

Stunted microbiota and opportunistic pathogen colonization in caesarean-section birth

Yan Shao¹, Samuel C. Forster^{1,2,3}, Evdokia Tsaliki⁴, Kevin Vervier¹, Angela Strang⁴, Nandi Simpson⁴, Nitin Kumar¹, Mark D. Stares¹, Alison Rodger⁴, Peter Brocklehurst⁵, Nigel Field^{4*} & Trevor D. Lawley^{1*}

Immediately after birth, newborn babies experience rapid colonization by microorganisms from their mothers and the surrounding environment¹. Diseases in childhood and later in life are potentially mediated by the perturbation of the colonization of the infant gut microbiota². However, the effects of delivery via caesarean section on the earliest stages of the acquisition and development of the gut microbiota, during the neonatal period (≤ 1 month), remain controversial^{3,4}. Here we report the disrupted transmission of maternal *Bacteroides* strains, and high-level colonization by opportunistic pathogens associated with the hospital environment (including *Enterococcus*, *Enterobacter* and *Klebsiella* species), in babies delivered by caesarean section. These effects were also seen, to a lesser extent, in vaginally delivered babies whose mothers underwent antibiotic prophylaxis and in babies who were not breastfed during the neonatal period. We applied longitudinal sampling and whole-genome shotgun metagenomic analysis to 1,679 gut microbiota samples (taken at several time points during the neonatal period, and in infancy) from 596 full-term babies born in UK hospitals; for a subset of these babies, we collected additional matched samples from mothers (175 mothers paired with 178 babies). This analysis demonstrates that the mode of delivery is a significant factor that affects the composition of the gut microbiota throughout the neonatal period, and into infancy. Matched large-scale culturing and whole-genome sequencing of over 800 bacterial strains from these babies identified virulence factors and clinically relevant antimicrobial resistance in opportunistic pathogens that may predispose individuals to opportunistic infections. Our findings highlight the critical role of the local environment in establishing the gut microbiota in very early life, and identify colonization with antimicrobial-resistance-containing opportunistic pathogens as a previously underappreciated risk factor in hospital births.

The acquisition and development of the gut microbiota during early life follow successive waves of exposures to and colonization by microorganisms, which shapes the longer-term composition and function of the microbiota⁵. Events in early life—including delivery by caesarean section^{1,4,6–8}, formula feeding^{7,9} and antibiotic exposure^{7,10}—that could perturb the composition of the gut microbiota are associated with the development of childhood asthma and atopy^{11–13}. Although recent studies^{7,8,10,14,15} have provided substantial insights into the development of the gut microbiota during the first three years of life, many of these studies have been limited by the taxonomic resolution provided by 16S rRNA gene profiling, small sample size or limited sampling during the first month of life (the neonatal period). High-resolution metagenomic studies of large, longitudinal cohorts are required to establish the effects of events in early life on the assembly of the gut microbiota, and any associated risks—particularly for the neonatal period, during which pioneering microorganisms could influence the subsequent development of the microbiota and immune system^{16,17}.

To characterize the trajectory of the acquisition and development of the gut microbiota during the neonatal period, we enrolled 596 healthy,

full-term babies (39.5 ± 1.37 weeks of gestation, 314 vaginal births and 282 births via caesarean section) (Fig. 1a, Extended Data Table 1, Supplementary Table 1) through the Baby Biome Study (BBS). Faecal samples were collected from all babies at least once during their neonatal period (≤ 1 month) and 302 babies were resampled later, during infancy (8.75 ± 1.98 months). Maternal faecal samples were also obtained, from 175 mothers paired with 178 babies. Metagenomic analysis of the 1,679 faecal samples taken in total revealed the temporal dynamics of the development of the gut microbiota (Fig. 1b), as well as increased diversity of the microbiota with age (Extended Data Fig. 1a). The gut microbiota exhibited substantial heterogeneity between individuals, and substantial instability (intra-individual variability) during the first weeks of life (Extended Data Fig. 1b). Inter-individual differences explained 57% of the variation in microbial species composition (permutational multivariate analysis of variance (PERMANOVA), $P < 0.001$, 1,000 permutations); this was followed by age at sampling, which explained 5.7% of the variance ($P < 0.001$). These results indicate that the gut microbiota is highly dynamic and individualized during the neonatal period—even more so than was observed during infancy (Extended Data Fig. 1c).

To determine the effect of clinical covariates on the composition of the gut microbiota, we performed cross-sectional PERMANOVA and stratified by age. The mode of delivery was the most significant factor to drive variation in the gut microbiota during the neonatal period (Fig. 2a, Supplementary Table 2). Breastfeeding, as well as clinical covariates that are associated with hospital birth (such as the use of perinatal antibiotics and the duration of the stay in hospital), exhibited smaller effects (Supplementary Note 1). The largest effect of the mode of delivery was observed on day 4 ($R^2 = 7.64\%$, $P < 0.001$) (Extended Data Fig. 2); this effect dissipated with age but remained significant at the point of sampling in infancy ($R^2 = 1.00\%$, $P = 0.002$). No differences were observed between the maternal gut microbiotas by mode of delivery, or between the neonatal gut microbiotas after elective and emergency births via caesarean section (Supplementary Table 2).

Given the significant effect of the mode of delivery during the neonatal period, we next sought to understand how the composition and developmental trajectory of the microbiota were altered. Samples from babies delivered vaginally were enriched with species of *Bifidobacterium* (such as *Bifidobacterium longum* and *Bifidobacterium breve*), *Escherichia* (*Escherichia coli*), *Bacteroides* (*Bacteroides vulgatus*) and *Parabacteroides* (*Parabacteroides distasonis*), and these commensal genera comprised 68.3% (95% confidence interval, 65.7–71.0%) of the neonatal gut microbiota (Fig. 2b, Supplementary Table 3), consistent with recent observations in other cohorts^{4,8}. By contrast, the gut microbiotas of babies delivered by caesarean section were depleted of these commensal genera and instead were dominated by *Enterococcus faecalis*, *Enterococcus faecium*, *Staphylococcus epidermidis*, *Streptococcus parasanguinis*, *Klebsiella oxytoca*, *Klebsiella pneumoniae*, *Enterobacter cloacae* and *Clostridium perfringens*, all of which are commonly associated with the hospital environment¹⁸ and hospitalized preterm babies^{19–21}.

¹Host-Microbiota Interactions Laboratory, Wellcome Sanger Institute, Hinxton, UK. ²Centre for Innate Immunity and Infectious Diseases, Hudson Institute of Medical Research, Clayton, Victoria, Australia. ³Department of Molecular and Translational Sciences, Monash University, Clayton, Victoria, Australia. ⁴Institute for Global Health, University College London, London, UK. ⁵Birmingham Clinical Trials Unit, University of Birmingham, Birmingham, UK. *e-mail: nigel.field@ucl.ac.uk; tl2@sanger.ac.uk

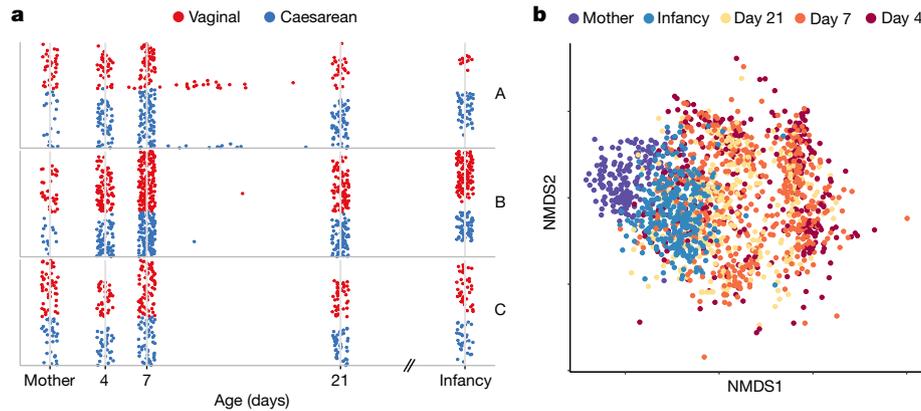


Fig. 1 | Developmental dynamics of the gut microbiota of newborn babies. **a**, Longitudinal metagenomic sampling of 1,679 gut microbiotas, from 771 BBS participants in 3 UK hospitals (labelled A, B and C). Each row corresponds to the time course of a subject. Five hundred and ninety-six babies were sampled during the neonatal period, primarily on day 4 ($n = 310$), day 7 ($n = 532$) and day 21 ($n = 325$), and in infancy

On day 4, species that belong to these genera accounted for 68.25% (95% confidence interval, 62.74–73.75%) of the total microbiota composition in babies delivered by caesarean section (Fig. 2b).

Previous studies have reported that—compared to babies delivered by caesarean section—the gut microbiotas of vaginally delivered babies were enriched in lactobacilli associated with the microbiota of the mother's vagina^{1,22}. Here, however, we observed no statistical difference in the prevalence (present at over 1% abundance in 11.9% and 15.7% of the microbiota of babies delivered vaginally or by caesarean section, respectively) or abundance of *Lactobacillus* between babies delivered vaginally (1.217%; 95% confidence interval, 0.81–1.621%) or by caesarean section (2.21%; 95% confidence interval, 1.54–2.88%). Instead, commensal species from the *Bacteroides* genus were detected at high abundance in the gut microbiota of 51.0% (160 out of 314) of vaginally delivered babies (mean relative abundance 8.13%; 95% confidence interval, 6.88–9.39%) (Extended Data Fig. 3). By contrast, *Bacteroides* species were low or absent in the gut microbiota of 99.6% (281 out of 282) babies delivered by caesarean section (mean relative abundance 0.43%; 95% confidence interval, 0.11–0.74%). In 60.6% (86 out of 142) of the babies delivered by caesarean section, this low-*Bacteroides* profile (see 'Classification of babies with the low-*Bacteroides* profile' in Methods for definition of this profile) persisted into infancy, by which point only species of *Bacteroides* were differentially abundant between the gut microbiotas of babies delivered vaginally and by caesarean section (Supplementary Table 3). Although we could not assess the independent effect of maternal exposure to antibiotics during delivery by caesarean section (as antibiotics were administered in all deliveries by caesarean section), among vaginally delivered babies we observed a significant association between the low-*Bacteroides* profile and maternal intrapartum antibiotic prophylaxis (odds ratio 1.77, 95% confidence interval, 1.17–2.71%, $P = 0.0074$), which also accounted for the greatest amount of variation in the gut microbiota of vaginally delivered babies ($R^2 = 5.88$ –13.6%) (Supplementary Table 2). These results expand on previous findings^{10,23}, and highlight the low-*Bacteroides* profile as a perturbation signature that is associated with delivery by caesarean section and with maternal intrapartum antibiotic prophylaxis in vaginal delivery.

The transmission of gastrointestinal bacteria from mothers to their babies is an underappreciated form of maternal kinship²⁴. To assess whether variation in the neonatal microbiota could be attributed to differential transmission of the maternal microbiota, we profiled the transmission of bacterial strains across 178 mother–baby dyads. We show that the majority of transmissions of maternal microbial strains during the neonatal period occurred in vaginally delivered babies (74.39%), at much higher frequency than was observed for babies delivered by caesarean section (12.56%, Fisher's exact test, $P < 0.0001$)

(8.75 ± 1.98 months of age, $n = 302$), as well as 175 matched mothers. C-section, caesarean section. **b**, Nonmetric multidimensional scaling (NMDS) ordination of bacterial beta-diversities, measured by Bray–Curtis dissimilarity between species relative abundance profiles ($n = 1,679$ samples).

(Fig. 3a, Extended Data Fig. 4, Supplementary Table 4). *Bacteroides* spp., *Parabacteroides* spp., *E. coli* and *Bifidobacterium* spp. were most frequently transmitted from mothers to babies through vaginal birth, consistent with previous observations^{4,25–27}. For *Bacteroides* species such as *B. vulgatus* (Fig. 3b), the lack of transmission continued far beyond the neonatal period in babies delivered by caesarean section²⁵:

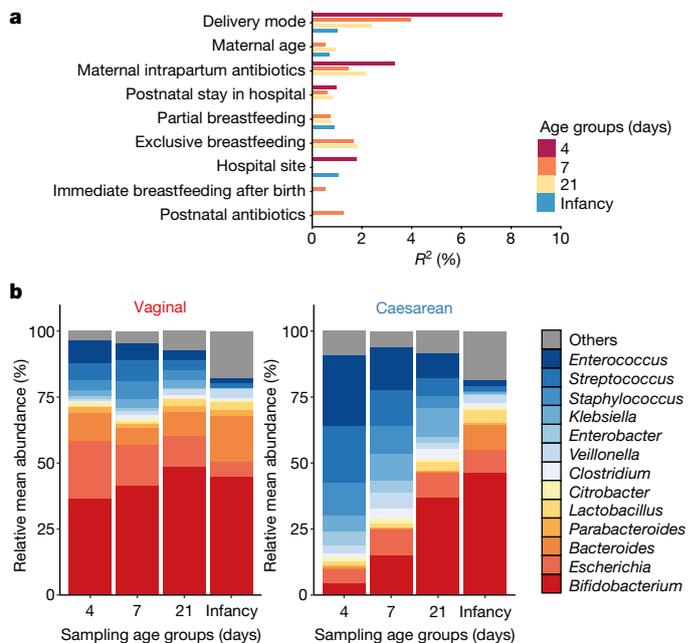


Fig. 2 | Perturbed composition and development of the neonatal gut microbiota associated with delivery by caesarean section. **a**, Bar plot illustrating the clinical covariates that are associated with variation in the neonatal gut microbiota on day 4 ($n = 310$ individuals), day 7 ($n = 532$ individuals), day 21 ($n = 325$ individuals) and in infancy ($n = 302$ individuals). Only significant associations in cross-sectional tests are shown. Covariates are ranked by the number of significant effects observed across sampling-age groups. The proportion of explained variance (R^2) and significance were determined by PERMANOVA on between-sample Bray–Curtis dissimilarity. **b**, Longitudinal changes in the mean relative abundance of genera of faecal bacteria, sampled on day 4, day 7, day 21 and in infancy, for genera with >1% mean relative abundance across all samples from the neonatal period. Vaginal deliveries, $n = 744$ samples from 310 babies; deliveries via caesarean section, $n = 725$ samples from 281 babies.

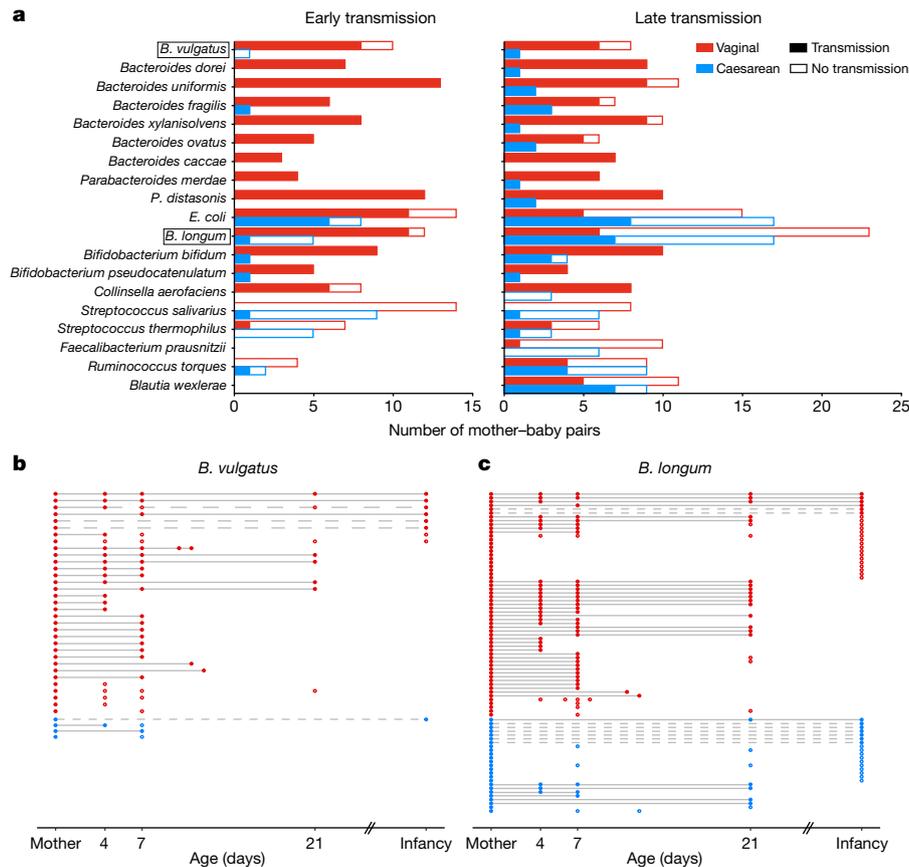


Fig. 3 | Disrupted transmission of maternal microbial strains in babies delivered by caesarean section. **a**, Early and late transmission of maternal microbial strains in mother–baby pairs (35 pairs for vaginally delivered babies and 24 pairs for babies delivered by caesarean section), longitudinally sampled during the neonatal (early) and infancy (late) period. Only the frequently shared species that were detected with sufficient coverage for strain analysis in more than ten pairs are shown. **b**, **c**, Transmission events of maternal strains of *B. vulgatus* (**b**)

and *B. longum* (**c**) in vaginally delivered babies and babies delivered via caesarean section, over time. In each row of mother–baby paired samples, each circle represents a detectable strain either identical to (filled) or distinct from (hollow) the maternal strain. Across the rows, identical strains are either linked by a solid line (which represents early transmission and persistence to infancy) or a dashed line (which indicates late transmission).

the transmission of *B. vulgatus* was rarely detected later in infancy. This is in contrast to the transmission pattern of other common early colonizers (such as *B. longum* (Fig. 3c) and *E. coli*), for which colonizations of maternal strains occurred more frequently later in infancy (Fisher’s exact test, $P = 0.0479$ and $P = 0.0226$, respectively). This result highlights the neonatal period as a critical early window of maternal transmission, as shown by the disrupted transmission of pioneering *Bacteroides* species that is evident in babies delivered by caesarean section (who show long-term absences of these *Bacteroides* species).

Babies delivered by caesarean section were deprived of maternally transmitted commensal bacteria, but had a substantially higher relative abundance of opportunistic pathogens that are commonly associated with the hospital environment. These enriched species included *E. faecalis*, *E. faecium*, *E. cloacae*, *K. pneumoniae*, *K. oxytoca* and *C. perfringens* (Fig. 4a, Supplementary Table 3), some of which are members of the ESKAPE (*E. faecium*, *Staphylococcus aureus*, *K. pneumoniae*, *Acinetobacter baumannii*, *Pseudomonas aeruginosa* and *Enterobacter* spp.) pathogens that are responsible for the majority of nosocomial infections²⁸. Their frequent colonization of the gut microbiota of newborn babies delivered by caesarean section was under-reported in previous cohorts^{3,8} that had insufficient statistical power (Supplementary Note 2). Among babies delivered by caesarean section, 83.7% carried opportunistic pathogen species during the neonatal period (as defined in ‘Classification of the opportunistic pathogen carriage’ in Methods), in comparison to 49.4% of the vaginally delivered babies (Fig. 4a). During the first 21 days of life, these opportunistic pathogens accounted for 30.4% (95% confidence interval, 27.86–32.96%) of

the species-level abundance in the gut microbiota of babies delivered by caesarean section, compared to 9.8% (95% confidence interval, 8.19–11.4%) in the vaginally delivered babies; the greatest difference was observed on day 4 (Extended Data Fig. 5a). Longitudinally, the difference in the abundance of all opportunistic pathogens combined persisted in the babies delivered by caesarean section who were resampled later during infancy (abundance in babies delivered by caesarean section of 2.8%, versus 1.6% in vaginally delivered babies; $P = 0.0375$, Welch’s *t*-test). The frequent and abundant carriage of opportunistic pathogens was also observed in vaginally delivered babies who had the low-*Bacteroides* profile (Extended Data Fig. 5b), and the absence of breastfeeding during the neonatal period was associated with a higher carriage of *C. perfringens*, *K. oxytoca* and *E. faecalis* (Supplementary Table 3).

Given the prevalent carriage of opportunistic pathogens in the metagenomes of the neonatal gut, we sought to validate the presence and viability of these pathogens with culturing. We undertook targeted large-scale culturing of 836 strains of opportunistic pathogens in the faecal samples of 177 babies (70 vaginally delivered babies and 107 babies delivered by caesarean section; a total of 741 isolates) and 38 mothers (95 isolates) using selective medium (Fig. 4b, Supplementary Table 5). Subsequent whole-genome sequencing and genomic characterization of *E. faecalis* ($n = 356$ isolates), *E. cloacae* ($n = 52$ isolates), *K. oxytoca* ($n = 150$ isolates) and *K. pneumoniae* ($n = 78$ isolates) enabled us to perform high-resolution phylogenetic analysis and to delineate the strain-specific carriage of virulence factors and genes associated with antimicrobial resistance (AMR).

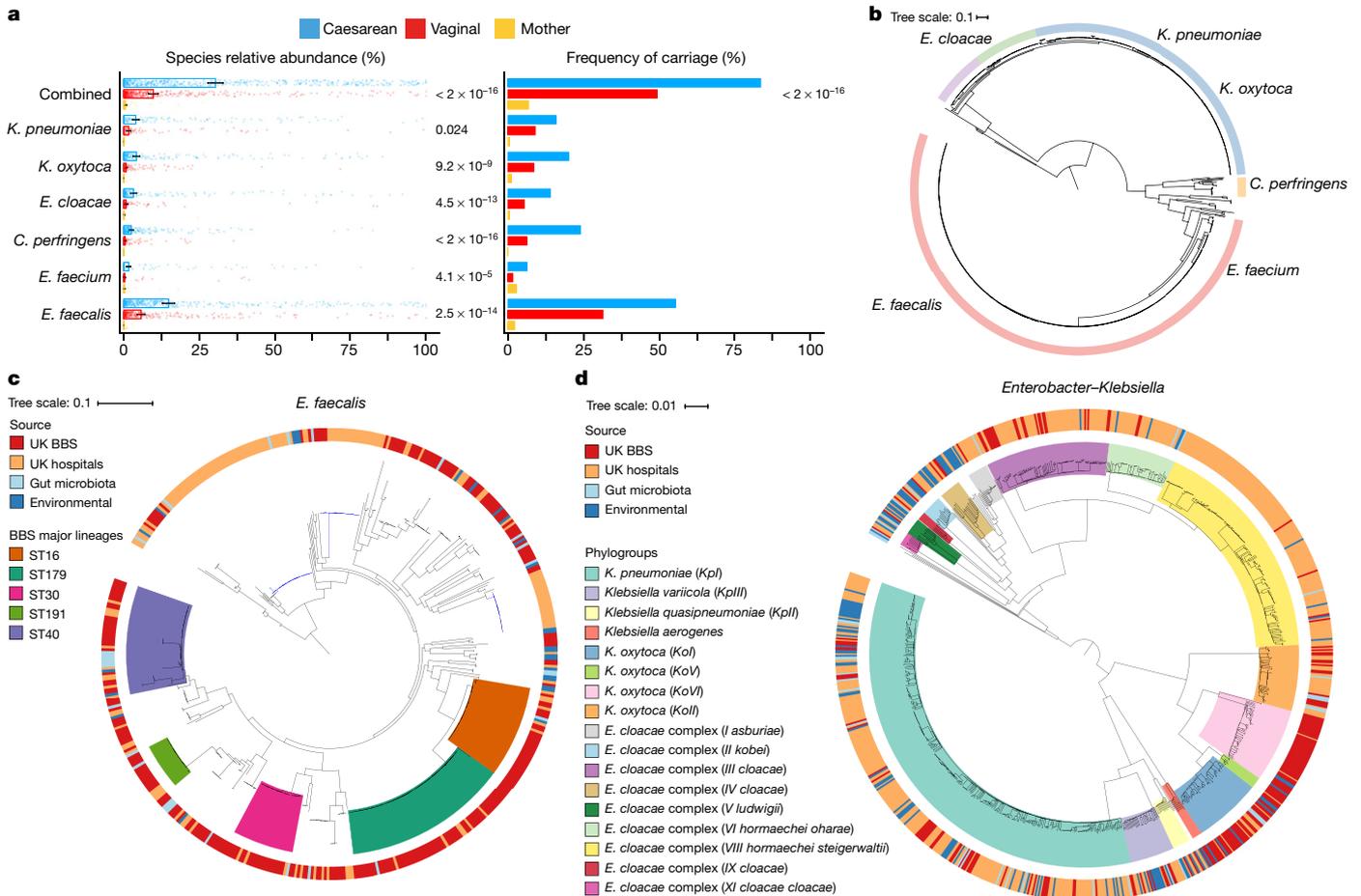


Fig. 4 | Extensive and frequent colonization of babies delivered by caesarean section with diverse opportunistic pathogens. **a**, The mean relative abundance and frequency ($>1\%$ mean relative abundance) of six opportunistic pathogen species that are enriched in babies delivered via caesarean section ($n = 596$ samples), compared to vaginally delivered babies ($n = 606$ samples) during the first 21 days of life, in the context of the maternal-level carriage ($n = 175$ individuals). Error bars indicate the 95% confidence interval of the mean relative abundance. The significance (P values indicated to the right of the bars) of the difference in mean relative abundance and combined pathogen-carriage frequency between babies delivered vaginally or via caesarean section was determined by two-sided Wilcoxon signed-rank test and Fisher's exact tests, respectively. **b**, Phylogenetic representation of 836 bacterial isolates cultured from raw faecal samples, including the abovementioned 6 opportunistic

Focusing on the most-prevalent opportunistic pathogen in babies delivered by caesarean section, we analysed the genomes of a diverse population of *E. faecalis* strains from the BBS in the context of publicly available genomes of the human gut microbiota and environmental strains of *E. faecalis* (Fig. 4c). We found that 53.9% of strains in the BBS were represented by 5 major lineages, each of which was distributed across vaginally delivered babies and babies delivered by caesarean section (and their mothers) in the three hospitals of the BBS (Extended Data Fig. 6a) and across patients in UK hospitals, but did not include high-risk UK epidemic lineages that are enriched in multi-drug resistance and virulence. Consistent with the phylogenetic placement of strains of the BBS with the human gastrointestinal and environmental strains, these non-epidemic *E. faecalis* strains exhibited comparable levels of carriage of genes related to AMR (Extended Data Fig. 6b–e, Supplementary Note 3). Similar to *E. faecalis*, *Enterobacter* and *Klebsiella* strains in the BBS also exhibited high-level population diversity, including the phylogenetic under-representation of epidemic lineages (Fig. 4d, Extended Data Fig. 7). These strains also showed levels of carriage of genes associated with AMR and virulence that were

pathogen species isolated from 5 major genera: *Enterococcus* spp. (red, $n = 451$ isolates); *Clostridium* spp. (yellow, $n = 24$ isolates); *Klebsiella* spp. (blue, $n = 235$ isolates), *Enterobacter* spp. (green, $n = 52$ isolates) and *Escherichia* spp. (purple, $n = 41$ isolates). **c**, Phylogeny of *E. faecalis* isolates of the BBS ($n = 282$ isolates) in the context of public isolates from the UK hospitals ($n = 168$ isolates), the human gut microbiota ($n = 28$ isolates) and environmental sources ($n = 27$ isolates) with the high-risk UK epidemic lineage branches coloured in blue. Midpoint-rooted maximum-likelihood tree is based on single-nucleotide polymorphisms in 1,656 core genes. **d**, Diverse strain populations of the *Enterobacter-Klebsiella* complex among the BBS collection ($n = 202$ isolates), in the context of the UK hospital ($n = 604$), the human gut microbiota ($n = 37$) and environmental strains ($n = 120$).

indicative of non-epidemic lineages that circulate in the hospital environment and healthy populations, rather than epidemic lineages that are hypervirulent and enriched for extended-spectrum β -lactamases (Extended Data Fig. 8, Supplementary Note 3). Given the previous isolation of the major BBS lineages in hospitalized patients with bloodstream infections, and their AMR and virulence capabilities, any level of opportunistic pathogen carriage represents a considerable risk of opportunistic infections, especially for the babies delivered by caesarean section who have a high prevalence (83.7%) of carriage.

Although there is insufficient evidence from metagenomics and the whole-genome sequencing of cultured isolates to rule out a maternal origin for the opportunistic pathogens (Supplementary Note 4), the absence of lineage-specific colonization suggests that exposure to the hospital environment is the primary factor that drives colonization by opportunistic pathogens in the babies of the BBS. Although our study was not designed for the retrospective sampling of the hospital environmental sources, opportunistic pathogens are frequently found in the hospital environment; hospital-born babies have been shown to carry the same bacteria as are present in operating rooms²⁹ and neonatal intensive care units³⁰.

Our longitudinal, whole-genome sequencing characterization of the human gut microbiota during the previously undersampled neonatal period (≤ 1 month) enables us to consolidate recent findings that have shown that mode of delivery is a major factor that shapes the gut microbiota in the first few weeks of life⁴, with a diminished effect that persists into infancy^{14,15}. The disrupted transmission of the maternal gastrointestinal bacteria (particularly pioneering *Bacteroides* species) through delivery by caesarean section and maternal intrapartum antibiotic prophylaxis predisposed newborn babies to colonization by clinically important opportunistic pathogens that circulate in the hospital environment. However, the clinical consequences of the perturbations of early-life microbiota and the carriage of immunogenic pathogens during this critical window of immune development remain to be determined. This highlights the need for large-scale, long-term cohort studies that also sample home births³¹ to better understand the consequence of the perinatal factors in hospital birth and establish whether perturbation of the neonatal microbiota negatively affects health outcomes in childhood and later life.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-019-1560-1>

Received: 4 December 2018; Accepted: 16 August 2019;

- Dominguez-Bello, M. G. et al. Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. *Proc. Natl Acad. Sci. USA* **107**, 11971–11975 (2010).
- Tamburini, S., Shen, N., Wu, H. C. & Clemente, J. C. The microbiome in early life: implications for health outcomes. *Nat. Med.* **22**, 713–722 (2016).
- Chu, D. M. et al. Maturation of the infant microbiome community structure and function across multiple body sites and in relation to mode of delivery. *Nat. Med.* **23**, 314–326 (2017).
- Wampach, L. et al. Birth mode is associated with earliest strain-conferred gut microbiome functions and immunostimulatory potential. *Nat. Commun.* **9**, 5091 (2018).
- Koenig, J. E. et al. Succession of microbial consortia in the developing infant gut microbiome. *Proc. Natl Acad. Sci. USA* **108**, 4578–4585 (2011).
- Stokholm, J. et al. Caesarean section changes neonatal gut colonization. *J. Allergy Clin. Immunol.* **138**, 881–889.e2 (2016).
- Bokulich, N. A. et al. Antibiotics, birth mode, and diet shape microbiome maturation during early life. *Sci. Transl. Med.* **8**, 343ra82 (2016).
- 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).
- Baumann-Dudenhoeffer, A. M., D'Souza, A. W., Tarr, P. I., Warner, B. B. & Dantas, G. Infant diet and maternal gestational weight gain predict early metabolic maturation of gut microbiomes. *Nat. Med.* **24**, 1822–1829 (2018).
- 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).
- Arrieta, M.-C. et al. Early infancy microbial and metabolic alterations affect risk of childhood asthma. *Sci. Transl. Med.* **7**, 307ra152 (2015).
- Fujimura, K. E. et al. Neonatal gut microbiota associates with childhood multisensitized atopy and T cell differentiation. *Nat. Med.* **22**, 1187–1191 (2016).
- Stokholm, J. et al. Maturation of the gut microbiome and risk of asthma in childhood. *Nat. Commun.* **9**, 141 (2018).
- Stewart, C. J. et al. Temporal development of the gut microbiome in early childhood from the TEDDY study. *Nature* **562**, 583–588 (2018).
- Vatanen, T. et al. The human gut microbiome in early-onset type 1 diabetes from the TEDDY study. *Nature* **562**, 589–594 (2018).
- Vatanen, T. et al. Variation in microbiome LPS immunogenicity contributes to autoimmunity in humans. *Cell* **165**, 1551 (2016).
- Olin, A. et al. Stereotypic immune system development in newborn children. *Cell* **174**, 1277–1292.e14 (2018).
- Lax, S. et al. Bacterial colonization and succession in a newly opened hospital. *Sci. Transl. Med.* **9**, eaah6500 (2017).
- Stewart, C. J. et al. Preterm gut microbiota and metabolome following discharge from intensive care. *Sci. Rep.* **5**, 17141 (2015).
- Gibson, M. K. et al. Developmental dynamics of the preterm infant gut microbiota and antibiotic resistance. *Nat. Microbiol.* **1**, 16024 (2016).
- Raveh-Sadka, T. et al. Evidence for persistent and shared bacterial strains against a background of largely unique gut colonization in hospitalized premature infants. *ISME J.* **10**, 2817–2830 (2016).
- Dominguez-Bello, M. G. et al. Partial restoration of the microbiota of cesarean-born infants via vaginal microbial transfer. *Nat. Med.* **22**, 250–253 (2016).
- Jakobsson, H. E. et al. Decreased gut microbiota diversity, delayed *Bacteroidetes* colonisation and reduced Th1 responses in infants delivered by caesarean section. *Gut* **63**, 559–566 (2014).
- Funkhouser, L. J. & Bordenstein, S. R. Mom knows best: the universality of maternal microbial transmission. *PLoS Biol.* **11**, e1001631 (2013).
- Nayfach, S., Rodriguez-Mueller, B., Garud, N. & Pollard, K. S. An integrated metagenomics pipeline for strain profiling reveals novel patterns of bacterial transmission and biogeography. *Genome Res.* **26**, 1612–1625 (2016).
- Ferretti, P. et al. Mother-to-infant microbial transmission from different body sites shapes the developing infant gut microbiome. *Cell Host Microbe* **24**, 133–145.e5 (2018).
- Yassour, M. et al. Strain-level analysis of mother-to-child bacterial transmission during the first few months of life. *Cell Host Microbe* **24**, 146–154.e4 (2018).
- Boucher, H. W. et al. Bad bugs, no drugs: no ESKAPE! An update from the Infectious Diseases Society of America. *Clin. Infect. Dis.* **48**, 1–12 (2009).
- Shin, H. et al. The first microbial environment of infants born by C-section: the operating room microbes. *Microbiome* **3**, 59 (2015).
- Brooks, B. et al. The developing premature infant gut microbiome is a major factor shaping the microbiome of neonatal intensive care unit rooms. *Microbiome* **6**, 112 (2018).
- Combellick, J. L. et al. Differences in the fecal microbiota of neonates born at home or in the hospital. *Sci. Rep.* **8**, 15660 (2018).

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

© The Author(s), under exclusive licence to Springer Nature Limited 2019

METHODS

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

Study population. The study was approved by the NHS London – City and East Research Ethics Committee (REC reference 12/LO/1492). 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, through the BBS (previously known as Life Study enhancement pilot study³²) from May 2014 to December 2017. Mothers provided written informed consent for their participation, and for the participation of their children, in the study. The study was performed in compliance with all relevant ethical regulations.

Sample collection. Faecal samples were collected from babies, with at least 1 sample in the first 21 days of life (referred to as neonatal period, primarily on day 4, 7 or 21). For a subset of babies who provided neonatal samples, a follow-up collection of a faecal sample was performed between 4 to 12 months of their lives (referred to as the infancy period). Maternal faecal samples were collected in the maternity unit before or after delivery, or stool was collected during delivery by midwives. Baby samples were collected at home by mothers, and returned to the processing laboratory by post at ambient temperature within 24 h. On arrival at the laboratory, all faecal samples were immediately stored at 4 °C for an average of 2.41 days (95% confidence interval 2.06–2.76 days) before further processing. Samples were aliquoted into 6 vials, 4 of which were stored at -80 °C for raw faeces biobanking; the other two vials were processed immediately for DNA extraction. Although this sample storage protocol (no preservation buffer for room temperature and 4 °C storage) was shown to be robust to technical variation in microbiome measurements at the time of study design (Supplementary Note 5), state-of-the-art preservation methods should be used in future large-scale microbiome studies to minimize the potential effect of sample storage on the microbiota composition³³. DNA was extracted from 30 mg of faecal samples as previously described³⁴. Negative controls using ultrapure water were included in parallel for each kit as well as each extraction batch, and DNA concentration was quantified to confirm it was contamination-free. Total DNA was eluted in 60 µl DNase/Pyrogen-free water, and stored at -80 °C until shipment to the Wellcome Sanger Institute for metagenomic sequencing.

Shotgun metagenomic sequencing and analysis. DNA samples, including negative controls, were quantified by PicoGreen dsDNA assay (Thermo Fisher), and samples with >100 ng DNA material proceeded to paired-end (2 × 125 bp) metagenomics sequencing on the HiSeq 2500 v4 platform. Low-quality bases were trimmed (SLIDINGWINDOW:4:20), and reads below 87 nucleotides in length (70% of original read length) were removed (MINLEN:87) using Trimmomatic³⁵. To remove potential human contaminants, quality-trimmed reads were screened against the human genome (GRCh38) with Bowtie2 v.2.3.0³⁶. On average, 22.4 (95% confidence interval 22.1–22.6) million raw reads were generated per sample. There were 19.3 (95% confidence interval 19.1–19.6) million reads (87.3% of the raw reads) per sample that passed decontamination and quality-trimming steps for downstream analysis. Sequencing depth was accounted for as a potential technical confounding factor in analyses of microbiota species and strain measurements, and significant species association with clinical covariates (Supplementary Note 6). Taxonomic classification from metagenomics reads was performed using Kraken v.1.0³⁷, a *k*-mer-based sequence classification approach against genomes of the Human Gastrointestinal Bacteria Genome Collection³⁸. Bracken v.1.0³⁹ was run on the Kraken classification output to estimate taxonomic abundance down to the species level. Metagenomic samples were compared at the genus and species levels by relative abundance. A cut-off of 100 Kraken-assigned paired-end reads (which corresponds to 0.001% relative abundance, given the sampling depth of about 10 million paired-end reads) was applied to determine metagenomic species detection. To assess whether the trade-off between the observed level of *Bacteroides* and opportunistic pathogens was an artefact of compositional effects, the proportion of abundances and reads that corresponded to *Bacteroides* were removed separately, before relative-abundance normalization. In the normalized datasets, the statistical enrichment of opportunistic pathogen species in babies delivered via caesarean section was consistent with the observation with the original data. The R packages phyloseq⁴⁰ and microbiome⁴¹ were used for metagenomic data analysis, and results were visualized using ggplot2⁴² in RStudio.

Classification of babies with the low-*Bacteroides* profile. For each baby, the median relative abundance of the *Bacteroides* genus was calculated across samples from the neonatal period. On the basis of the previously described threshold¹⁰, babies with a median abundance of less than 0.1% were assigned to the low-*Bacteroides* profile.

Classification of the opportunistic pathogen carriage. Total opportunistic pathogen load is estimated by calculating the combined, median relative abundance of the six differentially abundant opportunistic pathogen species (*C. perfringens*,

E. cloacae, *E. faecalis*, *E. faecium*, *K. oxytoca* and *K. pneumoniae*) per individual across their neonatal-period samples, and independently for the infancy-period and maternal samples. To prioritize relatively high-level opportunistic pathogen carriage that was feasible for downstream strain-cultivation experiments, individuals with a median abundance of over 1% total opportunistic pathogen load were defined as a positive carriage.

Analysis of transmission of maternal microbial strains. Strain transmissions in mother–baby paired samples were determined using a single-nucleotide-variant-calling method⁴³. StrainPhlAn was run on pre-processed metagenomes to generate consensus species-specific marker genes for phylogenetic reconstruction of all detectable strains (one dominant strain per sample), using default parameters and with the options ‘-alignment_program mafft’ and ‘-relaxed_parameters3’, as previously described²⁶. No significant variation in sequencing depth that had any effect on the coverage-dependent detection of microbiota species and strains was observed between babies delivered vaginally or by caesarean section, across age groups (Supplementary Note 6). For each species and strain with sufficient coverage for strain profiling, we generated a species-specific phylogenetic tree using RAxML⁴⁴. As previously described²⁶, the strain distance for each pair of mother–baby sample strains was computed by calculating the pairwise normalized phylogenetic distance on the corresponding species tree.

To define strain-transmission events, a previously described²⁶ conservative threshold of 0.1 on the strain distance value was used. The detectable strains in a given pair of mother–baby samples were considered to be identical (strain distance of less than 0.1, indicating transmission) or distinct (a strain distance of greater than 0.1, indicating no transmission). For all mother–baby pairs shown in Extended Data Fig. 4, an early transmission event was counted once per species per mother–baby pair, considering the detected transmission (or evidence for no transmission) at the earliest time point (primary transmission), irrespective of the subsequent transmission events in any later neonatal-period samples. For a subset of mother–baby pairs with both neonatal- and infancy-period sampled (Fig. 3a), late transmission events were counted separately, including cases of no early transmission owing to insufficient coverage (no detectable strains). To highlight the transmission pattern shared by phylogenetically related species, a neighbour-joining⁴⁵ tree of the eligible species was constructed on the basis of the mash distance matrix⁴⁶ of the respective reference genomes included in the StrainPhlAn database (Supplementary Table 4). The same approach and strain-distance threshold (core-genome single-nucleotide polymorphisms (SNPs)) were applied to the cultured strains to count the number of identical and distinct strains within mother–baby and longitudinal paired samples.

Statistical analysis. To calculate the effect of clinical covariates on the composition of the gut microbiota, we stratified by age groups and then assessed the proportion of explained variance (R^2 from PERMANOVA) in between-sample, species-level Bray–Curtis dissimilarity for each clinical covariate, using the adonis function from the R package vegan⁴⁷. PERMANOVA is mostly unaffected by group dispersion effects in balanced designs⁴⁸ (such as comparisons between mode of delivery); for unbalanced designs (such as breastfeeding comparisons) that are more sensitive to group dispersion effects, the group variance homogeneity condition was validated using the betadisper function. Group dispersions were not significantly different (betadisper $P < 0.05$) in all comparisons, which lent support to the significant—albeit visibly weak—effects of breastfeeding as reported by PERMANOVA. Samples with missing metadata for the given clinical covariate were excluded before running each cross-sectional analysis. Effect sizes and statistical significance were determined by 1,000 permutations, and P values were corrected for multiple testing using the Benjamini–Hochberg false-discovery rate (of 5%). MaAsLin⁴⁹ was used for the adjustment of covariates when determining the significance of species associated with a specific variable, while accounting for potentially confounding covariates, as previously described^{14,15}. All the covariates that were tested in the PERMANOVA were included in the adjustment, along with the sequencing depth used as a fixed effect. The default MaAsLin parameters were applied (maximum percentage of samples with missing metadata of 10%, minimum percentage relative abundance of 0.01%, $P < 0.05$, $q < 0.25$).

Bacterial isolation and whole-genome sequencing. Raw faecal samples from neonates, stored in the biobank laboratory at -80 °C, were requested on the basis of faecal carriage of targeted species over 1% relative abundance in metagenomes. Selected frozen faecal aliquots, where available (>100 ng), were couriered on dry ice to the Wellcome Sanger Institute within 6 h of shipment from the biobank laboratory. Bacterial isolates were cultured using the following culture media: *E. faecium* ChromoSelect Agar Base (Sigma-Aldrich) for *Enterococcus* spp., CP ChromoSelect Agar (Sigma-Aldrich) for *Clostridium* spp., Coliform ChromoSelect Agar (Sigma-Aldrich) and *Klebsiella* ChromoSelect Selective Agar (Sigma-Aldrich) for species of Enterobacteriaceae. Between two and five colonies per sample were picked for full-length 16S rRNA gene sequencing to confirm species identification, as previously described⁵⁰. Bacterial isolates with species identifications that were congruent with metagenomic identification were re-streaked and purified for genomic DNA

extraction using DNeasy 96 kit. DNA sequencing was performed on the Illumina HiSeq X, generating paired-end reads (2×151 bp). Multiple strains per species per faecal sample were also sequenced on the basis of variation across the full-length 16S rRNA sequences. Bacterial genomes were assembled and annotated using the previously described pipeline⁵¹. Genome assemblies were subjected to quality-checking and contaminant-screening with CheckM⁵² and Mash Screen⁵³, respectively. Where applicable, the suspected contaminant (non-target organism) sequences were confirmed, and filtered out via raw read-mapping using Bowtie2 v.2.3.0³⁶, before re-assembly.

Bacterial phylogenetic analysis. The phylogenetic analysis of the complete diverse species collection was conducted by extracting the amino acid sequences of 40 universal core marker genes^{54,55} from the BBS bacterial culture collection using Spec1⁵⁶. The protein sequences were concatenated and aligned with MAFFT v.7.2040, and maximum-likelihood trees were constructed using RAxML⁴⁴ with default settings. The four most prevalent opportunistic pathogen species (*E. faecalis*, *E. cloacae*, *K. oxytoca* and *K. pneumoniae*) in the BBS collection were further analysed in the context of the public genomes (Supplementary Table 5), including the UK-hospital strain collections^{57–60}, the gut microbiota-cultured strains from the Human Gastrointestinal Bacteria Genome Collection³⁸ and the Culturable Genome Reference⁶¹, and the environmental strains in the Genome Taxonomy Database (v.86)⁶². To generate phylogenetic trees of individual species, the public genome assemblies were combined with the assemblies of the study isolates, annotated with Prokka⁶³, and a pan-genome was estimated using Roary⁶⁴. In situations in which multiple identical strains (no difference in SNPs in the species core genome) were cultured from the same faecal sample, only one representative strain was included in the species phylogenetic trees. A 95% identity cut-off was used, and core genes were defined as those in 99% of isolates (unless stated otherwise). A maximum-likelihood tree of the SNPs in the core genes was created using RAxML⁴⁴ and 100 bootstraps. To illustrate the population structure of the closely related *Enterobacter* and *Klebsiella* strain isolates, FastANI⁶⁵ was used to estimate the pairwise average nucleotide-identity distance between all public and BBS genome assemblies, which was then used as an input to generate a neighbour-joining with BIONJ⁴⁵. All phylogenetic trees were visualized in iTOL⁶⁶. Sequence types were determined using MLSTcheck⁶⁷, which was used to compare the assembled genomes against the MLST database for the corresponding species.

Detecting virulence and resistance genes. ABRicate (v.0.8.13, <https://github.com/tseemann/abricate>) was used to screen for known, acquired resistance genes and virulence factors against bacterial genome assemblies. For genes related to AMR, a comprehensive BLAST database that integrates 5,556 non-redundant sequences in the NCBI Bacterial Antimicrobial Resistance Reference Gene Database (PRJNA313047), the Comprehensive Antibiotic Resistance Database (v.2.0.3)⁶⁸, ARG-ANNOT⁶⁹ and ResFinder⁷⁰ was queried. Three thousand two hundred and two non-redundant, experimentally validated core virulence genes in VFDB (version 5 October 2018)⁷¹ were included to build a BLAST database for virulence-factor screening.

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

Data availability

All sequencing data generated and analysed in this study have been deposited in the European Nucleotide Archive under accession numbers ERP115334 and ERP024601. The raw faecal samples and bacterial isolates are available from the corresponding authors upon request.

32. Field, N. et al. Infection and immunity from a lifecourse perspective: Life Study Enhancement. *The Lancet* **382**, S35 (2013).
33. Vandeputte, D., Tito, R. Y., Vanleeuwen, R., Falony, G. & Raes, J. Practical considerations for large-scale gut microbiome studies. *FEMS Microbiol. Rev.* **41**, S154–S167 (2017).
34. 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).
35. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
36. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
37. Wood, D. E. & Salzberg, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* **15**, R46 (2014).
38. Forster, S. C. et al. A human gut bacterial genome and culture collection for improved metagenomic analyses. *Nat. Biotechnol.* **37**, 186–192 (2019).
39. Lu, J., Breitwieser, F. P., Thielen, P. & Salzberg, S. L. Bracken: estimating species abundance in metagenomes data. *PeerJ Comput. Sci.* **3**, e104 (2017). <https://doi.org/10.7717/peerj-cs.104>.
40. McMurdie, P. J. & Holmes, S. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS ONE* **8**, e61217 (2013).

41. Lahti, L. & Shetty, S. Tools for microbiome analysis in R, version 1.1.10013 <https://github.com/microbiome/microbiome/> (2017).
42. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* (Springer, 2016).
43. Truong, D. T., Tett, A., Pasolli, E., Huttenhower, C. & Segata, N. Microbial strain-level population structure and genetic diversity from metagenomes. *Genome Res.* **27**, 626–638 (2017).
44. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
45. Gascuel, O. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Mol. Biol. Evol.* **14**, 685–695 (1997).
46. Ondov, B. D. et al. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.* **17**, 132 (2016).
47. Oksanen, J., Blanchet, F. G., Kindt, R. & Legendre, P. vegan: community ecology package, R package version 2.2–0 <https://cran.r-project.org/package=vegan> (2014).
48. Anderson, M. J. & Walsh, D. C. I. PERMANOVA, ANOSIM, and the Mantel test in the face of heterogeneous dispersions: what null hypothesis are you testing? *Ecol. Monogr.* **83**, 557–574 (2013).
49. Morgan, X. C. et al. Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome Biol.* **13**, R79 (2012).
50. Browne, H. P. et al. Culturing of ‘unculturable’ human microbiota reveals novel taxa and extensive sporulation. *Nature* **533**, 543–546 (2016).
51. Page, A. J. et al. Robust high-throughput prokaryote *de novo* assembly and improvement pipeline for Illumina data. *Microb. Genom.* **2**, e000083 (2016).
52. 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).
53. Ondov, B. D. et al. Mash Screen: high-throughput sequence containment estimation for genome discovery. Preprint at <https://www.biorxiv.org/content/10.1101/557314v1> (2019).
54. Sorek, R. et al. Genome-wide experimental determination of barriers to horizontal gene transfer. *Science* **318**, 1449–1452 (2007).
55. Ciccarelli, F. D. et al. Toward automatic reconstruction of a highly resolved tree of life. *Science* **311**, 1283–1287 (2006).
56. Mende, D. R., Sunagawa, S., Zeller, G. & Bork, P. Accurate and universal delineation of prokaryotic species. *Nat. Methods* **10**, 881–884 (2013).
57. Raven, K. E. et al. Genome-based characterization of hospital-adapted *Enterococcus faecalis* lineages. *Nat. Microbiol.* **1**, 15033 (2016).
58. Moradigaravand, D., Reuter, S., Martin, V., Peacock, S. J. & Parkhill, J. The dissemination of multidrug-resistant *Enterobacter cloacae* throughout the UK and Ireland. *Nat. Microbiol.* **1**, 16173 (2016).
59. Moradigaravand, D., Martin, V., Peacock, S. J. & Parkhill, J. Population structure of multidrug resistant *Klebsiella oxytoca* within hospitals across the UK and Ireland identifies sharing of virulence and resistance genes with *K. pneumoniae*. *Genome Biol. Evol.* **9**, 574–584 (2017).
60. Moradigaravand, D., Martin, V., Peacock, S. J., & Parkhill, J. Evolution and epidemiology of multidrug-resistant *Klebsiella pneumoniae* in the United Kingdom and Ireland. *MBio* **8**, e01976–e16 (2017).
61. Zou, Y. et al. 1,520 reference genomes from cultivated human gut bacteria enable functional microbiome analyses. *Nat. Biotechnol.* **37**, 179–185 (2019).
62. Parks, D. H. et al. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* **36**, 996–1004 (2018).
63. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).
64. Page, A. J. et al. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **31**, 3691–3693 (2015).
65. Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T. & Aluru, S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat. Commun.* **9**, 5114 (2018).
66. Letunic, I. & Bork, P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res.* **44**, W242–W245 (2016).
67. Page, A. J., Taylor, B. & Keane, J. A. Multilocus sequence typing by blast from *de novo* assemblies against PubMLST. *J. Open Source Software* **8**, 118 (2016).
68. Jia, B. et al. CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. *Nucleic Acids Res.* **45**, D566–D573 (2017).
69. Gupta, S. K. et al. ARG-ANNOT, a new bioinformatic tool to discover antibiotic resistance genes in bacterial genomes. *Antimicrob. Agents Chemother.* **58**, 212–220 (2014).
70. Zankari, E. et al. Identification of acquired antimicrobial resistance genes. *J. Antimicrob. Chemother.* **67**, 2640–2644 (2012).
71. Chen, L. et al. VFDB: a reference database for bacterial virulence factors. *Nucleic Acids Res.* **33**, D325–D328 (2016).

Acknowledgements This work was supported by the Wellcome Trust (WT101169MA) and Wellcome Sanger core funding (WT098051). Y.S. is supported by a Wellcome Trust PhD Studentship. S.C.F. is supported by the Australian National Health and Medical Research Council (1091097, 1159239 and 1141564) and the Victorian Government’s Operational Infrastructure Support Program. We thank the participating families for their time and contribution to the BBS; the research midwives at recruiting hospitals for recruitment and clinical metadata collection; N. Moreno, H. Ali, S. Bibi and A. Takyi for raw-sample processing; the Core Sequencing and Pathogen Informatics teams at the Wellcome Sanger Institute for informatics support; and H. Browne and A. Almeida for critical feedback of the manuscript.

Author contributions S.C.F., A.R., P.B., N.F. and T.D.L. conceived and designed the project. S.C.F., E.T., N.K. and M.D.S. carried out the pilot study, and designed sample collection and processing protocols, overseen by N.F. and T.D.L. E.T., A.S., N.S. and N.F. managed participant recruitment and coordinated clinical metadata collection; Y.S. performed bacterial culturing and DNA extraction with assistance from M.D.S. Y.S. generated and analysed the data with assistance from K.V. Y.S., S.C.F., N.F. and T.D.L. wrote the manuscript. All authors read and approved the manuscript.

Competing interests The authors declare no competing interests.

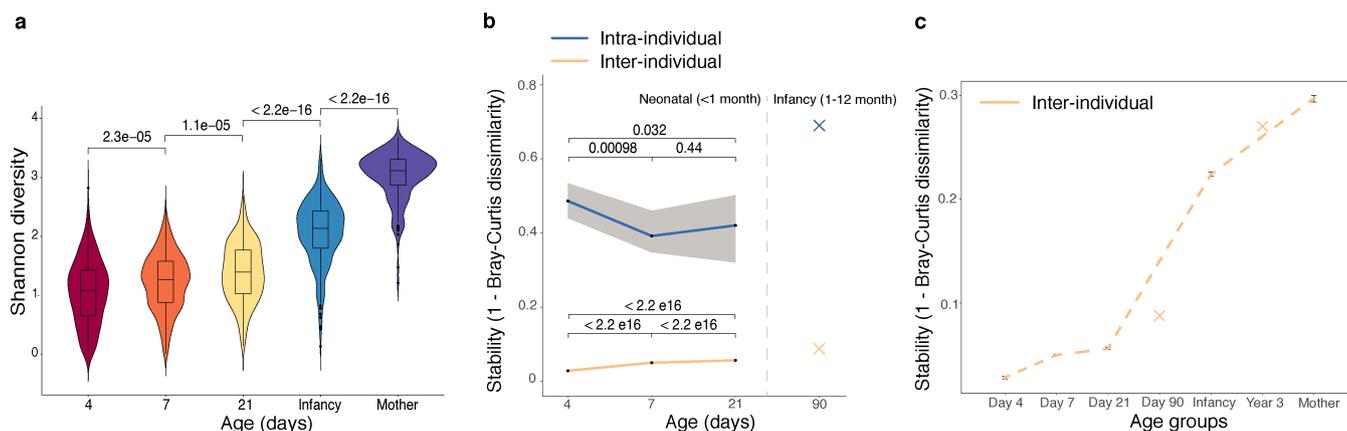
Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-019-1560-1>.

Correspondence and requests for materials should be addressed to N.F. or T.D.L.

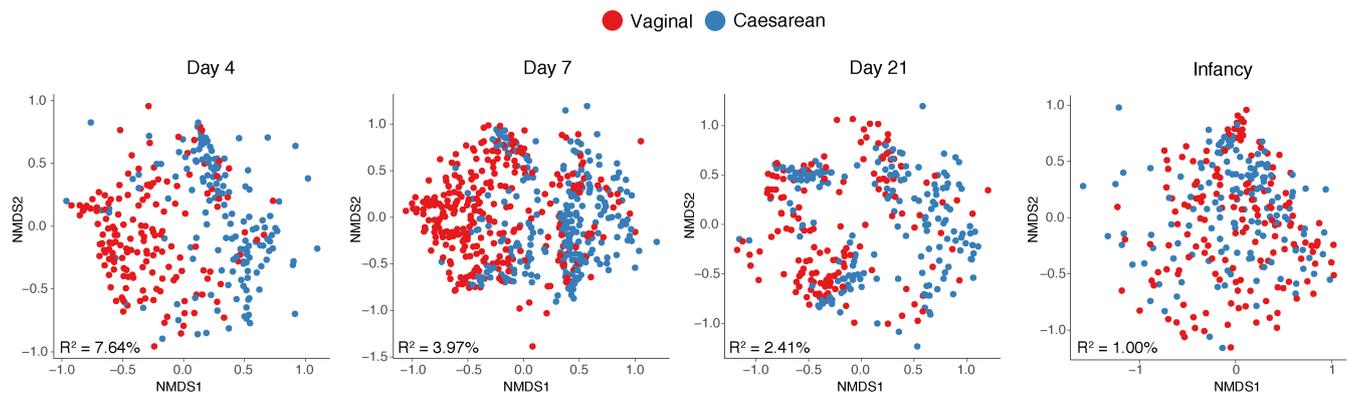
Peer review information *Nature* thanks Serena Manara, Xavier Ramnik, Nicola Segata and Paul Wilmes for their contribution to the peer review of this work.

Reprints and permissions information is available at <http://www.nature.com/reprints>.



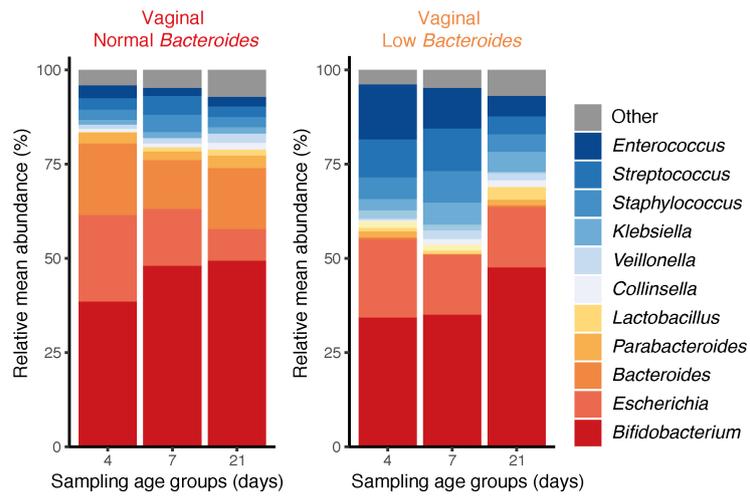
Extended Data Fig. 1 | Neonatal gut microbiota exhibits high volatility and individuality. **a**, Microbiota alpha diversity (Shannon diversity index) increased over developmental time. The violin plot outlines the kernel probability density; the width of the shaded area represents the proportion of the data shown. Centre line shows median; box limits indicate the 25th and 75th percentiles; whiskers extend $1.5\times$ the interquartile range from the 25th and 75th percentiles; and outliers are represented by dots. The gut microbiotas of babies on day 4 ($n = 310$ individuals), 7 ($n = 532$ individuals) and 21 ($n = 325$ individuals), and in infancy ($n = 302$ individuals), as well as from matched mothers ($n = 175$), are plotted. **b**, **c**, Stability of the gut microbiota, stratified by inter-

individual (day 4, $n = 310$ individuals; day 7, $n = 532$ individuals; and day 21, $n = 325$ individuals) and intra-individual comparisons in sliding time windows (day 4 to 7, $n = 274$ individuals and day 7 to 21, $n = 285$ individuals) during the neonatal period (**b**), in the context of the overall infancy period (**c**). Microbiota stability measurements from the TEDDY¹⁵ study (the earliest measurements on day 90, and at year 3) are plotted in crosses. Solid lines show the median per time window. Shaded areas show the 99% confidence interval, estimated using binomial distribution. Error bars indicate median absolute deviation. The significance of the difference between groups was determined by two-sided Wilcoxon rank-sum test.



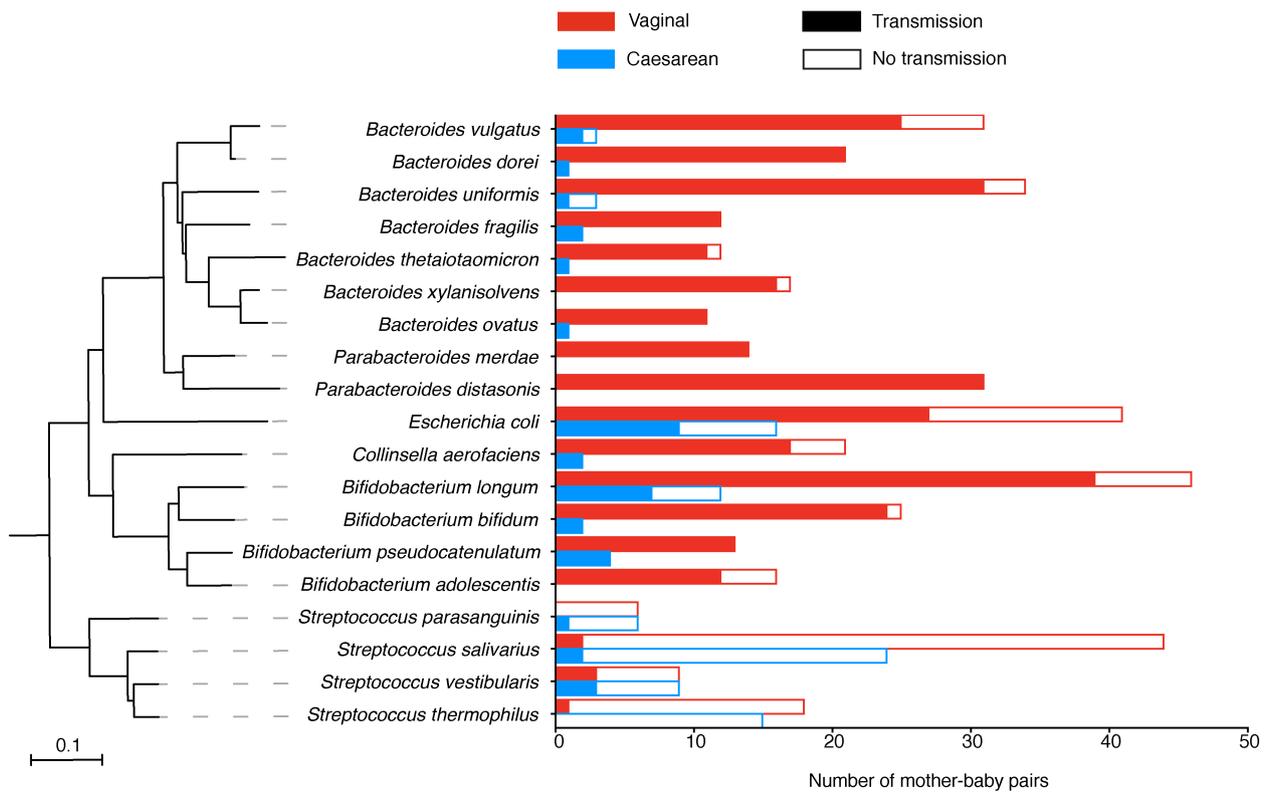
Extended Data Fig. 2 | Microbiota variation associated with mode of delivery in the neonatal period and infancy. Non-metric multidimensional scaling ordination of Bray–Curtis dissimilarity between the species relative-abundance profiles of the gut microbiota sampled from babies on day 4 (vaginally delivered, $n = 157$ babies; delivered by caesarean section, $n = 153$ babies), day 7 (vaginally delivered, $n = 280$ babies; delivered by caesarean section, $n = 252$ babies), day 21

(vaginally delivered, $n = 147$ babies; delivered by caesarean section, $n = 178$ babies) and during infancy (vaginally delivered, $n = 160$ babies; delivered by caesarean section, $n = 142$ babies). The microbial variation explained by the mode of delivery is represented by the PERMANOVA R^2 value (bottom left), and is significant across four cross-sectional PERMANOVA tests. False-discovery-rate-corrected P values are reported in Supplementary Table 2.



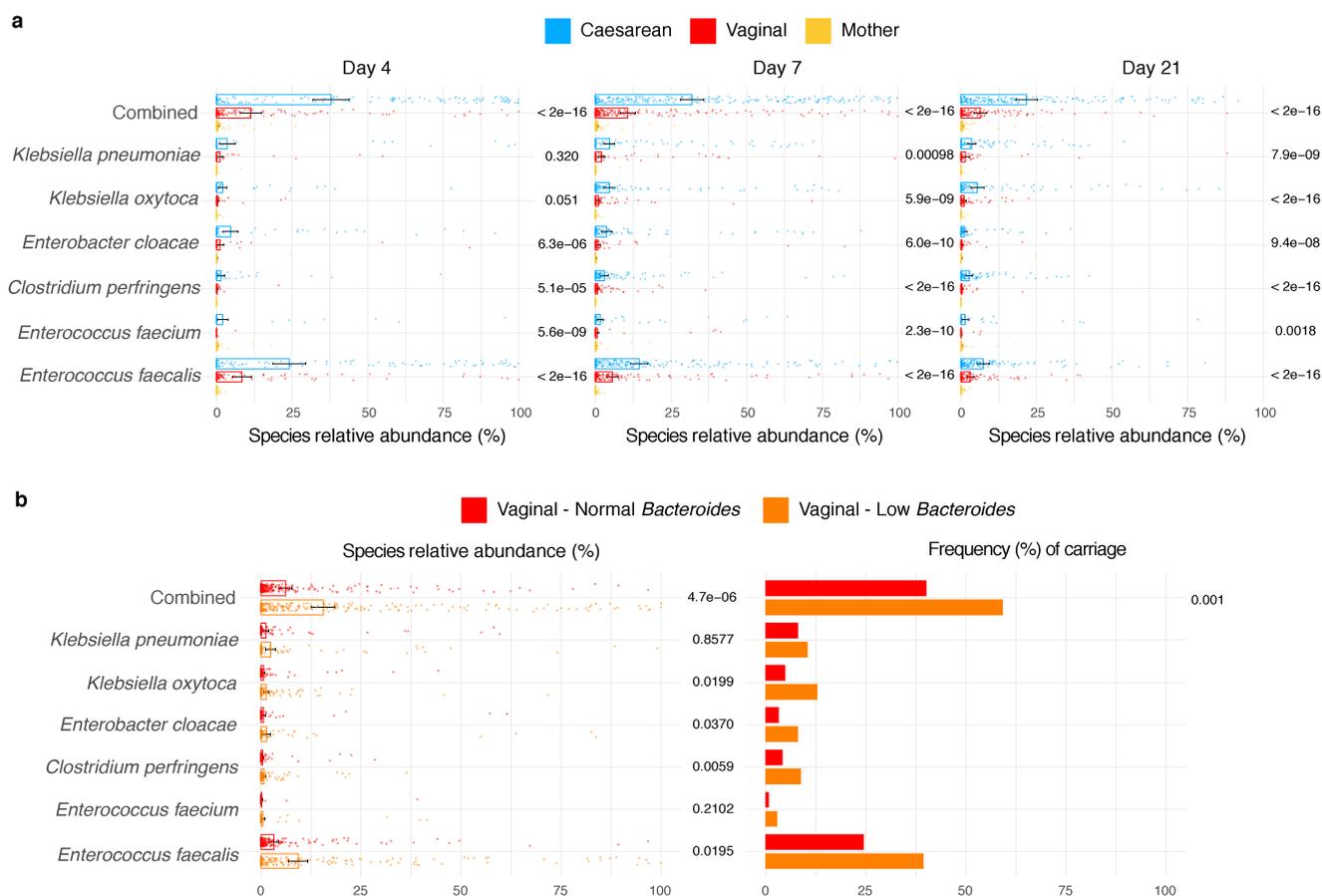
Extended Data Fig. 3 | Microbial succession in the neonatal gut microbiota of vaginally delivered babies. Bar plots show longitudinal changes in the mean relative abundance of faecal bacteria at the genus level on day 4, day 7 and day 21, for genera with >1% mean relative abundance

across all neonatal samples. Left, $n = 316$ samples from 160 vaginally delivered babies detected with *Bacteroides*. Right, $n = 290$ samples from 154 vaginally delivered babies with the low-*Bacteroides* profile (defined in 'Classification of babies with the low-*Bacteroides* profile' in Methods).



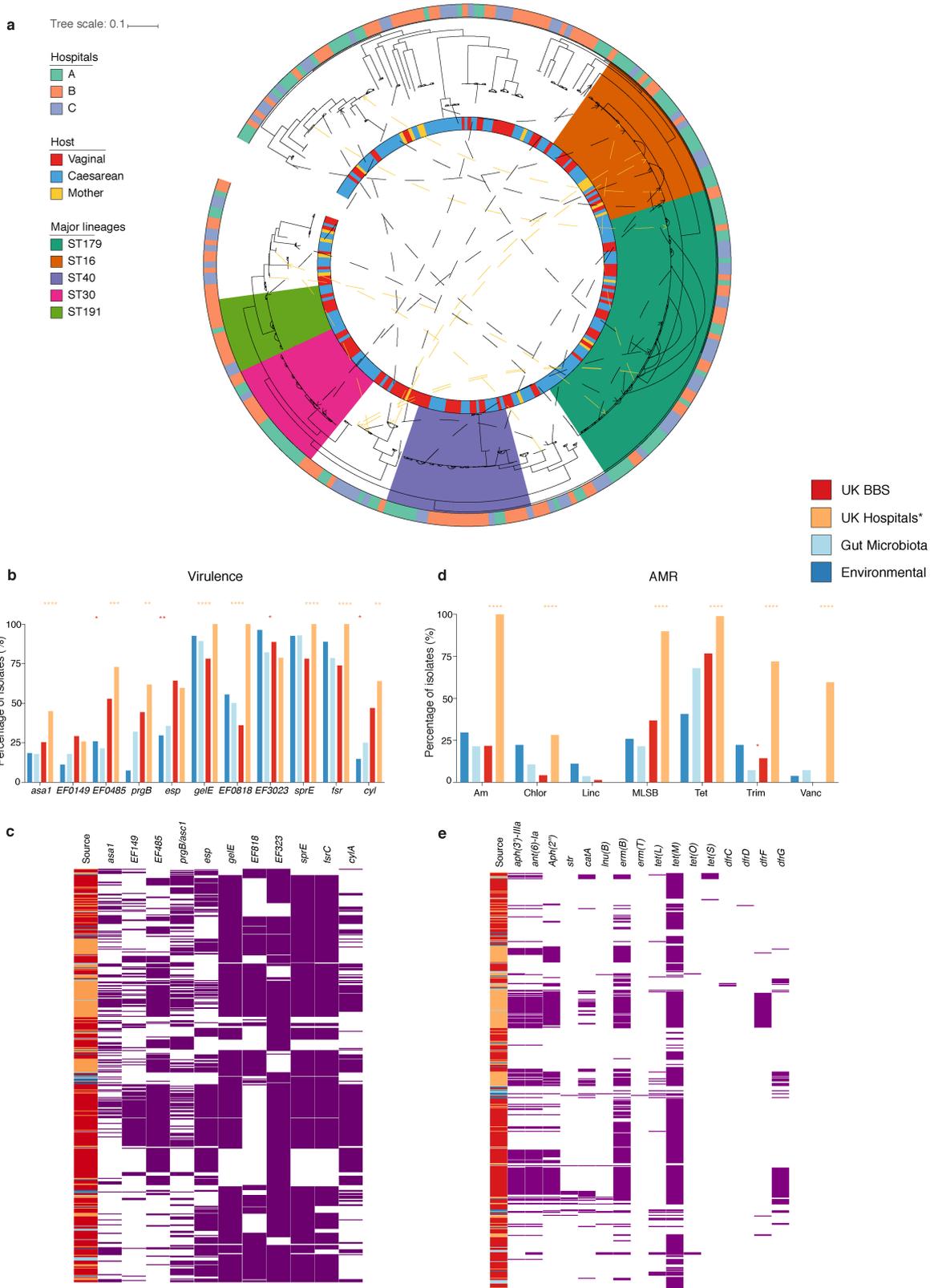
Extended Data Fig. 4 | Transmission of maternal microbial strains during the early neonatal period. Transmissions of maternal microbial strains across 178 mother–baby pairs (for 112 vaginally delivered babies, and 66 babies delivered by caesarean section) who sampled at least once during the early neonatal period. Only the frequently shared species that were detected with sufficient coverage for strain analysis in more than ten pairs are shown. The neighbour-joining tree is constructed on the

basis of the pairwise mash distances of the respective reference genomes. Phylogenetically related species shared a similar pattern for the timing of transmissions; for example, the frequent transmission of *Bacteroides* spp., *Parabacteroides* spp. and *Bifidobacterium* spp. in vaginally delivered babies and the lack of species of these genera in babies delivered by caesarean section, and the fact that most *Streptococcus* species were transmitted from environmental sources other than the maternal gut microbiota.



Extended Data Fig. 5 | Frequency and abundance of opportunistic pathogens in the neonatal gut microbiota. a, b, Babies delivered by caesarean section, and vaginally delivered babies with the low-*Bacteroides* profile, more frequently carried opportunistic pathogens (as defined in ‘Classification of the opportunistic pathogen carriage’ in Methods) and at higher level of species relative abundance during the first 21 days of life, as compared to vaginally delivered babies (a) and vaginally delivered babies with the normal *Bacteroides* profile (b), respectively. There was a significantly different presence in the neonatal samples within each major neonatal-period sampling group (day 4 ($n = 310$ individuals), day 7 ($n = 532$ individuals) and day 21 ($n = 325$ individuals))—in terms of mean relative abundance and frequency—of six known opportunistic pathogens

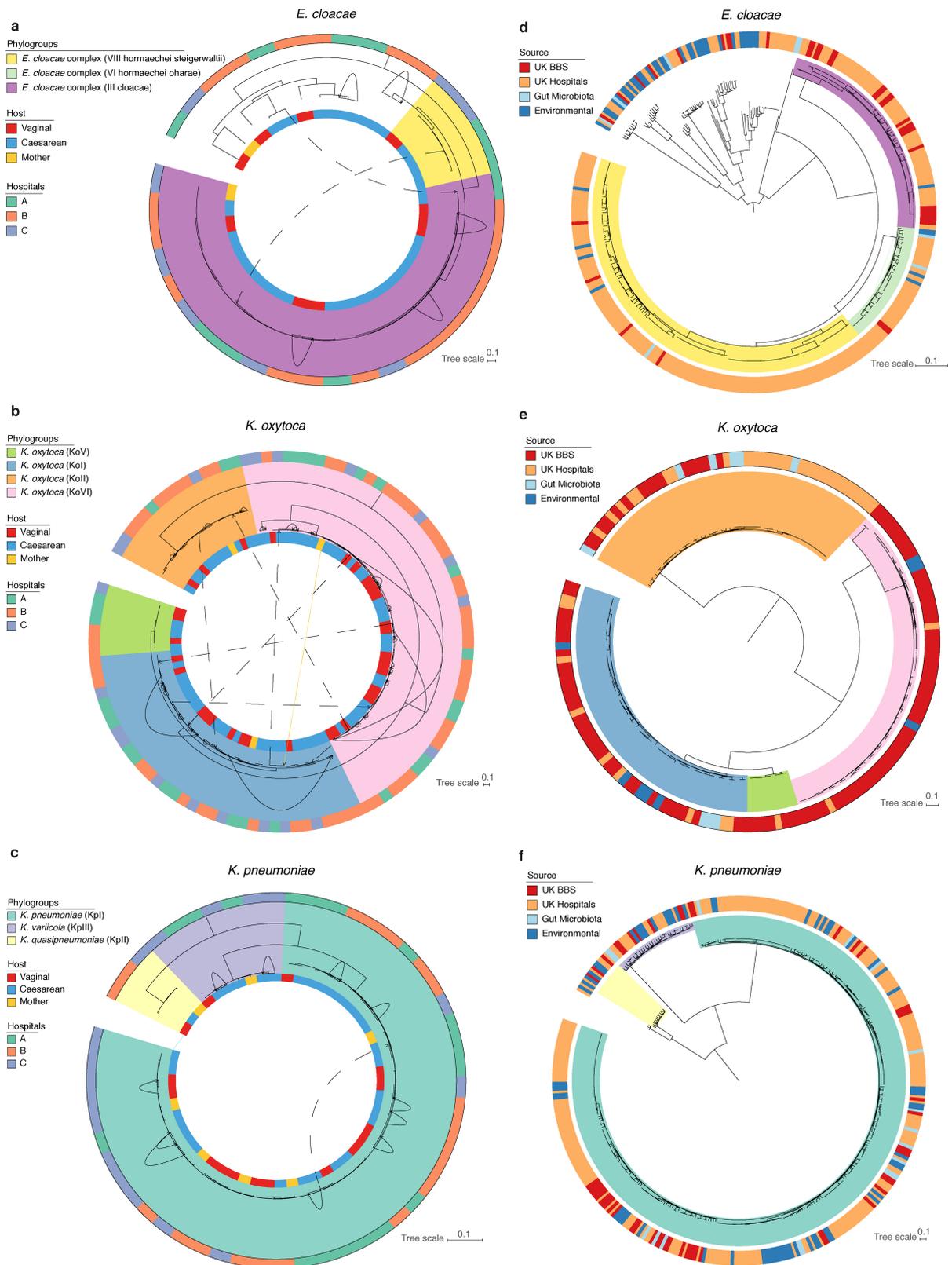
that are associated with the hospital environment, and rarely carried by adults ($n = 175$ mothers) (b). The numbers of individuals sampled in the neonatal period were 314 (vaginally delivered), 160 (vaginally delivered, and with a normal level of *Bacteroides*) and 154 (vaginally delivered, with the low-*Bacteroides* profile). Error bars indicate the 95% confidence interval of the mean relative abundance. The significance (P values indicated to the right of the bars) of the difference in mean species relative abundance and combined-pathogen carriage (defined in ‘Classification of the opportunistic pathogen carriage’ in Methods) frequency was obtained by applying two-sided Wilcoxon signed-rank test and Fisher’s exact test, respectively.



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

Extended Data Fig. 6 | Phylogeny and pathogenicity potential of *E. faecalis* strains of the BBS. **a**, Phylogenetic tree of *E. faecalis* strains of the BBS ($n = 282$ strains, isolated from 269 faecal samples of 160 subjects). The midpoint-rooted maximum-likelihood phylogeny is based on SNPs in 1,827 core genes. Five major lineages (>10 representatives in strains of the BBS; ST179, $n = 60$; ST16, $n = 30$; ST40, $n = 27$; ST30, $n = 21$; and ST191, $n = 14$) were identified within UK hospital collections, distributed across three hospitals in this study and with no phylogroup limited to any single hospital. Solid lines between strains indicate intra-subject strain persistence ($n = 92$ strains in 67 babies). Dashed lines indicate phylogenetically distinct strains that were isolated from longitudinal samples ($n = 18$) or mother–baby paired samples (yellow, $n = 10$); arrows indicate the direction of the potential transmission (early-to-later or mother-to-baby). In situations in which multiple identical strains (no SNP difference in species core genome) were isolated from the same faecal sample, only one representative strain was included in the species phylogenetic tree (total number of strains, $n = 356$). **b–e**, Prevalence of virulence-related genes (**b**, **c**) and AMR-related genes (grouped

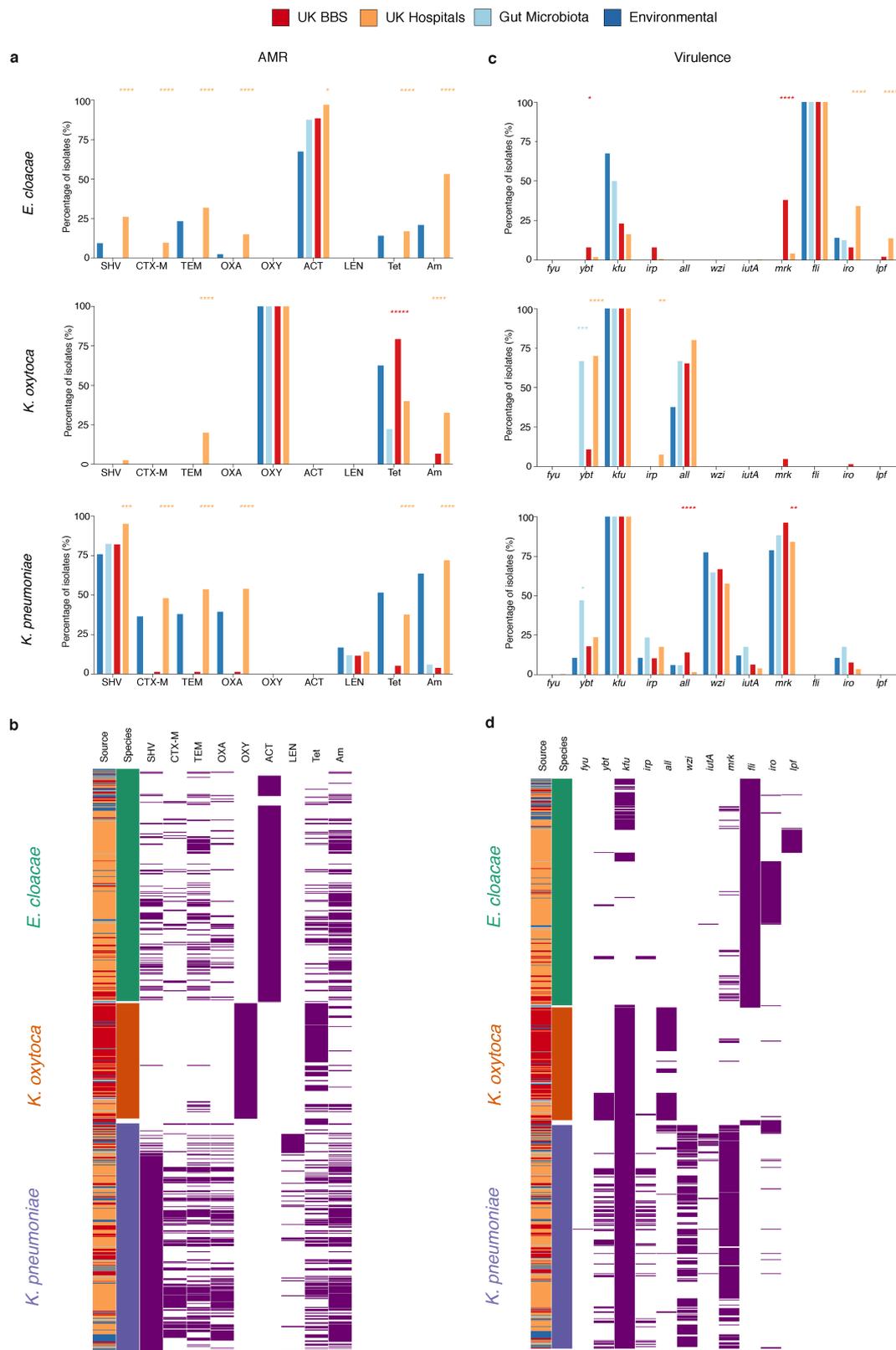
by antibiotic class) (**d**, **e**) detected in *E. faecalis* strains of the BBS. Significance results shown are coloured according to the group with higher frequency of detected genes, by two-sided Fisher's exact test between the groups of the public gut microbiota strains ($n = 28$) versus strains of the BBS ($n = 356$), and strains of the BBS versus the epidemic strains in UK hospitals ($n = 89$; tree branches coloured blue in Fig. 4c). **** $P < 0.0001$, *** $P < 0.001$, ** $P < 0.01$, * $P < 0.05$. Virulence-related genes: *asa1*, *EF0149*, *EF0485* and *prgB*, aggregation substance; *esp*, enterococcal surface protein; genes that encode exoenzymes: *gelE*, gelatinase; *EF0818* and *EF3023*, hyaluronidase (spreading factor); *sprE*, serine protease; and *fsr*, quorum sensing system; toxin-encoding gene: *cyl*, cytolysin. Genes detected across all isolates (*dfrE*, *efrA*, *efrB*, *emeA* and *lsaA*) are not shown. AMR-related genes: Am, aminoglycosides (*aph3''-III*, *ant(6)-Ia*, *aph(2'')* and *str*); Chlor, chloramphenicol (*catA*); Linc, lincosamides (*lnuB*); MLSB, macrolide, lincosamide and streptogramin B (*ermB* or *ermT*); Tet, tetracycline (*tetL*, *tetM*, *tetO* and *tetS*); Trim, trimethoprim (*dfrC*, *dfrD*, *dfrF* or *dfrG*); and Vanc, vancomycin.



Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | Phylogenies of *E. cloacae*, *K. oxytoca* and *K. pneumoniae* strains. **a–f**, Midpoint-rooted core-genome maximum-likelihood trees of *E. cloacae* complex, *K. oxytoca* and *K. pneumoniae* strains isolated in this study (**a–c**) and in the context of public genomes (**d–f**). **a–c**, Number of strains of *E. cloacae* (**a**) ($n = 37$, isolated from 37 faecal samples of 30 subjects, 1,861 core genes), *K. oxytoca* (**b**) ($n = 107$, isolated from 90 faecal samples of 62 subjects, 2,910 core genes) and *K. pneumoniae* strains (**c**) ($n = 53$, isolated from 47 faecal samples of 35 subjects, 3,471 core genes) of the BBS. Solid lines between strains indicate the intra-subject strain persistence (*E. cloacae*, $n = 5$ strains in 5 babies; *K. oxytoca*, $n = 25$ strains in 18 babies; and *K. pneumoniae*, $n = 11$ strains in 8 babies). Dashed lines indicate phylogenetically distinct strains isolated from longitudinal samples (*E. cloacae*, $n = 2$ strains in 2 individuals; *K. oxytoca*, $n = 7$ strains in 6 subjects; and *K. pneumoniae*, $n = 1$ strain in 1 individual); arrows indicate the direction of potential transmission (early-to-later samples). In situations in which multiple

identical strains (no difference in SNPs in species core genome) were isolated from the same faecal sample, only one representative strain was included in the species phylogenetic tree (number of non-redundant BBS strains: *E. cloacae*, $n = 52$; *K. oxytoca*, $n = 150$; *K. pneumoniae*, $n = 78$). For each species, the main phylogroups identified with UK hospital collections are shown (*E. cloacae*, III and VIII; *K. oxytoca*, KoI, KoII, KoV and KoVI; *K. pneumoniae*, KpI, KpII and KpIII); these were distributed across three hospitals in this study, with no phylogroup limited to any single hospital. **d–f**, The number of public genomes included in the phylogenetic analysis of *E. cloacae* (**d**) (UK hospitals, $n = 314$; gut microbiota, $n = 8$; environmental sources, $n = 43$; 1,484 core genes), *K. oxytoca* (**e**) (UK hospitals, $n = 40$; gut microbiota, $n = 9$; environmental sources, $n = 8$; 3,399 core genes) and *K. pneumoniae* strains (**f**) (UK hospitals, $n = 250$; gut microbiota, $n = 17$; environmental sources, $n = 66$; 2,510 core genes).



Extended Data Fig. 8 | See next page for caption.

Extended Data Fig. 8 | Carriage of AMR and virulence genes in *Klebsiella* and *Enterobacter* strains. **a–d**, Frequency and heat maps of isolates for putative AMR-related (**a**, **b**) and virulence-related genes (grouped by antibiotic class) (**c**, **d**) that are most-frequently detected in strains of the UK hospital collection of *E. cloacae* (green), *K. oxytoca* (orange) and *K. pneumoniae* (blue). Significance results shown are coloured according to the group with higher frequency of detected genes, by two-sided Fisher's exact test between the groups of the public gut microbiota strains (*E. cloacae*, $n = 8$; *K. oxytoca*, $n = 9$; and *K. pneumoniae*, $n = 17$) versus strains in the BBS (*E. cloacae*, $n = 52$; *K. oxytoca*, $n = 150$; and *K. pneumoniae*, $n = 78$), and strains in the BBS versus strains in UK hospitals (*E. cloacae*, $n = 314$; *K. oxytoca*, $n = 40$;

K. pneumoniae, $n = 250$). **** $P < 0.0001$, *** $P < 0.001$, ** $P < 0.01$, * $P < 0.05$. AMR-related genes: extended-spectrum β -lactamases, SHV (*bla_{SHV}*), CTX-M (*bla_{CTX-M}*) and TEM (*bla_{TEM}*); other β -lactamases, OXA (*bla_{OXA}*), OXY (*bla_{OXY}*), ACT (*bla_{ACT}*) and LEN (*bla_{LEN}*); Tet, tetracycline (*tetA* and *tetR*); Am, aminoglycosides (*aac(3)*, *aac(6')*, *aad* and *str*). Virulence-related genes: iron acquisition, *fyu*; yersiniabactin, *ybt*; iron transporter permease, *kfu*; iron regulatory proteins, *irp*; allantoin metabolism, *all*; capsule, *wzi*; aerobactin siderophore receptor, *iutA*; fimbriae and biofilm formation, *mrk*; flagella biosynthesis, *fli*; siderophore production, *iro*; and fimbrial chaperones, *lpf*. Genes detected across all isolates are not shown.

Extended Data Table 1 | The main clinical characteristics of the BBS cohort

Sampling age group (subjects)	
All babies	596
Day 4	310
Day 7	532
Day 21	325
Infancy	302
Mother	175
Mode of delivery – Caesarean section	
All babies	282 (47.3%)
Day 4	153 (49.4%)
Day 7	252 (47.4%)
Day 21	178 (54.8%)
Infancy	142 (47.0%)
Mother	65 (37.1%)
Mode of feeding – Non-breastfeeding	
Day 4	52 (17.5%)
Day 7	83 (16.8%)
Day 21	59 (19.0%)
Infancy	33 (13.4%)
Antibiotics	
Intrapartum antibiotic prophylaxis (vaginal delivery)	23 (7.3%)

The complete clinical metadata of the study participants are reported in Supplementary Table 1.

Reporting Summary

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

Statistics

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

n/a Confirmed

- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever 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

Bowtie2 v2.3.0, vegan_2.5-5, phyloseq_1.28.0, microbiome_1.6.0, ggplot2_3.1.1, ggpubr_0.1.8, Maaslin_0.0.5, R v3.6.0, Trimmomatic v0.35, Kraken v1.0, Bracken v1.0, MetaPhlAn v2.6, StrainPhlAn v1.0, CheckM v1.0.12, FastANI v1.1, Roary 1.007001, Prokka v1.13, RAxML v8.0.0, BIONJ, MAFFT v7.2040, FastTree v2.1.10, ABRicate v0.8.13, MLSTcheck v2.1.1630910, Mash v2.1, iTOL v4

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/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

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

All sequencing data generated and analysed in this study have been deposited in the European Nucleotide Archive under accession numbers ERP115334 and ERP024601. Raw faecal samples and bacterial isolates are available from the corresponding authors upon request.

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](https://www.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	No sample size calculation was employed as this is an observational study. Participants were recruited on a voluntary basis in hospitals. This study presents the largest gut metagenome dataset of newborn babies (596 individuals, 1202 samples)
Data exclusions	None. All sequencing data that passed the pre-established sequencing quality control (described in Methods) were included for analysis.
Replication	No replication was employed as this is an observational study.
Randomization	Randomization was not employed as this is an observational study. Participants were recruited on a voluntary basis in hospitals.
Blinding	No blinding used as this is an observational study.

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
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input type="checkbox"/>	<input checked="" type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data

Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

Human research participants

Policy information about [studies involving human research participants](#)

Population characteristics	Study population are 771 mothers (mean 33.82±4.99 years of age) and their term newborn babies (49.4% female, 50.6% male) born in three NHS hospitals in the UK (London and Leicester). Babies were sampled in their neonatal period at 4 to 21 days of age, and further sampled in later infancy at 4 to 12 months of age.
Recruitment	Participants were recruited and consented before or after birth whilst on the labour ward, by their attending midwives. All pregnant women at the labour ward to give birth were eligible to participate, except those under 16 years of age and non-UK residents intending to return abroad immediately after delivery.
Ethics oversight	The study was approved by the NHS London - City and East Research Ethics Committee (REC reference 12/LO/1492).

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