

Stunted microbiota and opportunistic pathogen colonization in caesarean-section birth

Shao, Yan; Forster, Samuel C; Tsaliki, Evdokia; Vervier, Kevin; Strang, Angela; Simpson, Nandi; Kumar, Nitin; Stares, Mark D; Rodger, Alison; Brocklehurst, Peter; Field, Nigel; Lawley, Trevor D

DOI:

[10.1038/s41586-019-1560-1](https://doi.org/10.1038/s41586-019-1560-1)

License:

None: All rights reserved

Document Version

Peer reviewed version

Citation for published version (Harvard):

Shao, Y, Forster, SC, Tsaliki, E, Vervier, K, Strang, A, Simpson, N, Kumar, N, Stares, MD, Rodger, A, Brocklehurst, P, Field, N & Lawley, TD 2019, 'Stunted microbiota and opportunistic pathogen colonization in caesarean-section birth', *Nature*, vol. 574, no. 7776, pp. 117-121. <https://doi.org/10.1038/s41586-019-1560-1>

[Link to publication on Research at Birmingham portal](#)

General rights

Unless a licence is specified above, all rights (including copyright and moral rights) in this document are retained by the authors and/or the copyright holders. The express permission of the copyright holder must be obtained for any use of this material other than for purposes permitted by law.

- Users may freely distribute the URL that is used to identify this publication.
- Users may download and/or print one copy of the publication from the University of Birmingham research portal for the purpose of private study or non-commercial research.
- User may use extracts from the document in line with the concept of 'fair dealing' under the Copyright, Designs and Patents Act 1988 (?)
- Users may not further distribute the material nor use it for the purposes of commercial gain.

Where a licence is displayed above, please note the terms and conditions of the licence govern your use of this document.

When citing, please reference the published version.

Take down policy

While the University of Birmingham exercises care and attention in making items available there are rare occasions when an item has been uploaded in error or has been deemed to be commercially or otherwise sensitive.

If you believe that this is the case for this document, please contact UBIRA@lists.bham.ac.uk providing details and we will remove access to the work immediately and investigate.

1 Stunted microbiota and opportunistic pathogen colonisation associated with 2 C-section birth

3 Yan Shao¹, Samuel C. Forster^{1,2,3}, Evdokia Tsaliki⁴, Kevin Vervier¹, Angela Strang⁴, Nandi Simpson⁴,
4 Nitin Kumar¹, Mark D. Stares¹, Alison Rodger⁴, Peter Brocklehurst⁵, Nigel Field⁴ §,
5 Trevor D. Lawley^{1,§}

6 ¹Host-Microbiota Interactions Laboratory, Wellcome Sanger Institute, Hinxton, United Kingdom

7 ²Centre for Innate Immunity and Infectious Diseases, Hudson Institute of Medical Research, Clayton, Victoria, Australia

8 ³Department of Molecular and Translational Sciences, Monash University, Clayton, Victoria, Australia

9 ⁴Institute for Global Health, University College London, London, United Kingdom

10 ⁵Birmingham Clinical Trials Unit, University of Birmingham, Birmingham, United Kingdom

11

12 §corresponding authors

13 Trevor D. Lawley: Wellcome Sanger Institute, Hinxton, United Kingdom, CB10 1SA, Phone 01223 495 391, Fax
14 01223 495 239, Email: tl2@sanger.ac.uk

15 Nigel Field: Institute for Global Health, University College London, London, United Kingdom, WC1N 1EH,
16 Email: nigel.field@ucl.ac.uk

17

18 Running title: Caesarean section predisposes neonates to healthcare-associated opportunistic pathogen colonisation

19 Keywords: gastrointestinal microbiome, early-life microbiota colonisation, clinical metagenomics, neonatal, c-
20 section, intrapartum antibiotic prophylaxis, paediatric, opportunistic pathogens, antimicrobial resistance (AMR),
21 Enterococcus, Klebsiella

22 **Abstract**

23 Immediately after birth, newborn babies experience rapid colonisation by microorganisms from their
24 mothers and the surrounding environment¹. Diseases in childhood and later in life are potentially mediated
25 through perturbation of the infant gut microbiota colonisations². However, the impact of modern clinical
26 practices, such as caesarean section delivery and antibiotic usage, on the earliest stages of gut microbiota
27 acquisition and development during the neonatal period (≤ 1 month) remains controversial^{3,4}. Here we report
28 disrupted maternal transmission of *Bacteroides* strains and high-level colonisation by healthcare-associated
29 opportunistic pathogens, including *Enterococcus*, *Enterobacter* and *Klebsiella* species, in babies delivered
30 by caesarean section (C-section), and to a lesser extent, in those delivered vaginally with maternal antibiotic
31 prophylaxis or not breastfed during the neonatal period. Applying longitudinal sampling and whole-genome
32 shotgun metagenomic analysis on 1,679 gut microbiotas of 772 full term, UK-hospital born babies and
33 mothers, we demonstrate that the mode of delivery is a significant factor impacting gut microbiota
34 composition during the neonatal period that persists into infancy (1 month - 1 year). Matched large-scale
35 culturing and whole-genome sequencing (WGS) of over 800 bacterial strains cultured from these babies
36 identified virulence factors and clinically relevant antimicrobial resistance (AMR) in opportunistic
37 pathogens that may predispose to opportunistic infections. Our findings highlight the critical early roles of
38 the local environment (i.e. mother and hospital) in establishing the gut microbiota in very early life, and
39 identifies colonisation with AMR carrying, healthcare-associated opportunistic pathogens as a previously
40 unappreciated risk factor.

41 **Main**

42 The acquisition and development of the early-life gut microbiota follow successive waves of
43 microbial exposures and colonisation that shapes the longer-term microbiota composition and function⁵.
44 Early life events, including Caesarean section delivery^{1,6}, formula feeding^{7,8} and antibiotic exposure^{8,9} that
45 could perturb the gut microbiota composition are associated with the development of childhood asthma and
46 atopy¹⁰⁻¹². While recent studies^{8,9,13-15} have provided substantial insights into the gut microbiota
47 development during the first 3 years of life, many were limited by the taxonomic resolution provided by
48 16S rRNA gene profiling, small sample size or limited sampling during the first month of life (neonatal
49 period). High-resolution metagenomic studies of large, longitudinal cohorts are required to establish the
50 impact and risks of early life events on the gut microbiota assembly, particularly during the neonatal period
51 where pioneering microbes could influence subsequent microbiota and immune system development^{16,17}.

52 To characterise the trajectory of gut microbiota acquisition and development during the neonatal
53 period, we enrolled 596 healthy, term babies (39.5 ± 1.37 gestation weeks, 314 vaginal and 282 C-section
54 births, Fig. 1a, Extended Data Table 1) through the Baby Biome Study (BBS). Faecal samples were
55 collected from all babies at least once during their neonatal period (<1 month) with 302 babies re-sampled
56 later in infancy (8.75 ± 1.98 months). Maternal faecal samples were also obtained from 175 mothers paired
57 with 178 babies. Metagenomic analysis of 1,679 faecal samples from 772 babies and mothers revealed
58 temporal dynamics of the gut microbiota development (Fig. 1b) and increased diversity with age (Extended
59 Data Fig. 1a). Strikingly, the gut microbiotas exhibited substantial heterogeneity (inter-individual) and
60 instability (intra-individual) during the first weeks of life (Extended Data Fig. 1b). Inter-individual
61 differences explained 57% of the microbial taxonomic variation (Permutational multivariate analysis of
62 variance (PERMANOVA), $P < 0.001$, 1,000 permutations), followed by sampling age at 5.7% of the
63 variance ($P < 0.001$). These results indicate that the gut microbiotas were highly dynamic and
64 individualised during the neonatal period, even more than observed in infancy (Extended Data Fig. 1c).

65 To determine the impact of clinical covariates on the composition of the gut microbial community,
66 we performed cross-sectional PERMANOVA, stratified by age. Mode of delivery was the most significant
67 factor driving gut microbiota variation during the neonatal period (Fig. 2a, Supplementary Table 2), while
68 other clinical covariates associated with hospital birth (e.g. perinatal antibiotics, duration of hospital stay)

69 and breastfeeding exhibited smaller effects (Supplementary Note 1). The largest effect of delivery mode
70 was observed on day 4 (Fig. 2b, $R^2=7.64\%$, $P<0.001$), which dissipated with age but remained significant
71 at the point of infancy sampling ($R^2=1.00\%$, $P<0.01$). No difference was observed in maternal gut
72 microbiotas by delivery modes or neonatal gut microbiotas between elective and emergency C-section
73 births (Supplementary Table 3).

74 Given the significant effect of the mode of delivery during the neonatal period, we next sought to
75 understand how the microbiota composition and developmental trajectory were altered. Samples from
76 babies delivered vaginally were enriched with *Bifidobacterium* (e.g. *B. longum*, *B. breve*), *Escherichia* (*E.*
77 *coli*) and *Bacteroides/Parabacteroides* species (e.g. *B. vulgatus*, *P. distasonis*) with these commensal
78 genera comprising 68.3% (95% CI 65.7-71.0%) of the neonatal gut microbial communities (Fig. 2c,
79 Supplementary Table 5), which validated the recent observations in other cohorts^{4,13}. In contrast, the gut
80 microbiota of C-section delivered babies were depleted of these commensal genera and instead were
81 dominated by *Enterococcus* (*E. faecalis*, *E. faecium*), *Staphylococcus epidermis*, *Streptococcus*
82 *parasanguinis*, *Klebsiella* (*K. oxytoca*, *K. pneumoniae*), *Enterobacter cloacae* and *Clostridium perfringens*,
83 which are commonly associated with hospital environments¹⁸ and hospitalised preterm babies¹⁹⁻²¹. On day
84 4, species belonging to these genera accounted for 68.25% (95% CI 62.74-73.75%) of the total microbiota
85 composition in C-section delivered babies (Fig. 2c).

86 Previous studies reported that, compared to C-section delivered babies, the gut microbiotas of
87 vaginally delivered babies were enriched in lactobacilli associated with the mother's vaginal microbiota^{1,22}.
88 However, here we observed no statistical difference in the prevalence (vaginal 11.9% vs C-section 15.7%
89 present at over 1% abundance) or abundance of *Lactobacillus* between vaginally (1.217%, 95% CI 0.81-
90 1.621%) or C-section (2.21%, 95% CI 1.54-2.88%) delivered babies. Rather, commensal species from the
91 *Bacteroides* genus were detected at high abundance in the gut microbiota of 49.0% (154/314) of vaginally
92 delivered babies (mean relative abundance 8.13%, 95% CI 6.88-9.39%, Extended Data Fig. 3). In contrast,
93 *Bacteroides* species were low or absent in 99.6% (281/282) C-section delivered babies (mean relative
94 abundance 0.43%, 95% CI 0.11-0.74). In 60.6% (86/142) of the C-section babies, this low-*Bacteroides*
95 profile (defined in Methods) persisted into infancy, when *Bacteroides* became the only differentially
96 abundant species between vaginally and C-section delivered babies (Supplementary Table 5). Although we

97 could not assess the independent effect of maternal antibiotic exposure during C-section delivery as
98 antibiotics were administered in all C-section deliveries, among vaginally delivered babies we observed a
99 statistically significant association between the low-*Bacteroides* profile with maternal intrapartum
100 antibiotic prophylaxis (IAP, OR=1.77, 95% CI: 1.17-2.71, P=0.0074), which also accounted for the greatest
101 amount of gut microbiota variation in vaginally delivered babies ($R^2=5.88-13.6\%$, Supplementary Table 4).
102 These results expand on previous findings^{9,23} and further highlight a low-*Bacteroides* profile as the
103 perturbation signature associated with C-section and maternal IAP in vaginal delivery.

104 Maternal transmission of gastrointestinal bacteria to their babies is an underappreciated form of
105 kinship²⁴. To assess if the neonatal microbiota variation could be attributed to differential transmission of
106 maternal microbiota, we profiled the bacterial strain transmission across 178 mother-baby dyads. We show
107 that the majority of maternal strain transmissions during the neonatal period occurred in vaginally delivered
108 babies (74.39%), at much higher frequency in comparison with those delivered by C-section (12.56%,
109 Fisher's exact test, P<0.0001, Fig 3a, Extended Data Fig. 4, Supplementary Tables 6-7). *Bacteroides* spp.,
110 *Parabacteroides* spp., *E. coli* and *Bifidobacterium* spp. were most frequently transmitted from mothers to
111 babies through vaginal birth, in agreement with previous observation in smaller cohorts^{4,25-27}. For
112 *Bacteroides* species such as *B. vulgatus* (Fig. 3b), the lack of transmission continued far beyond the neonatal
113 period in C-section born babies²⁵ with the late transmission of *B. vulgatus* rarely detected later in infancy.
114 This is in contrast to the transmission pattern of other common early colonisers such as *B. longum* (Fig. 3c)
115 and *E. coli*, for which colonisations of maternal strains occurred more frequently later in infancy (Fisher's
116 exact tests, P=0.0479 and P=0.0226, respectively). This result highlights the neonatal period as a critical
117 early window of maternal transmission with the disrupted transmission of pioneering *Bacteroides* species
118 evident in C-section babies with long-term *Bacteroides* absence.

119 While C-section babies were deprived of maternally transmitted commensal bacteria, they had a
120 substantially higher relative abundance of opportunistic pathogens commonly associated with the
121 healthcare environment. These enriched species included *E. faecalis*, *E. faecium*, *E. cloacae*, *K.*
122 *pneumoniae*, *K. oxytoca* and *C. perfringens* (Fig. 4a, Supplementary Table 5), some of which are members
123 of the ESKAPE pathogens responsible for the majority of nosocomial infections²⁸. Indeed, their frequent
124 gut microbiota colonisation in C-section newborns was under-reported in previous smaller cohorts^{3,13} with

125 insufficient statistical power (Supplementary Note 2). Among C-section born babies, 83.7% carried
126 opportunistic pathogen species during the neonatal period (as defined in Methods), in comparison to 49.4%
127 of the vaginally born babies (Fig. 4a). During the first 21 days of life, these healthcare-associated
128 opportunistic pathogens accounted for 30.4% (95% CI 27.86-32.96%) of the species level abundance in the
129 gut microbiota of C-section babies, compared to 9.8% (95% CI 8.19-11.4%) in the vaginal babies, with the
130 greatest difference observed on day 4 (Extended Data Fig. 5a). Longitudinally, the difference in combined
131 opportunistic pathogen abundance persisted in the C-section babies re-sampled later in infancy (C-section
132 2.8% versus vaginal 1.6%, $P=0.0375$, Welch's t-test). Interestingly, frequent and abundant carriage of
133 opportunistic pathogens were also observed in low-*Bacteroides* vaginally delivered babies (Extended Data
134 Fig. 5b), while the absence of breastfeeding during the neonatal period was associated with a higher carriage
135 of *C. perfringens*, *K. oxytoca* and *E. faecalis* (Supplementary Table 5).

136 Given the prevalent carriage of opportunistic pathogens in the neonatal gut metagenomes, we sought
137 to validate their presence and viability with culturing. We undertook targeted large-scale culturing of 836
138 opportunistic pathogen strains in the faecal samples of 177 babies (70 vaginal and 107 C-section babies,
139 total 741 isolates) and 38 mothers (95 isolates) using selective media (Fig. 4b, Supplementary Table 8).
140 Subsequent WGS and genomic characterisation of *E. faecalis* (n=356), *E. cloacae* (n=52), *K. oxytoca*
141 (n=150) and *K. pneumoniae* (n=78) allowed us to perform high-resolution phylogenetic analysis and to
142 delineate strain-specific carriage of AMR genes and virulence factors.

143 Focusing on the most prevalent opportunistic pathogen in C-section born babies, we analysed the
144 genomes of a diverse population of BBS *E. faecalis* strains in the context of publicly available genomes of
145 human and environmental strains (Fig. 4c). We found that 53.9% of the BBS strains were represented by
146 five major lineages, each of which was distributed across vaginal and C-section babies and mothers in the
147 three BBS hospitals (Extended Data Fig. 6a) and UK hospital patients, but did not include high-risk UK
148 epidemic lineages enriched in multi-drug resistance (MDR) and virulence²⁹. In congruence with the
149 phylogenetic placement of the BBS strains with the human gastrointestinal and environmental strains, these
150 non-epidemic *E. faecalis* exhibited comparable levels of carriage of AMR genes (Extended Data Fig. 6b,
151 Supplementary Note 3). Similar to *E. faecalis*, the BBS *Enterobacter* and *Klebsiella* strains also exhibited
152 high-level population diversities with the phylogenetic under-representation of epidemic lineages (Fig. 4d,

153 Extended Data Fig. 7), and levels of AMR and virulence gene carriage indicative of non-epidemic lineages
154 circulating in hospital environments and healthy populations, rather than hypervirulent and ESBL-enriched
155 epidemic lineages³⁰⁻³² (Extended Data Fig. 8, Supplementary Note 3). Given the prior isolation of the major
156 BBS lineages in hospitalised patients and their AMR and virulence capabilities, any level of opportunistic
157 pathogen carriage represents a significant risk of future infections, especially for the C-section born babies
158 with high prevalence (83.7%) of carriage.

159 Whilst there is insufficient evidence from metagenomics and cultured isolate WGS that indicates
160 an apparent maternal origin of the opportunistic pathogens (Supplementary Note 4), the absence of lineage-
161 specific colonisation suggests hospital environmental exposure as the primary factor driving opportunistic
162 pathogen colonisation of the BBS babies. Although our study was not designed for retrospective sampling
163 of the hospital environmental sources, opportunistic pathogens are frequently found in hospital
164 environments, where hospital-born babies have been shown to carry the same bacteria present in operating
165 rooms³³ and neonatal intensive care units³⁴.

166 Undertaking the largest, longitudinal WGS characterisation of the human gut microbiota in the
167 previously under-sampled neonatal period (≤ 1 month), we consolidate the recent findings that mode of
168 delivery is a major factor shaping the gut microbiota in the first few weeks of life⁴, with the diminished
169 effect persisting into infancy^{14,15}. The disrupted transmission of the maternal gastrointestinal bacteria,
170 particularly the pioneering *Bacteroides* species in birth via C-section and maternal IAP, predisposed
171 newborn babies to colonisation by clinically important opportunistic pathogens circulating in healthcare
172 and hospital environments. However, the clinical consequences of the early life microbiota perturbations
173 and carriage of immunogenic pathogens during this critical window of immune development remain to be
174 determined. This highlights the need for large-scale, long-term cohort studies that also sample home births³⁵
175 to better understand the consequence of hospital birth and establish if neonatal microbiota perturbation
176 negatively impacts health outcomes in childhood and later life.

177 **Figure legends**

178 **Fig. 1: Developmental dynamics of the neonatal gut microbiota.**

179 **a**, Longitudinal metagenomic sampling of 1,679 early-life gut microbiotas of 772 individuals from
180 three participating hospitals (A, B, C) of the Baby Biome Study. Each row corresponds to the time
181 course of a subject, comprising 596 babies sampled during the neonatal period primarily on day 4
182 (n=310), 7 (n=532) and 21 (n=325), in infancy (8.75 ± 1.98 months of age, n = 302), and from matched
183 mothers (n = 175). **b**, Non-metric multidimensional scaling (NMDS) ordination of Bray–Curtis
184 dissimilarity (n = 917) between the species relative abundance profiles of the gut microbiota sampled
185 from babies sampled on day 4, day 7, day 21, in infancy and from mothers (n = 175).

186 **Fig. 2: Perturbed neonatal gut microbiota composition and development associated with the**
187 **mode of delivery**

188 **a**, Bar plot illustrating the clinical covariates associated with the neonatal gut microbiota variations on
189 day 4 (n=310), day 7 (n=532), day 21 (n=325) and in infancy (n=302). Only the statistically
190 significant associations in cross-sectional tests are shown. Covariates are ranked by the number
191 statistically significant effect observed across sampling age groups. The proportion of explained
192 variance (R^2) and statistical significance were calculated using PERMANOVA on between-sample
193 Bray-Curtis distances. **b**, Non-metric multidimensional scaling (NMDS) ordination of Bray–Curtis
194 dissimilarity between the species relative abundance profiles of the gut microbiota sampled from
195 babies on day 4 (vaginal delivery, n=157; C-section delivery, n=153), day 7 (vaginal delivery, n=280;
196 C-section delivery, n=252), day 21 (vaginal delivery, n=147; C-section delivery, n=178), during
197 infancy (vaginal delivery, n=160; C-section delivery, n = 142) and from mothers (vaginal delivery,
198 n=110; C-section delivery, n=65). Microbial variation explained by the mode of delivery in each
199 cross-section test is shown in the bottom left. All statistical tests were significant with PERMANOVA
200 R^2 and q-values reported in Supplementary Table 2. **c**, Longitudinal changes in the mean relative
201 abundance (RA) of faecal bacteria at the genus level sampled on day 4, 7, 21 days of life and in
202 infancy, for genera with > 1% RA across all neonatal period samples. Vaginal delivery, n=744 from
203 310 babies; C-section delivery, n=725 from 281 babies.

204 **Fig. 3: Disrupted maternal strain transmission in C-section-delivered babies.**

205 **a**, Early and late transmission of the maternal strains in mother-baby pairs (vaginal: 35, C-section: 24)
206 longitudinally sampled during the neonatal (early) and infancy (late) period. Only the frequently
207 shared species detected with sufficient coverage for strain analysis in more than 10 pairs are shown.
208 Phylogenetically related species shared transmission pattern. **b, c** Transmission events of maternal *B.*
209 *vulgatus* (**b**) and *B. longum* (**c**) strains in vaginally delivered, and C-section delivered babies over
210 time. In each row of mother-baby paired samples, each circle represents a detectable strain either
211 identical (filled) to or distinct from (hollow) the maternal strain. Across the rows, identical strains are
212 linked by a solid line representing early transmission and persistence to infancy, while the dashed line
213 indicates late transmission.

214 **Fig. 4: Extensive and frequent colonisation of C-section delivered babies with diverse**
215 **opportunistic pathogen species previously associated with healthcare infection.**

216 **a**, The mean relative abundance (RA) and frequency (>1% RA) of six opportunistic pathogen species
217 enriched C-section born babies (n=596), compared to vaginal-born babies (n=606) during the first 21
218 days of life, in the context of the maternal level carriage (n=175). Error bars indicate the 95% CI of the
219 mean relative abundance. Statistical significance of the differences in RA and frequency was determined
220 by Holm's-adjusted Wilcoxon and Fisher's exact tests, respectively. ***P < 0.001, **P < 0.01, *P < 0.05
221 **b**, Phylogenetic representation of 836 bacterial strains cultured from raw faecal samples, including six
222 opportunistic pathogens isolated five major genera: *Enterococcus spp.* (red, n=451); *Clostridium spp.*
223 (yellow, n=24); *Klebsiella spp.* (blue, n=235), *Enterobacter spp.* (green, n=52) and *Escherichia spp.*
224 (purple, n=41). **c**, Phylogeny of the BBS *E. faecalis* isolates (n=282) in the context of public isolates
225 from UK hospitals (n=168), the healthy human gut microbiotas (n=28) and environmental sources
226 (n=27) with the high-risk UK epidemic lineage (CC2/CC28/CC388) branches coloured in blue.
227 Midpoint-rooted maximum likelihood tree is based on SNPs in 1,656 core genes. **d**, Diverse
228 *Enterobacter-Klebsiella* complex populations among the BBS collection (n=202), in the context of UK
229 hospital (n=604), the healthy human gut microbiotas (n=37) and environmental sources (n=120).

230 **Methods**

231 **Study population**

232 The study was approved by the NHS London - City and East Research Ethics Committee (REC reference
233 12/LO/1492). Participants were recruited at the Barking, Havering and Redbridge University Hospitals
234 NHS Trust (BHR), the University Hospitals Leicester NHS Trust (LEI), and the University College London
235 Hospitals NHS Foundation Trust (UCLH), through the Baby Biome Study (previously Life Study
236 enhancement pilot study) from May 2014 to December 2017. Mothers provided written, informed consent
237 to participate and for their children to participate in the study.

238 **Sample collection**

239 Faecal samples were collected from babies with at least one sample in the first 21 days of life, primarily on
240 day 4, 7 or 21. For a subset of babies who provided neonatal samples, a follow-up faecal sample collection
241 was performed between 4 to 12 months of their lives. Maternal faecal samples were collected in the
242 maternity unit before or after delivery, or stool was collected during delivery by midwives. Baby samples
243 were collected at home by mothers and returned to the processing laboratory by post at ambient temperature
244 within 24 hours. On arrival at the lab, all faecal samples were immediately stored at 4°C for an average of
245 2.41 days (95% CI 2.06-2.76 days) before further processing. Samples were aliquoted into six vials, four
246 of which were stored at -80°C for raw faeces biobanking while the other two vials were processed
247 immediately for DNA extraction. Although this sample storage protocol (no preservation buffer for room
248 temperature and 4°C storage) was shown to be robust to technical variation in microbiome profiles at the
249 time of study design (Supplementary Note 5), state-of-the-art sampling methods should be utilised in future
250 large-scale microbiome to minimise the potential effect of sample storage on the microbiota composition³⁶.
251 DNA was extracted from 30 mg of faecal samples as described in the BBS collection and processing
252 protocol³⁷. Negative controls using ultrapure water was included in parallel for each kit as well as each
253 extraction batch, and DNA concentration quantified to confirm contamination free. Total DNA was eluted
254 in 60µl DNase/Pyrogen-free water, and stored at -80°C until shipment to the Wellcome Sanger Institute for
255 metagenomic sequencing.

256 **Shotgun metagenomic sequencing and analysis**

257 DNA samples, including negative controls, were quantified by PicoGreen dsDNA assay (Thermo Fisher),
258 and samples with >100 ng DNA material proceeded to paired-end (2 x 125bp) metagenomics sequencing
259 on the HiSeq 2500 v4 platform. Low-quality bases were trimmed (*SLIDINGWINDOW:4:20*), and reads
260 below 87 nucleotides (70% of original read length) were removed (*MINLEN:87*) using Trimmomatic³⁸. To
261 remove potential human contaminants, quality trimmed reads were screened against the human genome
262 (GRCh38) with Bowtie2 v2.3.0³⁹. On average, 22.4 (95% CI 22.1-22.6) million raw reads were generated
263 per sample. 19.3 (95% CI 19.1-19.6) million reads (87.3% of the raw reads) per sample passed
264 decontamination and quality trimming steps for downstream analysis. Sequencing depth was accounted for
265 as a potential technical confounding factor in analyses of microbiota species and strain measurements, and
266 significant species association with clinical covariates (Supplementary Note 6). Taxonomic classification
267 from metagenomics reads was performed using Kraken v1.0⁴⁰, a k-mer based sequence classification
268 approach against the Human Gastrointestinal Bacteria Genome Collection (HGG) genomes⁴¹. Bracken
269 v1.0⁴² was run on the Kraken classification output to estimate taxonomic abundance down to the species
270 level. Metagenomic samples were compared at the genus and species levels by relative abundance. A cut-
271 off of 100 Kraken-assigned paired-end reads (corresponds to 0.001% relative abundance given the sampling
272 depth of ~10 million paired-end reads) was applied to determine metagenomic species detection. To assess
273 whether the trade-off between the observed level of *Bacteroides* and opportunistic pathogens was an
274 artefact of compositional effects, the proportion of abundances and reads corresponding to *Bacteroides*
275 were removed separately, prior to relative abundance normalisation. In the normalised datasets, the
276 statistical enrichment of opportunistic pathogen species in C-section babies was consistent with the
277 observation with the original data. The R packages *phyloseq*⁴³ and *microbiome*⁴⁴ was used for metagenomic
278 data analysis and results visualised using *ggplot2*⁴⁵ in RStudio.

279 **Classification of the low-*Bacteroides* babies**

280 For each baby, the median relative abundance of the *Bacteroides* genus was calculated across the neonatal
281 period samples. Based on the threshold described previously⁹, babies with a median abundance of less than
282 0.1% were assigned low-*Bacteroides* status.

283

284 **Classification of the opportunistic pathogen carriage**

285 Total opportunistic pathogen load is estimated by calculating the median relative abundance of combined
286 opportunistic pathogen species (*C. perfringens*, *E. cloacae*, *E. faecalis*, *E. faecium*, *K. oxytoca*, *K.*
287 *pneumoniae*) per individual across their neonatal period samples, and independently for the infancy period
288 and maternal samples. To prioritise on relatively high-level opportunistic pathogen carriage feasible for
289 downstream strain cultivation experiments, individuals with a median abundance of over 1% total
290 opportunistic pathogen load were defined as a positive carriage.

291 **Maternal strain transmission analysis**

292 Strain transmissions in mother-baby paired samples were determined using a single-nucleotide variant
293 calling method⁴⁶. StrainPhlAn was run on pre-processed metagenomes to generate consensus species-
294 specific marker genes for phylogenetic reconstruction of all detectable strains (one dominant strain per
295 sample), using default parameters and with the options "--alignment_program mafft" and "--
296 relaxed_parameters3" as previously described²⁶. No statistically significant variation in sequencing depth
297 was observed between vaginal and C-section born subjects across age groups that had any impact on
298 coverage-dependent microbiota species and strains detection (Supplementary Note 6). For each species and
299 strains with sufficient coverage for strain profiling, we generated a species-specific phylogenetic tree using
300 RAxML⁴⁷. As previously described²⁶, the strain distance for each pair of mother-baby sample strains was
301 computed by calculating the pairwise normalised phylogenetic distance on the corresponding species tree.
302 To define strain transmission events, a previously described²⁶, conservative threshold of 0.1 on the strain
303 distance value was used. The detectable strains in a given pair of mother-baby samples were considered
304 identical (strain distance less than 0.1, transmission) or distinct (strain distance greater than 0.1, no
305 transmission). For all mother-baby pairs shown in Extended Data Fig. 4, early transmission event was
306 counted once per species per mother-baby pair, considering the detected transmission (or evidence for no
307 transmission) at the earliest time point (primary transmission), irrespective of the subsequent transmission
308 events in any later neonatal period samples. For a subset of mother-baby pairs with both neonatal and
309 infancy period sampled (shown in Fig. 3a), late transmission events were counted separately, including
310 cases of no early transmission due to insufficient coverage (no detectable strains). To highlight the
311 transmission pattern shared by phylogenetically related species, a neighbour-joining tree of the eligible

312 species was constructed based on the mash distance matrix⁴⁹ of the respective reference genomes
313 included in the StrainPhlAn database (Supplementary Table 9). The same approach and strain distance
314 threshold (core-genome SNPs) were applied to the cultured strains to count the number of identical and
315 distinct strains within mother-baby and longitudinal paired samples.

316 **Statistical analysis**

317 To calculate the effect of clinical covariates on the gut microbiota composition, we stratified by age groups
318 and then assessed the proportion of explained variance (R^2 from PERMANOVA) in Bray-Curtis distance
319 for each clinical covariate, using the *adonis* from the R package *vegan*⁵⁰. While PERMANOVA is mostly
320 unaffected by group dispersion effects in balanced designs⁵¹ (e.g. mode of delivery comparisons), for
321 unbalanced designs (e.g. breastfeeding comparisons) more sensitive to group dispersion effects, the group
322 variance homogeneity condition was validated using the *betadisper* function. Group dispersions were not
323 significantly different (*betadisper* $P < 0.05$) in all comparisons, which lent support to the statistically
324 significant, albeit visibly weak effects of breastfeeding as reported by PERMANOVA. Samples with
325 missing metadata (NA) for the given clinical covariate were excluded prior to running each cross-sectional
326 analysis. Effect sizes and statistical significance were determined by 1,000 permutations, and *P*-values
327 corrected for multiple testing using the Benjamini-Hochberg false discovery rate (FDR = 5%). Statistical
328 tests of between-group taxonomic abundance comparisons (Welch's t-test with p-values FDR-corrected)
329 were performed in the Statistical Analysis of Metagenomics Profiles program v2.0⁵². MaAsLin⁵³ was used
330 for adjustment of covariates when determining the significance of species associated with a specific
331 variable while accounting for potentially confounding covariates, as previously described^{14,15}. All the
332 covariates tested in the PERMANOVA were included in the adjustment along with the sequencing depth
333 used as fixed effects. The default MaAsLin parameters were applied (maximum percentage of samples NA
334 in metadata 10%, minimum percentage relative abundance 0.01%, $P < 0.05$, $q < 0.25$).

335 **Bacterial isolation and whole-genome sequencing**

336 Raw faecal samples from neonates stored in the biobank lab at -80°C were requested based on faecal
337 carriage of targeted species over 1% relative abundance in metagenomes. Selected frozen faecal aliquots,
338 where available (> 100 ng) were couriered on dry ice to the Wellcome Sanger Institute within 6 hours of
339 shipment from the biobank lab. Bacterial isolates were cultured using the following culture media:

340 *Enterococcus faecium* ChromoSelect Agar Base (Sigma-Aldrich) for *Enterococcus* spp., CP ChromoSelect
341 Agar (Sigma-Aldrich) for *Closteridium* spp., Coliform ChromoSelect Agar (Sigma-Aldrich) and *Klebsiella*
342 ChromoSelect Selective Agar (Sigma-Aldrich) for species of *Enterobacteriaceae*. Between 2-5 colonies
343 per sample were picked for full-length 16S rRNA gene sequencing to confirm species identification, as
344 described previously⁵⁴. Bacterial isolates with species identification congruent with metagenomic
345 identification were re-streaked and purified for genomic DNA extraction using DNeasy 96 kit. DNA
346 sequencing was performed on the Illumina HiSeq X, generating paired-end reads (2 x 151bp). Multiple
347 strains per species per faecal sample were also sequenced based on variation across the full-length 16S
348 rRNA sequences. Bacterial genomes were assembled and annotated using the pipeline described
349 previously⁵⁵. Genome assemblies were subjected to quality check and contaminant screening with
350 CheckM⁵⁶ and Mash⁵⁷, respectively. Where applicable, the suspected contaminant (non-target organism)
351 sequences were confirmed and filtered out via raw read mapping using Bowtie2 v2.3.0, prior to re-
352 assembly.

353 **Bacterial phylogenetic analysis**

354 The phylogenetic analysis of the complete diverse species collection was conducted by extracting the amino
355 acid sequence of 40 universal core marker genes^{58,59} from the BBS bacterial culture collection using
356 SpecI⁶⁰. The protein sequences were concatenated and aligned with MAFFT v.7.2040, and maximum-
357 likelihood trees were constructed using RAxML⁴⁷ with default settings. Four most prevalent BBS collection
358 opportunistic pathogen species *E. faecalis*, *E. cloacae*, *K. oxytoca* and *K. pneumoniae* were further analysed
359 in context of the public genomes (Supplementary Table 10), including the UK hospital strain collections²⁹⁻
360 ³², the gut microbiota-cultured strains from the HGG and the Culturable Genome Reference (CGR)⁶¹
361 collections, and the environmental strains on the Genome Taxonomy Database (GTDB, v86)⁶². To generate
362 phylogenetic trees of individual species, the public genome assemblies were combined with the assemblies
363 of the study isolates, annotated with Prokka⁶³, and a pangenome estimated using Roary⁶⁴. Where multiple
364 identical strains (no SNP difference in species core-genome) were cultured from the same faecal sample,
365 only one representative strain was included in the species phylogenetic trees. A 95% identity cut-off was
366 used, and core genes were defined as those in 99% of isolates unless otherwise stated. A maximum
367 likelihood tree of the SNPs in the core genes was created using RAxML⁴⁷ and 100 bootstraps. To illustrate

368 the population structure of the closely related *Enterobacter* and *Klebsiella* strain isolates, FastANI⁶⁵ was
369 used to estimate the pairwise average nucleotide identity distance between all public and BBS genome
370 assemblies, which was then used as an input to generate a neighbour-joining with BIONJ⁶⁶. All
371 phylogenetic trees were visualised in iTOL⁶⁷. Sequence types were determined using MLSTcheck⁶⁸, which
372 was used to compare the assembled genomes against the MLST database for the corresponding species.

373 **Detecting virulence and resistance genes**

374 ABRicate (v0.8.13, <https://github.com/tseemann/abricate>) was used to screen for known, acquired
375 resistance genes and virulence factors against bacterial genome assemblies. For AMR genes, a
376 comprehensive BLAST database integrating 5,556 non-redundant sequences in the NCBI Bacterial
377 Antimicrobial Resistance Reference Gene Database (PRJNA313047), CARD v2.0.3, ARG-ANNOT and
378 ResFinder was queried against. 3,202 non-redundant experimentally validated core virulence genes in
379 VFDB (version 5 Oct 2018) were included to build a BLAST database for virulence factor screening.

380 **Acknowledgements**

381 This work was supported by the Wellcome Trust (WT101169MA) and Wellcome Sanger core funding
382 (WT098051). Y.S. is supported by a Wellcome Trust PhD Studentship. S.C.F. is supported by the
383 Australian National Health and Medical Research Council [1091097, 1159239 and 1141564] and the
384 Victorian Government's Operational Infrastructure Support Program. We are grateful to the participating
385 families for their time and contribution to the Baby Biome Study, the research midwives at recruiting
386 hospitals for recruitment and clinical metadata collection, Nadia Moreno, Henna Ali, Samra Bibi and Alfred
387 Takyi for raw sample processing. We thank the Core Sequencing and Pathogen Informatics teams at the
388 Wellcome Sanger Institute for informatics support, H. Browne and A. Almeida for critical feedback of the
389 manuscript.

390 **Contributions**

391 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.
392 carried out the pilot study, and designed sample collection and processing protocols, overseen by N.F. and
393 T.D.L.; E.T., A.S., N.S. and N.F. managed participant recruitment and coordinated clinical metadata
394 collection; Y.S. performed bacterial culturing and DNA extraction with assistance from M.D.S.; Y.S.
395 generated and analysed the data with assistance from K.V.; Y.S., S.C.F., N.F. and T.D.L. wrote the
396 manuscript. All authors read and approved the manuscript.

397 **Competing interests**

398 The authors declare no competing financial interests.

399 **Corresponding authors**

400 Correspondence to Trevor D. Lawley or Nigel Field.

401 **Data availability**

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

405 **References**

- 406 1. Dominguez-Bello, M. G. *et al.* Delivery mode shapes the acquisition and structure of the
407 initial microbiota across multiple body habitats in newborns. *PNAS* **107**, 11971–11975
408 (2010).
- 409 2. Tamburini, S., Shen, N., Wu, H. C. & Clemente, J. C. The microbiome in early life:
410 implications for health outcomes. *Nat. Med.* **22**, 713–722 (2016).
- 411 3. Chu, D. M. *et al.* Maturation of the infant microbiome community structure and function
412 across multiple body sites and in relation to mode of delivery. *Nat. Med.* (2017).
413 doi:10.1038/nm.4272
- 414 4. Wampach, L. *et al.* Birth mode is associated with earliest strain-conferred gut microbiome
415 functions and immunostimulatory potential. *Nat Commun* **9**, 5091 (2018).
- 416 5. Koenig, J. E. *et al.* Succession of microbial consortia in the developing infant gut
417 microbiome. *Proc. Natl. Acad. Sci. U.S.A.* **108 Suppl 1**, 4578–4585 (2011).
- 418 6. Stokholm, J. *et al.* Cesarean section changes neonatal gut colonisation. *Journal of Allergy
419 and Clinical Immunology* **138**, 881–889.e2 (2016).
- 420 7. Baumann-Dudenhoeffer, A. M., D’Souza, A. W., Tarr, P. I., Warner, B. B. & Dantas, G.
421 Infant diet and maternal gestational weight gain predict early metabolic maturation of gut
422 microbiomes. *Nat. Med.* **5**, 178 (2018).
- 423 8. Bokulich, N. A. *et al.* Antibiotics, birth mode, and diet shape microbiome maturation during
424 early life. *Science Translational Medicine* **8**, 343ra82–343ra82 (2016).
- 425 9. Yassour, M. *et al.* Natural history of the infant gut microbiome and impact of antibiotic
426 treatment on bacterial strain diversity and stability. *Science Translational Medicine* **8**,
427 343ra81–343ra81 (2016).
- 428 10. Arrieta, M.-C. *et al.* Early infancy microbial and metabolic alterations affect risk of
429 childhood asthma. *Science Translational Medicine* **7**, 307ra152–307ra152 (2015).
- 430 11. Fujimura, K. E. *et al.* Neonatal gut microbiota associates with childhood multisensitized
431 atopy and T cell differentiation. *Nat. Med.* (2016). doi:10.1038/nm.4176
- 432 12. Stokholm, J. *et al.* Maturation of the gut microbiome and risk of asthma in childhood. *Nat
433 Commun* **9**, 141 (2018).
- 434 13. Bäckhed, F. *et al.* Dynamics and Stabilization of the Human Gut Microbiome during the First
435 Year of Life. *Cell Host & Microbe* **17**, 690–703 (2015).
- 436 14. Stewart, C. J. *et al.* Temporal development of the gut microbiome in early childhood from
437 the TEDDY study. *Nature* **562**, 583–588 (2018).
- 438 15. Vatanen, T. *et al.* The human gut microbiome in early-onset type 1 diabetes from the
439 TEDDY study. *Nature* **562**, 589–594 (2018).
- 440 16. Vatanen, T. *et al.* Variation in Microbiome LPS Immunogenicity Contributes to
441 Autoimmunity in Humans. *Cell* **165**, 1551 (2016).
- 442 17. Olin, A. *et al.* Stereotypic Immune System Development in Newborn Children. *Cell* **174**,
443 1277–1292.e14 (2018).
- 444 18. Lax, S. *et al.* Bacterial colonisation and succession in a newly opened hospital. *Science
445 Translational Medicine* **9**, eaah6500 (2017).
- 446 19. Stewart, C. J. *et al.* Preterm gut microbiota and metabolome following discharge from
447 intensive care. *Scientific Reports* **5**, 17141 (2015).
- 448 20. Gibson, M. K. *et al.* Developmental dynamics of the preterm infant gut microbiota and
449 antibiotic resistome. *Nature Microbiology* **1**, 1–10 (2016).
- 450 21. Raveh-Sadka, T. *et al.* Evidence for persistent and shared bacterial strains against a
451 background of largely unique gut colonisation in hospitalised premature infants. *The ISME
452 Journal* **10**, 2817–2830 (2016).
- 453 22. Dominguez-Bello, M. G. *et al.* Partial restoration of the microbiota of cesarean-born infants
454 via vaginal microbial transfer. *Nat. Med.* **22**, 250–253 (2016).
- 455 23. Jakobsson, H. E. *et al.* Decreased gut microbiota diversity, delayed Bacteroidetes
456 colonisation and reduced Th1 responses in infants delivered by caesarean section. *Gut* **63**,
457 559–566 (2014).

- 458 24. Funkhouser, L. J. & Bordenstein, S. R. Mom Knows Best: The Universality of Maternal
459 Microbial Transmission. *PLOS Biology* **11**, e1001631 (2013).
- 460 25. Nayfach, S., Rodriguez-Mueller, B., Garud, N. & Pollard, K. S. An integrated metagenomics
461 pipeline for strain profiling reveals novel patterns of bacterial transmission and
462 biogeography. *Genome Res.* **26**, 1612–1625 (2016).
- 463 26. Ferretti, P. *et al.* Mother-to-Infant Microbial Transmission from Different Body Sites Shapes
464 the Developing Infant Gut Microbiome. *Cell Host & Microbe* **24**, 133–145.e5 (2018).
- 465 27. Yassour, M. *et al.* Strain-Level Analysis of Mother-to-Child Bacterial Transmission during
466 the First Few Months of Life. *Cell Host & Microbe* **24**, 146–154.e4 (2018).
- 467 28. Boucher, H. W. *et al.* Bad bugs, no drugs: no ESKAPE! An update from the Infectious
468 Diseases Society of America. *Clin. Infect. Dis.* **48**, 1–12 (2009).
- 469 29. Raven, K. E. *et al.* Genome-based characterisation of hospital-adapted *Enterococcus faecalis*
470 lineages. *Nature Microbiology* **1**, 15033 (2016).
- 471 30. Moradigaravand, D., Reuter, S., Martin, V., Peacock, S. J. & Parkhill, J. The dissemination
472 of multidrug-resistant *Enterobacter cloacae* throughout the UK and Ireland. *Nature*
473 *Microbiology* **1**, 16173 (2016).
- 474 31. Moradigaravand, D., Martin, V., Peacock, S. J. & Parkhill, J. Population Structure of
475 Multidrug-Resistant *Klebsiella oxytoca* within Hospitals across the United Kingdom and
476 Ireland Identifies Sharing of Virulence and Resistance Genes with *K. pneumoniae*. *Genome*
477 *Biology and Evolution* **9**, 574–584 (2017).
- 478 32. Moradigaravand, D., Martin, V., Peacock, S. J., Parkhill, J. & Chiller, T. Evolution and
479 Epidemiology of Multidrug-Resistant *Klebsiella pneumoniae* in the United Kingdom and
480 Ireland. *MBio* **8**, e01976–16 (2017).
- 481 33. Shin, H. *et al.* The first microbial environment of infants born by C-section: the operating
482 room microbes. *Microbiome 2015 3:1* **3**, 59 (2015).
- 483 34. Brooks, B. *et al.* The developing premature infant gut microbiome is a major factor shaping
484 the microbiome of neonatal intensive care unit rooms. *Microbiome 2015 3:1* **6**, 112 (2018).
- 485 35. Combellick, J. L. *et al.* Differences in the fecal microbiota of neonates born at home or in the
486 hospital. *Scientific Reports* **8**, 15660 (2018).
- 487 36. Vandeputte, D., Tito, R. Y., Vanleeuwen, R., Falony, G. & Raes, J. Practical considerations
488 for large-scale gut microbiome studies. *FEMS Microbiol. Rev.* **41**, S154–S167 (2017).
- 489 37. Bailey, S. R. *et al.* A pilot study to understand feasibility and acceptability of stool and cord
490 blood sample collection for a large-scale longitudinal birth cohort. *BMC Pregnancy*
491 *Childbirth* **17**, 439 (2017).
- 492 38. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina
493 sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
- 494 39. Ben Langmead & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**,
495 357–359 (2012).
- 496 40. Wood, D. E. & Salzberg, S. L. Kraken: ultrafast metagenomic sequence classification using
497 exact alignments. **15**, R46 (2014).
- 498 41. Forster, S. C. *et al.* A human gut bacterial genome and culture collection for improved
499 metagenomic analyses. *Nature Biotechnology* **37**, 186–192 (2019).
- 500 42. Lu, J., Breitwieser, F. P., Thielen, P. & Salzberg, S. L. Bracken: estimating species
501 abundance in metagenomics data. *PeerJ Computer Science* **3**, e104 (2017).
- 502 43. McMurdie, P. J. & Holmes, S. phyloseq: An R Package for Reproducible Interactive
503 Analysis and Graphics of Microbiome Census Data. *PLOS ONE* **8**, e61217 (2013).
- 504 44. Lahti, L. & Shetty, S. Tools for microbiome analysis in R. Version 1.1.10013.
505 URL: <http://microbiome.github.com/microbiome>. (2017).
- 506 45. Wickham, H. *ggplot2: elegant graphics for data analysis*. (2016).
- 507 46. Truong, D. T., Tett, A., Pasolli, E., Huttenhower, C. & Segata, N. Microbial strain-level
508 population structure and genetic diversity from metagenomes. *Genome Res.* gr.216242.116
509 (2017). doi:10.1101/gr.216242.116
- 510 47. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large
511 phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).

- 512 48. Simonsen, M., Mailund, T. & Pedersen, C. N. S. in *Algorithms in Bioinformatics* **5251**, 113–
513 122 (Springer, Berlin, Heidelberg, 2008).
- 514 49. Ondov, B. D. *et al.* Mash: fast genome and metagenome distance estimation using MinHash.
515 *Genome Biology* 2014 15:3 **17**, 132 (2016).
- 516 50. Oksanen, J., Blanchet, F. G., Kindt, R. & Legendre, P. *R Package ‘vegan’: Community
517 Ecology Package. R Package version 2.2–0.* (2014).
- 518 51. Anderson, M. J. & Walsh, D. C. I. PERMANOVA, ANOSIM, and the Mantel test in the face
519 of heterogeneous dispersions: What null hypothesis are you testing? *Ecological Monographs*
520 **83**, 557–574 (2013).
- 521 52. Parks, D. H., Tyson, G. W., Hugenholtz, P. & Beiko, R. G. STAMP: statistical analysis of
522 taxonomic and functional profiles. *Bioinformatics* **30**, 3123–3124 (2014).
- 523 53. Morgan, X. C. *et al.* Dysfunction of the intestinal microbiome in inflammatory bowel disease
524 and treatment. *Genome Biology* 2014 15:3 **13**, R79 (2012).
- 525 54. Browne, H. P. *et al.* Culturing of ‘unculturable’ human microbiota reveals novel taxa and
526 extensive sporulation. *Nature* **533**, 543–546 (2016).
- 527 55. Page, A. J. *et al.* Robust high-throughput prokaryote de novo assembly and improvement
528 pipeline for Illumina data. *Microbial Genomics* **2**, e000083 (2016).
- 529 56. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM:
530 assessing the quality of microbial genomes recovered from isolates, single cells, and
531 metagenomes. *Genome Res.* **25**, 1043–1055 (2015).
- 532 57. Ondov, B. D. *et al.* Mash Screen: High-throughput sequence containment estimation for
533 genome discovery. *bioRxiv* 557314 (2019). doi:10.1101/557314
- 534 58. Sorek, R. *et al.* Genome-Wide Experimental Determination of Barriers to Horizontal Gene
535 Transfer. *Science* **318**, 1449–1452 (2007).
- 536 59. Ciccarelli, F. D. *et al.* Toward Automatic Reconstruction of a Highly Resolved Tree of Life.
537 *Science* **311**, 1283–1287 (2006).
- 538 60. Mende, D. R., Sunagawa, S., Zeller, G. & Bork, P. Accurate and universal delineation of
539 prokaryotic species. *Nat. Methods* **10**, 881–884 (2013).
- 540 61. Zou, Y. *et al.* 1,520 reference genomes from cultivated human gut bacteria enable functional
541 microbiome analyses. *Nature Biotechnology* **37**, 179–185 (2019).
- 542 62. Parks, D. H. *et al.* A standardised bacterial taxonomy based on genome phylogeny
543 substantially revises the tree of life. *Nature Biotechnology* **36**, 996 (2018).
- 544 63. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069
545 (2014).
- 546 64. Page, A. J. *et al.* Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics*
547 **31**, 3691–3693 (2015).
- 548 65. Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T. & Aluru, S. High-
549 throughput ANI Analysis of 90K Prokaryotic Genomes Reveals Clear Species Boundaries.
550 *bioRxiv* 225342 (2017). doi:10.1101/225342
- 551 66. Gascuel, O. BIONJ: an improved version of the NJ algorithm based on a simple model of
552 sequence data. *Molecular Biology and Evolution* **14**, 685–695 (1997).
- 553 67. Letunic, I. & Bork, P. Interactive tree of life (iTOL) v3: an online tool for the display and
554 annotation of phylogenetic and other trees. *Nucl. Acids Res.* **44**, W242–W245 (2016).
- 555 68. Page, A. J., Taylor, B., Softw, J. K. J. O. S.2016. Multilocus sequence typing by blast from
556 de novo assemblies against PubMLST. *theoj.org*

Figure 1

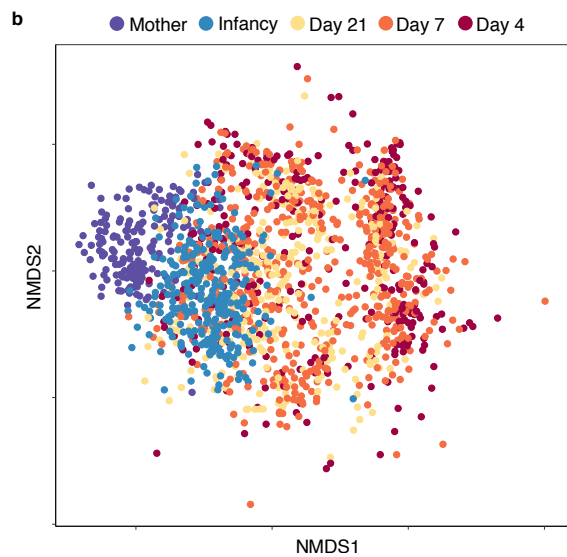
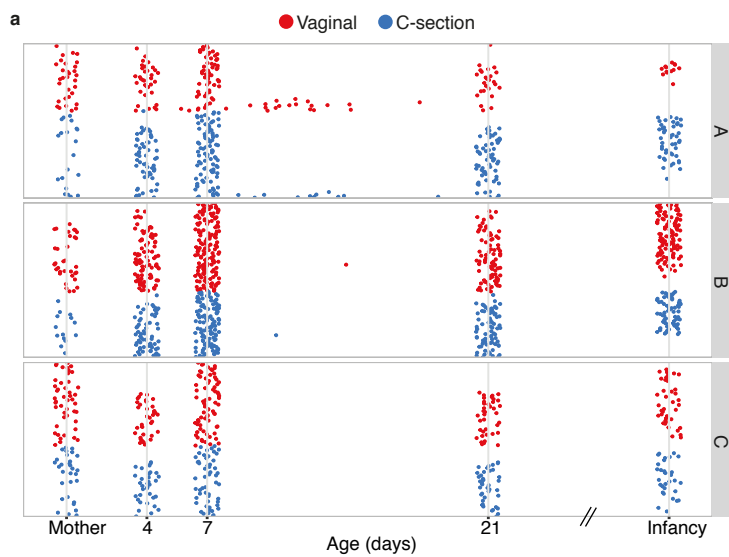


Figure 2

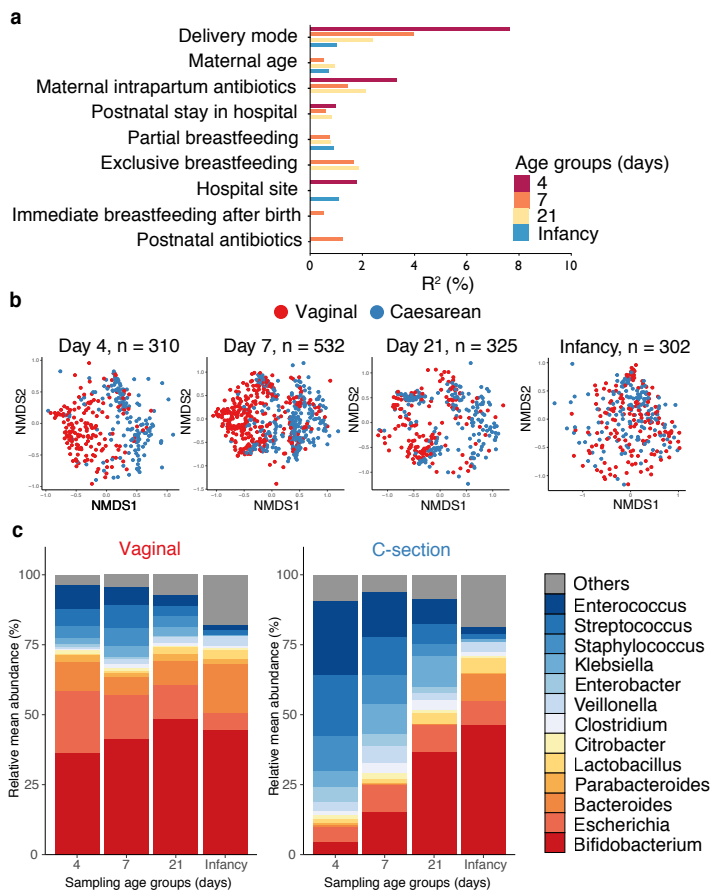


Figure 3

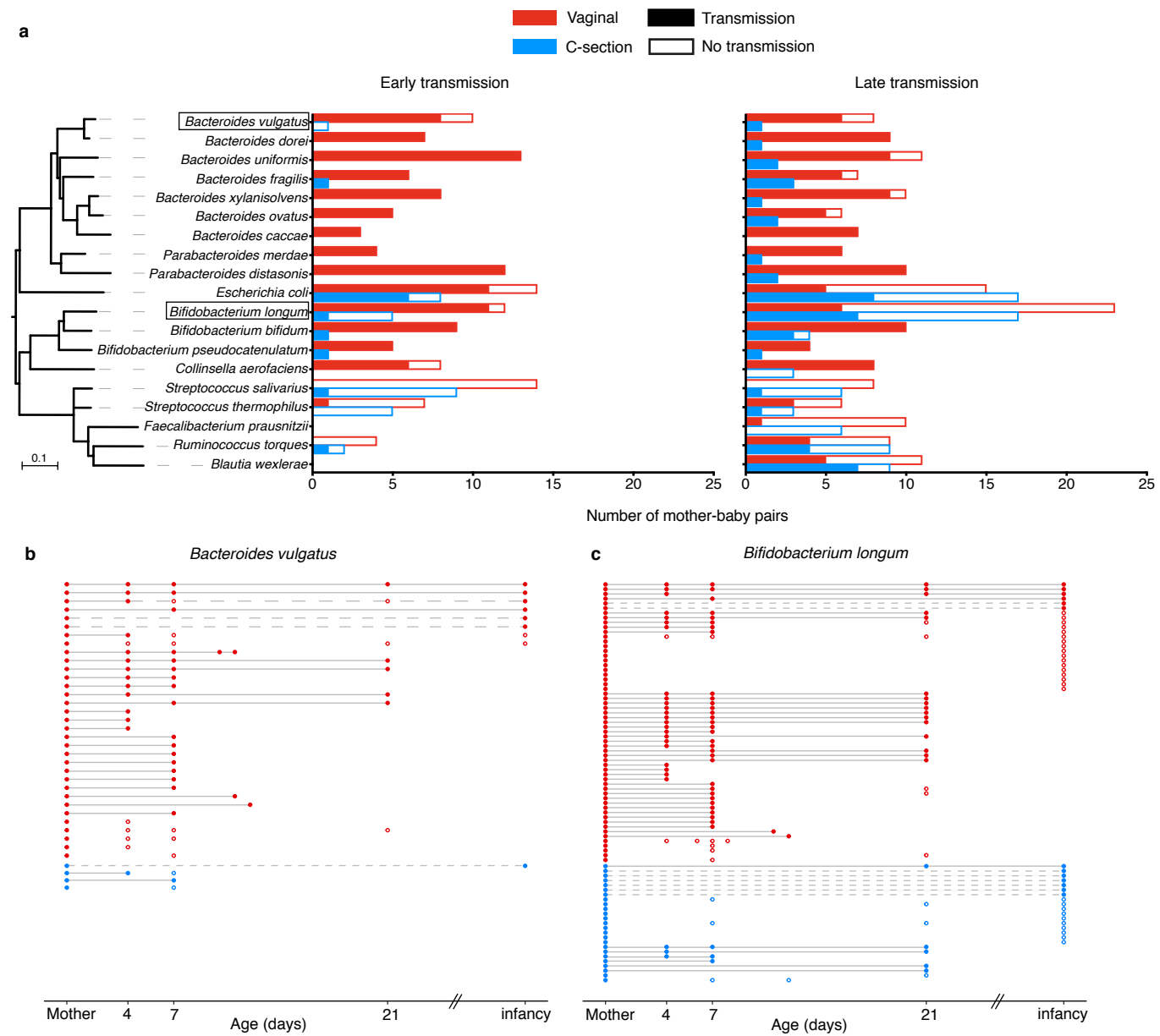


Figure 4