across three major time points in the neonatal period (day 4, N=360; day 7, N=1,149; day 21, N=350). PCoA plots, are color-coded by community state assignments and based on species-level JSD. Ellipses encapsulate 67% of the samples within each respective cluster.

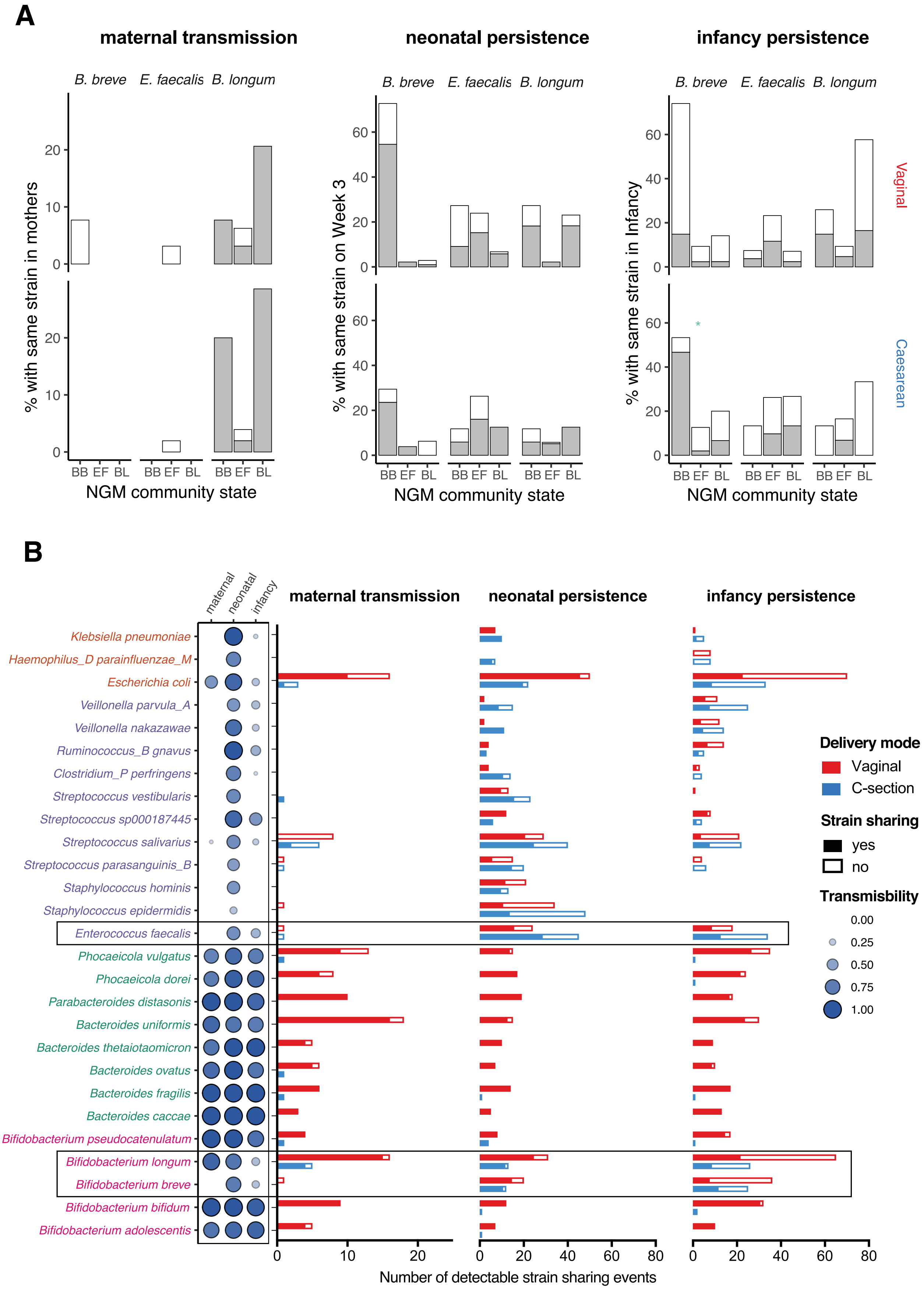


Extended Data Fig. 4 | Abundance and co-occurrence of the NGM community state driver species. PCoA plots depicted in Fig. 1, with arrows, illustrate the scale and direction of core NGM species (>1% mean abundance) driving the formation of NGM community states (clusters). The length of the arrows is scaled to reflect the degree of contribution to the variation in NGM composition, with the arrow points towards increasing species abundance. Species that frequently co-occur with the NGM driver species within their respective community states share the same arrow direction.



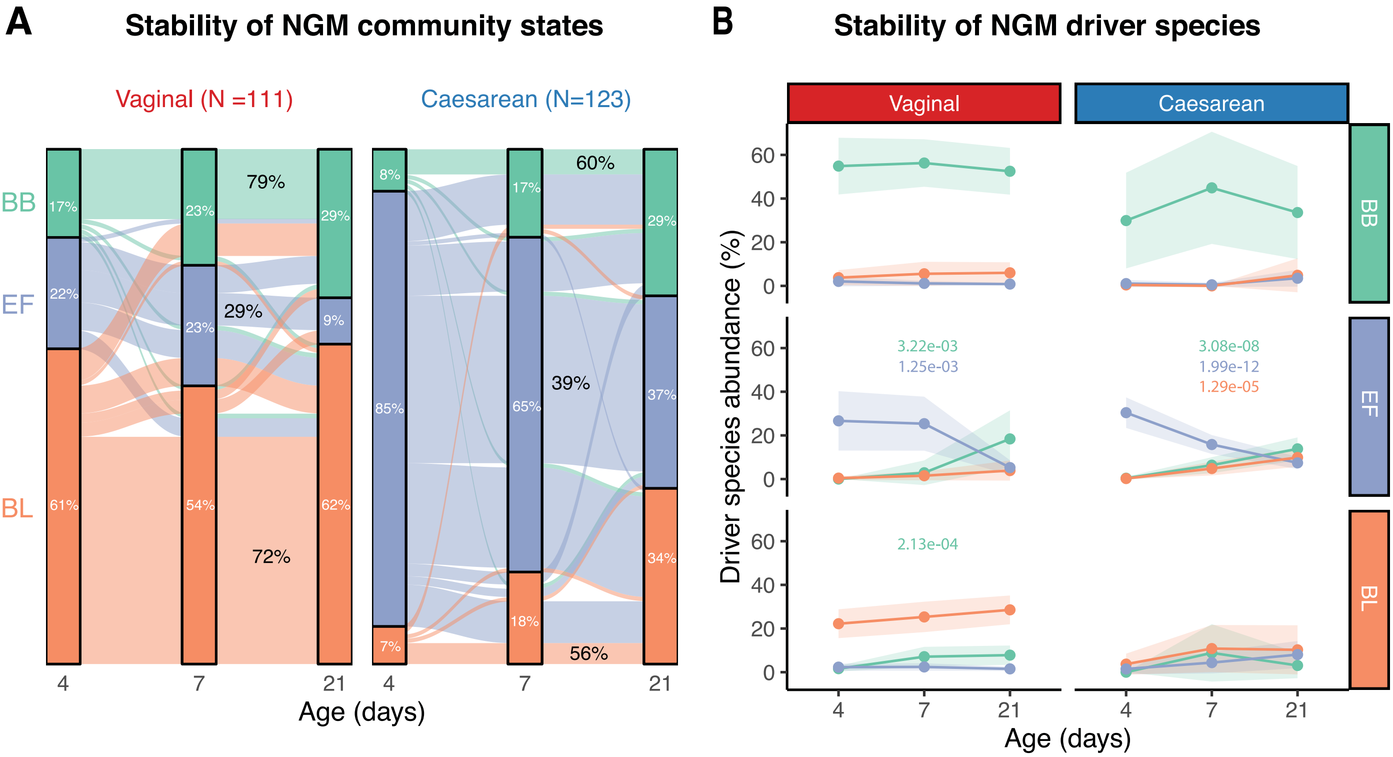
Extended Data Fig. 5 | Validation of NGM community states and driver species across geographies and lifestyles. All three NGM community states and the driver species (*B. breve, B. longum or B. infantis, and E. faecalis*) were independently detected in infant gut metagenomic cohorts (0–6 months) from diverse geographical regions and lifestyles. These include Europe (Sweden, days 4–12, N=37), the United States (TEDDY cohort, months 2–6, N=69), the Middle East (Israel, weeks 1–24, N=60), and South Asia (Bangladesh, months 0–2, N=234). In the Bangladeshi cohort, which is a non-industrialised and non-urban population, the *B. infantis* and *E. coli*-driven clusters are representative of the *B. longum* (closely related to *B. infantis*) and *E. faecalis* (also facultative anaerobe opportunistic pathogen) community states, respectively. The analysis and visualization methods are consistent with those described in Fig. 1a, b.

https://doi.org/10.1038/s41564-024-01804-9



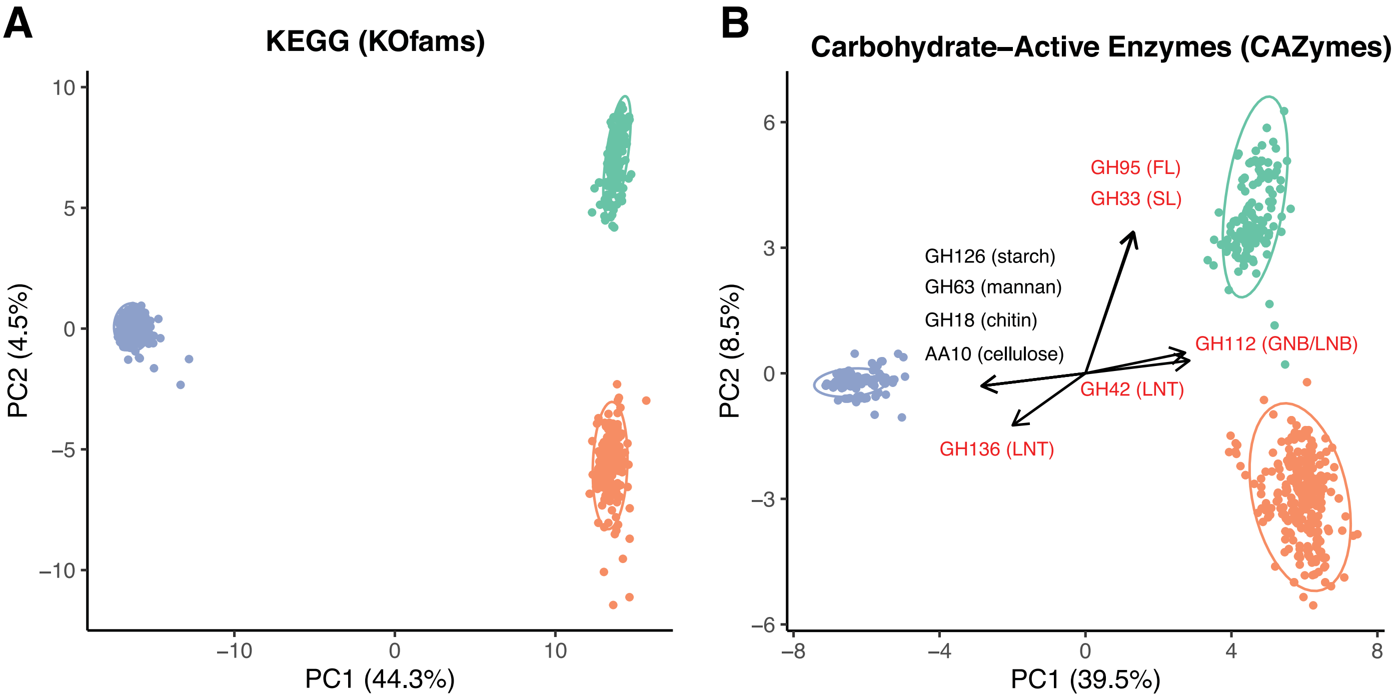
Extended Data Fig. 6 | See next page for caption.

Extended Data Fig. 6 | Strain-level dynamics and stability across NGM species. (a) Frequency of study participants detected with the same strains (in grey, otherwise in white) from their mother's faecal samples across NGM community states. To delineate transmission trends, the chart is categorized by birth mode and the three NGM driver species. Frequency of strain-sharing event (for example, maternal transmission in mother-baby pair or strain persistence within-individual longitudinal samples) is presented as raw counts of detectable strain sharing events normalized by the total number of subjects per birth mode and NGM community state (week 1). **(b)** Bar plots counting strain-sharing events across three settings: (Left) Maternal transmission in mother-infant dyads (183 subjects; 167 transmissions from 213 evaluated species-sample pairs). (Middle) Neonatal persistence via neonatal longitudinal sampling (359 subjects; 700 transmissions from 938 evaluated pairs). (Right) Infancy persistence from neonatal into infancy period (302 subjects; 464 transmissions from 920 evaluated pairs). When longitudinal samples were considered, strain sharing events were considered only once per subject per setting, using the time point with highest counts. Only species with ≥20 strain-sharing events detected across three settings are shown. Three community state driver species are highlighted in boxes. Transmission patterns often align with phylogeny: Actinomycetota/Actinobacteria (pink) and Bacteroidota/Bacteroidetes (green) typically transmit maternally during vaginal birth and persist into infancy. Conversely, Bacillota/Firmicutes (purple) and Pseudomonadota/Proteobacteria (orange) show lower maternal transmission rates and reduced neonatal persistence. Notable outliers include *E. coli* and *B. breve*. The size of bubbles represents the transmissibility of each species, which is its ratio of detected to potential strain-sharing events, as determined by StrainPhlAn4. Only subject pairs with sequencing depth sufficient for StrainPhlAn strain-level analyses are displayed; data points not shown are non-evaluable.

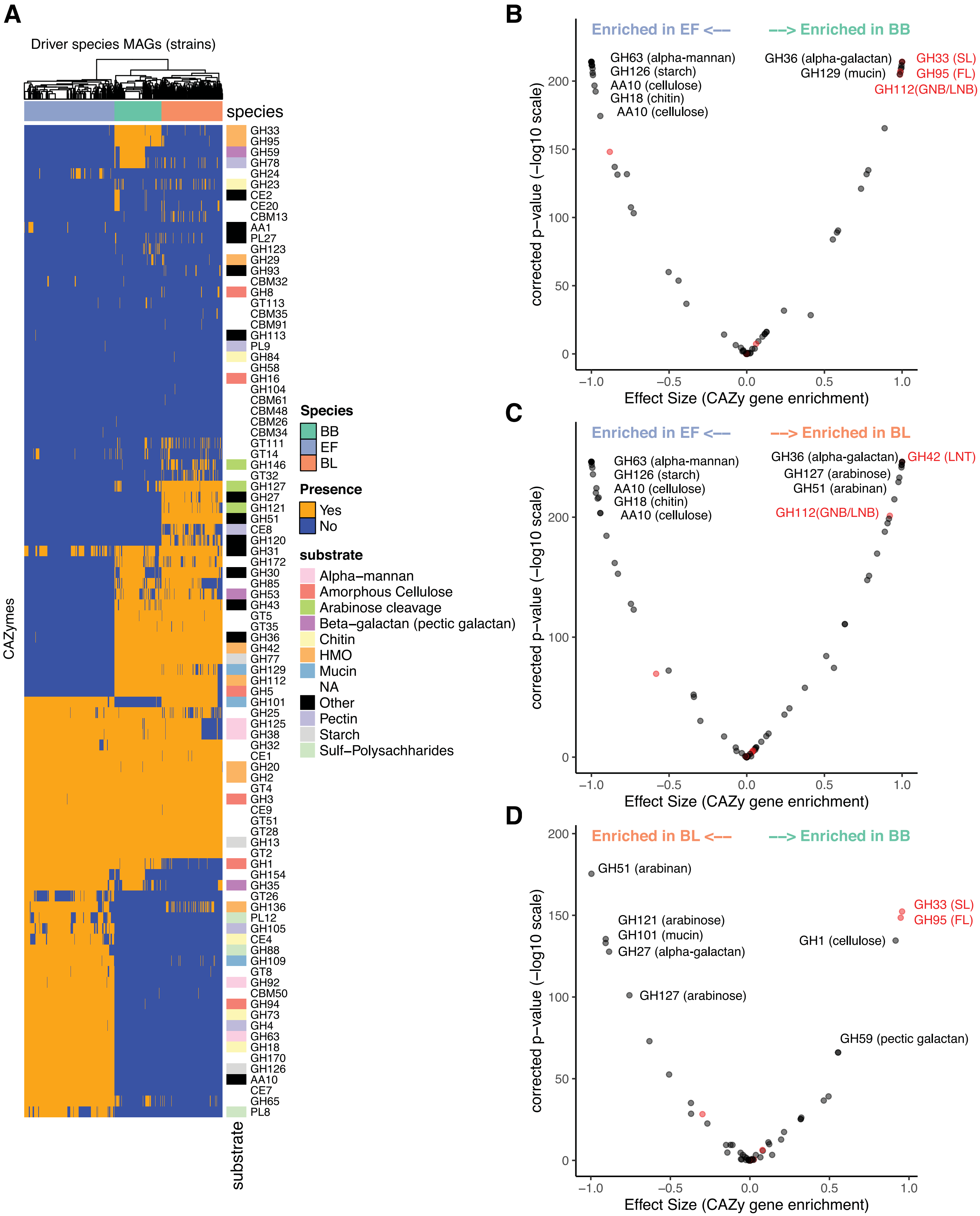


Extended Data Fig. 7 | Colonisation dynamics in neonatal longitudinal samples. (a) Overview of NGM community states of all subjects individually sampled on major neonatal period sampling points day 4, 7 or 21, stratified by birth mode. In VD, N=176/602/156 on day 4, 7, and 21, respectively; In CS, N=184/547/194 on day 4, 7, and 21, respectively. Total samples N=1859. **(b)** Longitudinal shifts in NGM community states and the levels of driver species from week 1 to week 3, based on subjects longitudinally sampled across days 4, 7, and 21, N=234; VD, N=111; CS, N=123). Community states that remained consistent from first (day 4) to the final neonatal longitudinal sampling (day 7 or/and 21), is depicted as a percentage of their starting pool size (labelled in black). Subjects that began with either BB or BL community state on day 4 were significantly more likely to remain in the same community state on day 7 and 21, compared to those that began with EF (pairwise chi-squared tests with FDR correction, q-values < 0.01). However, this trend was not observed as early as day 7 (global chi-squared test, p=0.7043). The colour scheme represents the community states or driver species: BB and *B. breve* in green; EF and *E. faecalis* in purple; BL and *B. longum* in orange. Statistical differences in species abundance between longitudinal samples was determined using paired ANOVA test (two-sided) with FDR correction. Boxplot center line and red point indicate the median and mean, respectively; box limits indicate the upper and lower quartiles; and whiskers indicate 1.5× the interquartile range.

https://doi.org/10.1038/s41564-024-01804-9



Extended Data Fig. 8 | Species-driven functional divergence in NGM community states. (a-b) Principal Component Analysis (PCA) of community state driver enterotype species genomes. Groupings are based on the presence of genes tied to the full metabolic repertoire using **(a)** KEGG orthologs (KOfams) and carbon metabolism via **(b)** Carbohydrate-Active enZYmes (CAZymes). Each dot denotes an individual strain: *B. longum* (BL, N=342) in orange, *B. breve* (BB, N=267) in green, and *E. faecalis* (EF, N=507) in blue. Ellipses encapsulate the 95% confidence intervals. Arrows showcase the contribution of select CAZy genes to principal components (details in Extended Data Fig. 8). CAZy genes for human milk oligosaccharides (HMOs) utilisation are highlighted in red.



Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | Carbon metabolism of NGM community state driver species. (a) A heatmap displays the clustering of carbohydrate-active enzymes (CAZymes) across the genomes of three driver species. Genes are coloured based on their corresponding carbohydrate substrate categories. **(b-d)** Volcano plots depict differentially enriched CAZymes in each driver species, comparing **(b)** BB vs. EF, **(c)** BL vs. EF, and **(d)** BB vs. BL. The effect size represents the difference in the proportion of genes between species. P-values are adjusted using Fisher's exact test (two-sided) with FDR correction. Genes related to HMO metabolism are marked in red. Significantly enriched genes are labelled for clarity. Arrows at the top indicate the direction of species enrichment in each comparison. *B. longum* (BL, N=342) is shown in orange, *B. breve* (BB, N=267) in green, and *E. faecalis* (EF, N=507) in blue.

nature portfolio

Corresponding author(s): Yan Shao
Trevor D. Lawley

Last updated by author(s): Jul 5, 2024

Reporting Summary

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

Statistics

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

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

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

Software and code

Policy information about availability of computer code

Data collection	No software was used for data collection.
Data analysis	All software used for data analyses is publicly available, and listed as follows: InStrain v1.3.0, bowtie2 v2.3.5, StrainPhlAn4 v4.0.6, phyloseq v1.48, RStudio v4.1.0, R packages cluster v2.1.4, DirichletMultinomial v1.4, tidyverse v2.0.0, ggalluvial v0.12.5, lmerTest v3.1-3, Prism 9, seqkit v2.4.0, shovill v1.1.0, SPAdes v3.15.5, MEGAHIT v1.1.3, bwa-mem v0.7.17, MetaBAT2 v2.13, MaxBin2 v2.2.4 and CONCOCT v0.4, samtools v1.5, metaWRAP v1.2, CheckM2 v1.0.1, GTDB-Tk v2.3.0., DRAM v1.4.5, ABRicate v1.0.1

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

Data

Policy information about availability of data

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

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Shotgun metagenomic sequencing data (after quality trimming and human decontamination) of the entire Baby Biome Study cohort have been deposited to the

European Nucleotide Archive under study accession number ERP115334. Bacterial genome assemblies for the three species analysed have been deposited to Zenodo (https://doi.org/10.5281/zenodo.12667210). Sample metadata and subject-level clinical metadata of de-identified study participants are provided in the Supplementary Tables. The raw faecal samples and bacterial isolates are available from the corresponding authors upon request.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	Sex of the study participants was collected in this study based on the clinical records at birth. Sex has been included as a variable in the molecular epidemiological analyses, with cohort-level data summarized in Table 1, and individual-level data listed in Supplementary Table 1.
Reporting on race, ethnicity, or other socially relevant groupings	Maternal ethnicity of the study participants was collected in this study based on a self-reported questionnaire at recruitment. Maternal ethnicity has been included as a variable in the molecular epidemiological analyses, with cohort-level data summarized in Table 1, and individual-level data listed in Supplementary Table 1.
Population characteristics	All clinical covariates of the study cohort are summarised in Table 1. A total of 20 variables were included in the final analyses based on clinical relevance, quality of data and completeness (N=6 maternal, N=8 perinatal or at time of delivery, N=5 postnatal, N=1 at the time of stool sample collection variables).
Recruitment	Participants were recruited on a voluntary basis in the study hospitals. Mothers provided written informed consent for their participation, and for the participation of their children, in the study.
Ethics oversight	The study was approved by the NHS London – City and East Research Ethics Committee (REC reference 12/LO/1492). The study was performed in compliance with all relevant ethical regulations.

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

Field-specific reporting

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

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

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

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	All fecal samples collected in this study were used if available. No statistical methods were used to pre-determine sample sizes, but this study already represents the largest dataset of longitudinal fecal metagenomes (n = 1,904; n = 2,387 including infancy samples) of newborn babies (n = 1,288).
Data exclusions	None. All sequencing samples that passed sequencing quality control were included for analysis.
Replication	All experimental data presented include replicates, with the number of biological replicates stated in the figure captions.
Randomization	Randomization was not employed in microbiome analyses as this is an observational study. Mice were allocated randomly into three replicate cages per experimental condition.
Blinding	No blinding used in microbiome analyses as this is an observational study. Biological counting experiments were blinded by another person than the experimenter before being counted as to avoid experimental bias.

Reporting for specific materials, systems and methods

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

Materials & experimental systems

n/a | Involved in the study

☒ ☐ Antibodies

☒ ☐ Eukaryotic cell lines

☒ ☐ Palaeontology and archaeology

☐ ☒ Animals and other organisms

☒ ☐ Clinical data

☒ ☐ Dual use research of concern

☒ ☐ Plants

Methods

n/a | Involved in the study

☒ ☐ ChIP-seq

☒ ☐ Flow cytometry

☒ ☐ MRI-based neuroimaging

Animals and other research organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research, and Sex and Gender in Research

Laboratory animals	Wild-type C57BL/6N mice; median age 6 weeks; SD 12 days.
Wild animals	This study did not involve wild animals.
Reporting on sex	No sex-based analyses have been performed. Cages of male and female mice were randomly allocated to experimental groups.
Field-collected samples	This study did not involve field-collected samples.
Ethics oversight	Mice were maintained under germ-free conditions at the Wellcome Sanger Institute Home Office-approved facility, with all procedures carried out in accordance with the United Kingdom Animals (Scientific Procedures) Act of 1986 under Home Office approval (PPL no. 80/2643).

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

Plants

Seed stocks	*Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.*
Novel plant genotypes	*Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.*
Authentication	*Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.*