ARTICLE

Corrected: Author Correction

https://doi.org/10.1038/s41586-019-1451-5

Human placenta has no microbiome but can contain potential pathogens

Marcus C. de Goffau[1,2,8], Susanne Lager[3,4,5,8], Ulla Sovio[3,4], Francesca Gaccioli[3,4], Emma Cook[3], Sharon J. Peacock[1,6,7], Julian Parkhill[1,2]*, D. Stephen Charnock-Jones[3,4,9] & Gordon C. S. Smith[3,4,9]*

We sought to determine whether pre-eclampsia, spontaneous preterm birth or the delivery of infants who are small for gestational age were associated with the presence of bacterial DNA in the human placenta. Here we show that there was no evidence for the presence of bacteria in the large majority of placental samples, from both complicated and uncomplicated pregnancies. Almost all signals were related either to the acquisition of bacteria during labour and delivery, or to contamination of laboratory reagents with bacterial DNA. The exception was *Streptococcus agalactiae* (group B Streptococcus), for which non-contaminant signals were detected in approximately 5% of samples collected before the onset of labour. We conclude that bacterial infection of the placenta is not a common cause of adverse pregnancy outcome and that the human placenta does not have a microbiome, but it does represent a potential site of perinatal acquisition of *S. agalactiae*, a major cause of neonatal sepsis.

Placental dysfunction is associated with common adverse pregnancy outcomes that determine a substantial proportion of the global burden of disease[1]. However, the cause of placental dysfunction in most cases is unknown. Several studies have used sequencing-based methods for bacterial detection (metagenomics and 16S rRNA gene amplicon sequencing), and have concluded that the placenta is physiologically colonized by a diverse population of bacteria (the 'placental microbiome') and that the nature of this colonization may differ between healthy and complicated pregnancies[2–4]. This contrasts with the view in the pre-sequencing era that the placenta was normally sterile[5]. However, several studies that applied sequencing-based methods informed by the potential for false-positive results due to contamination[6–8] have failed to detect a placental microbiome[9–12]. The aim of the present study was to determine whether pre-eclampsia, delivery of a small for gestational age (SGA) infant and spontaneous preterm birth (PTB) were associated with the presence or a pattern of bacterial DNA in the placenta and to determine whether there was evidence to support the existence of a placental microbiome. We used samples from a large, prospective cohort study of nulliparous pregnant women[13], and applied an experimental approach informed by the potential for false-positive results[14].

Experimental approach

We studied two cohorts of patients (Extended Data Fig. 1 and Supplementary Tables 1, 2). In cohort 1, babies were all delivered by pre-labour Caesarean section, and the cohort included 20 patients with pre-eclampsia, 20 SGA infants, and 40 matched controls. The placental biopsies were spiked with approximately 1,100 colony-forming units (CFUs) of *Salmonella bongori* (positive control) and samples were analysed using both deep metagenomic sequencing of total DNA (424 million reads on average per sample) and 16S rRNA gene amplicon sequencing. Cohort 2 included 100 patients with pre-eclampsia, 100 SGA infants, 198 matched controls (two controls were used twice) and 100 preterm births. All of these samples were analysed twice using 16S rRNA gene amplicon sequencing from DNA extracted by two different kits.

Cohort 1: metagenomics and 16S rRNA

The positive control (*S. bongori*, average 180 reads per sample, Extended Data Fig. 2a) was detected in all samples. Several other bacterial signals were also observed. Principal component analysis (PCA) (Fig. 1a) demonstrated that almost all of the variation in the metagenomics data (98%) was represented by principal components 1 (80%) and 2 (18%). This variation was driven by batch effects and not by case–control status (Fig. 1b). Any variation that is associated with processing batches, and not the sampling framework, must be due to contamination. A heat map (Fig. 1c) showed that eight out of the ten runs had a pronounced *Escherichia coli* signal (more than 20,000 reads in 64 samples, and 50–150 reads in 16 samples), a large collection of additional bacterial signals, and high levels of PhiX174 reads (group 1; Fig. 1c). Additional analyses mapping all *E. coli* reads from all samples together against the closest reference genome (WG5) showed that all *E. coli* reads belonged to the same strain (Extended Data Fig. 3) and are, therefore, due to contamination. All samples belonging to runs 4 and 5 (Fig. 1b) also had strong *Bradyrhizobium* and *Rhodopseudomonas palustris* signals (group 2 in PCA analysis). Runs 8 and 9 (group 3) lacked these strong signals. Two samples had strong human betaherpesvirus 6B (HHV-6B) signals (more than 10,000 read pairs; Fig. 1a–c), which reflected inheritance of the chromosomally integrated virus, affecting 0.5–1% of individuals in western populations[15].

We analysed the concordance between metagenomics and 16S rRNA gene amplicon sequencing in 79 samples from cohort 1 (Table 1, one 16S primer pair failed). The only signal consistently detected using both methods was *S. bongori*. An average of approximately 33,000 *S. bongori* reads (54% of total reads) were found by 16S rRNA amplicon sequencing (Extended Data Fig. 2b). *S. bongori* was not detected in the 16S negative controls (DNA extraction blanks; Table 1). The level of agreement between metagenomics and 16S rRNA for the other

[1]Wellcome Sanger Institute, Cambridge, UK. [2]Department of Veterinary Medicine, University of Cambridge, Cambridge, UK. [3]Department of Obstetrics and Gynaecology, University of Cambridge, National Institute for Health Research Biomedical Research Centre, Cambridge, UK. [4]Centre for Trophoblast Research (CTR), Department of Physiology, Development and Neuroscience, University of Cambridge, Cambridge, UK. [5]Department of Women's and Children's Health, Uppsala University, Uppsala, Sweden. [6]Department of Medicine, University of Cambridge, Cambridge, UK. [7]London School of Hygiene and Tropical Medicine, London, UK. [8]These authors contributed equally: Marcus C. de Goffau, Susanne Lager. [9]These authors jointly supervised this work: D. Stephen Charnock-Jones, Gordon C. S. Smith. *e-mail: jp369@cam.ac.uk; gcss2@cam.ac.uk

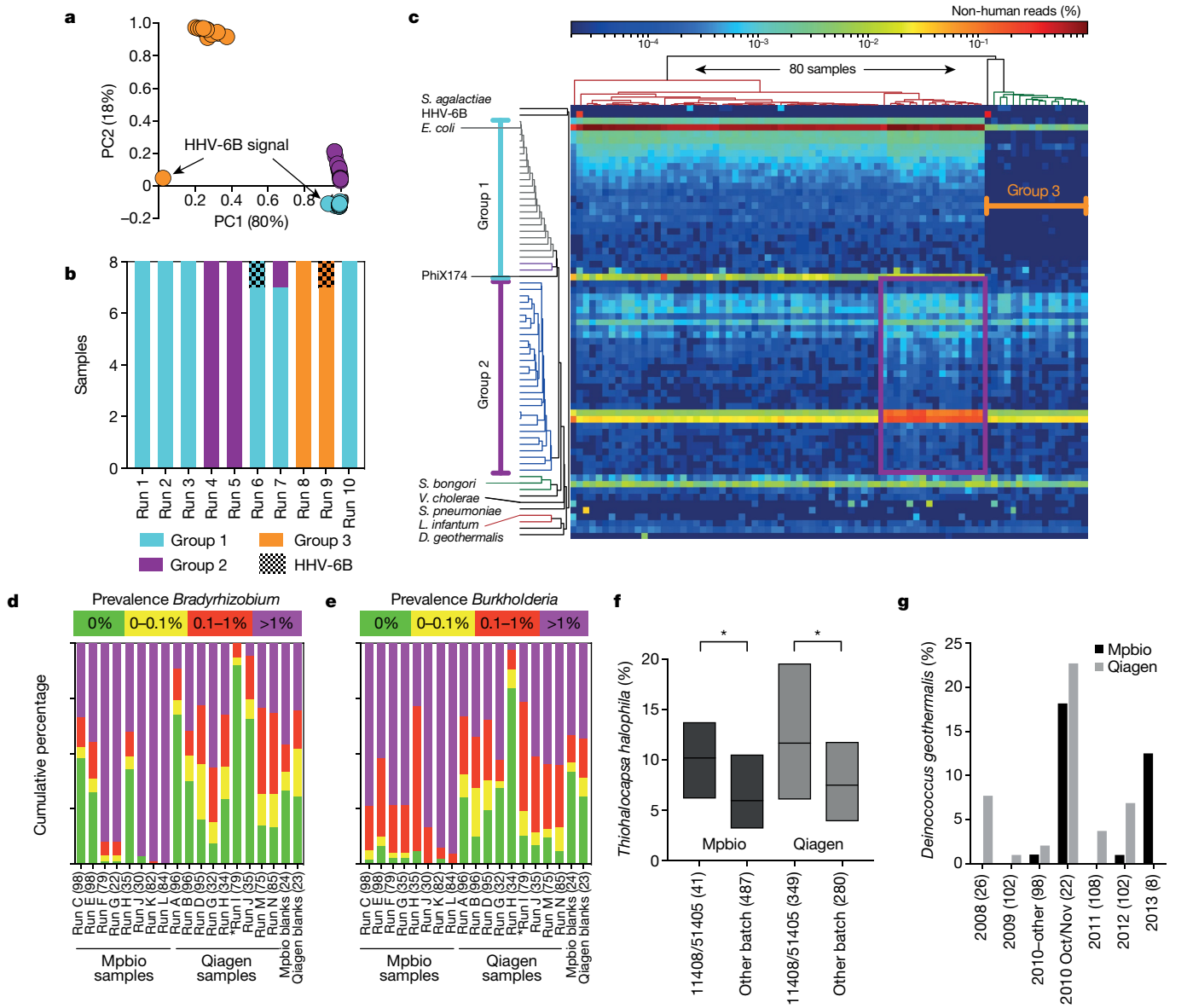


Fig. 1 | Batch effect detection in metagenomic and 16S rRNA amplicon sequencing data, cohort 1 samples. a–c, Summary of metagenomics data. **a**, PCA of summarized genus level identified by Kraken[25] output. **b**, MiSeq sequencing runs ($n = 8$ per run). **c**, Heat map of all non-human read abundance (see Extended Data Fig. 4). **d**, **e**, Read abundance by run and DNA isolation method (Mpbio or Qiagen) in chronological order for *Bradyrhizobium* (**d**) and *Burkholderia* (**e**). Scatterplots are shown in Extended Data Fig. 6. **f**, Associations between *Thiohalocapsa halophila* and Q5 buffer (lot 11508) or Taq polymerase (lot 51405). Interquartile range is shown; centre values denote medians. *$P < 0.001$ (Mann–Whitney *U*-test). **g**, *D. geothermalis* detection (>0.1% reads) by year of delivery. The number of samples in each group in **f** and **g** is shown in parentheses.

bacterial signals was assessed using the kappa statistic, scaled from 0 (no agreement) to 1 (perfect agreement). Only two signals demonstrated agreement (moderate-substantial) between the two methods: *S. agalactiae* and *Deinococcus geothermalis* (Table 1). The results were consistent when using different definitions of positive (Supplementary Table 3) and neither signal was detected in negative controls. The number of positive samples was too small for informative comparison of cases and controls.

Several bacterial signals associated with principal component 2, including the *Caulobacter*, *Methylobacterium* and *Burkholderia* genera, were also detected by 16S rRNA gene sequencing. However, the kappa statistics were low and these signals were also detected in negative controls (Table 1). *Vibrio cholerae* and *Streptococcus pneumoniae* signals were detected using metagenomics in 14 and 11 samples, respectively. However, neither was detected using 16S rRNA sequencing (Table 1). Assembly and analysis of these reads demonstrated that the closest matches were isolates from Bangladesh (PRJEB14661 *V. cholerae*) and the Global Pneumococcal Sequence Project (PRJEB31141 *S. pneumoniae*), which had been sequenced on the same pipelines at the Sanger Institute, indicating that these signals are due to cross-contamination during library preparation or sequencing (the same explanation applies for *Leishmania infantum*, Fig. 1c).

Cohort 2: duplicate 16S rRNA

By combining the data from two independent DNA isolation methods (the MP Biomedical kit, hereafter 'Mpbio', or Qiagen kit), we were able to visualize batch effects using PCA (Extended Data Fig. 5a) or visualize species individually (Fig. 1d–g) and analyse signal reproducibility. For example, *Bradyrhizobium* was detected nearly ubiquitously and in high abundance in some 16S rRNA sequencing runs, but was less frequently detected and in lower abundance in others (Fig. 1d, compare runs K and L with runs I and J). The *Burkholderia* genus, which has been suggested to have a role in PTB[3], had a higher signal in samples isolated using the Mpbio DNA isolation reagents than with the Qiagen kit, and also showed pronounced run-to-run variation (Fig. 1e). Furthermore, both *Bradyrhizobium* and *Burkholderia* were commonly detected in

Table 1 | Comparison of main signals using metagenomics with 16S rRNA amplicon sequencing

Species	Positive signals MG and 16S (79 = max)				MG reads in positive samples	16S reads in positive samples[a]	Concordance MG and 16S kappa score (*P* value)[b]	Part of an MG batch effect[c]	Presence 16S in negative controls Absent/weak/strong ($n = 5$)[d]
	Both[a]	MG only	16S only[a]	Neither					
Salmonella bongori	79	0	0	0	178	54%	NA	No	5/0/0
Escherichia coli	1	78	0	0	18,602	1.2%	0 (–)	Gr. 1 & 2	4/1/0
Shigella (genus)	0	75	0	4	254	NA	0 (–)	Gr. 1 & 2	5/0/0
Salmonella enterica	0	75	0	4	33	NA	0 (–)	Gr. 1 & 2	5/0/0
Cronobacter sakazakii	0	65	0	14	21	NA	0 (–)	Gr. 1 & 2	5/0/0
Bacillus subtilis	0	63	0	16	13	NA	0 (–)	Gr. 1 & 2	5/0/0
Yersinia pseudotuberculosis	0	59	0	20	3	NA	0 (–)	Gr. 1 & 2	5/0/0
Neisseria meningitidis	0	44	0	35	2	NA	0 (–)	Gr. 1 & 2	5/0/0
Bradyrhizobium (genus)	0	79	0	0	125	NA	0 (–)	Gr. 2	5/0/0
Rhodopseudomonas palustris	0	79	0	0	45	NA	0 (–)	Gr. 2	5/0/0
Caulobacter (genus)	12	67	0	0	14	1.4%	0 (–)	Gr. 2	1/3/1
Methylobacterium (genus)	9	69	0	1	8	2.4%	0.003 (0.36)	Gr. 2	1/4/0
Burkholderia (genus)	21	57	0	1	7	1.9%	0.009 (0.27)	Gr. 2	1/4/0
Propionibacterium acnes	66	13	0	0	20	4.8%	0 (–)	No	0/3/2
Streptococcus pneumoniae	0	11	0	68	115	NA	0 (–)	No	5/0/0
Vibrio cholerae	0	14	0	65	46	NA	0 (–)	No	5/0/0
Thiohalocapsa halophila	0	0	71	8	NA	4.2%	0 (–)	No	0/0/5
Stenotrophomonas maltophilia	5	51	1	22	2	1.9%	0.03 (0.24)	No	2/3/0
Acinetobacter baumanii	1	26	0	52	2	2.4%	0.05 (0.08)	No	4/1/0
Micrococcus luteus	1	46	0	32	15	2.0%	0.02 (0.20)	No	4/1/0
Gardnerella vaginalis	0	5	0	74	1	NA	0 (–)	No	4/1/0
Lactobacillus crispatus	0	4	0	75	1	NA	0 (–)	No	5/0/0
Deinococcus geothermalis	1	1	0	77	68	33%	0.66 (<0.0001)	No	5/0/0
Streptococcus agalactiae	3	4	0	72	8	13%	0.58 (<0.0001)	No	5/0/0

The average number of metagenomics (MG) and average percentage of 16S reads in positive samples are shown. Gr., group; NA, not applicable.
[a]16S rRNA amplicon sequencing signals higher than 1% are defined as positive.
[b]One-sided *P* values (kappa statistic).
[c]See Fig. 1 for definition of groups 1 and 2.
[d]Strong signals are defined as more than 1%.

the negative controls. Batch effects based on the use of particular polymerase chain reaction (PCR) reagent lots can also be visualized. For example, the association of *Thiohalocapsa halophila* with either the PCR reagent '5× Q5 buffer' (lot 11408) or 'Q5 Taq polymerase' (lot 51405), both of which were used to process the same 390 samples, is shown in Fig. 1f.

We used the kappa statistic to quantify the level of agreement between 16S rRNA amplicon sequencing of two DNA samples from the mode of delivery. $^*P < 0.05$, $^{***}P < 0.001$, Mann–Whitney *U*-tests were used where values below 1% are regarded as 0%. See Extended Data Fig. 6 for scatterplots. Percentage read count is based on the higher value for given species using Qiagen or Mpbio DNA isolation kit (using all 498 samples).

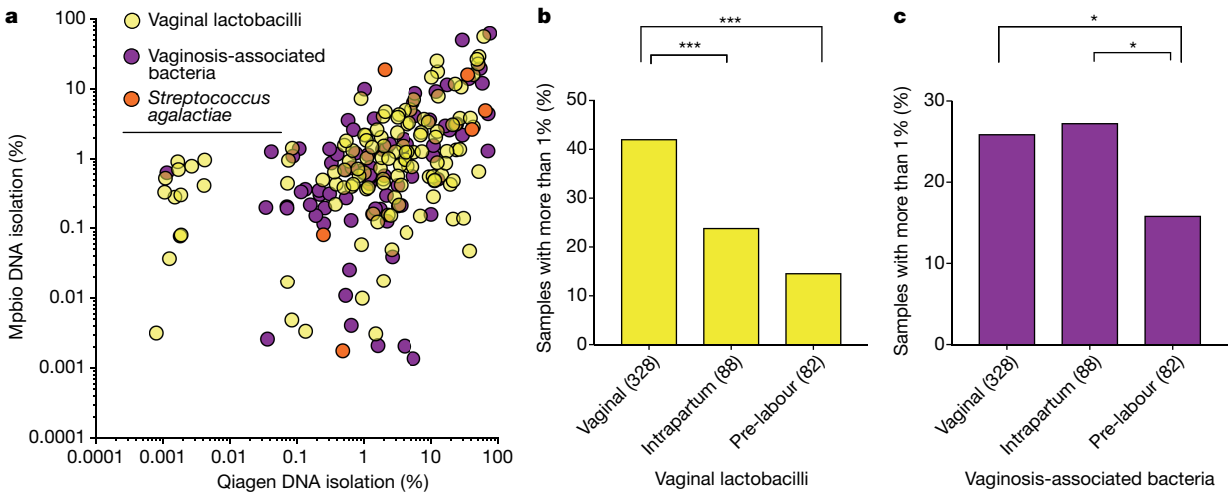


Fig. 2 | Mode of delivery and detection of vaginal bacteria by 16S rRNA amplicon sequencing. a, Concordant detection of vaginal lactobacilli and a combination of all vaginosis-associated bacteria by both Qiagen (*x* axis) and Mpbio (*y* axis) results in Spearman's rho correlation coefficients of 0.37 and 0.59, respectively, when analysing the top right quadrant only (>0.1%). **b**, **c**, Comparisons between vaginally associated bacteria and

Table 2 | Simplified overview on the nature of bacterial findings

	Signals					
	Independent of:					
	DNA extraction batch[a]	Date of delivery[b]	Mode of delivery	Not in negative controls[c]	Sample-associated[d]	Verified by meta-genomics[e]
Capable pathogens						
Streptococcus agalactiae	✓	✓	✓	✓	✓	✓
Listeria monocytogenes[f]	✓	✓	✓	✓	✓	–
Vaginal lactobacilli						
Lactobacillus crispatus	✓	✓	–	~	✓	~
Lactobacillus iners	✓	✓	–	~	✓	–
Lactobacillus gasseri	✓	✓	–	✓	✓	–
Lactobacillus jensenii	✓	✓	–	~	✓	–
Vaginosis-associated bacteria						
Gardnerella vaginalis	✓	✓	–	~	✓	–
Atopobium vaginae	✓	✓	–	~	✓	–
Ureaplasma (genus)	✓	✓	–	✓	✓	–
Prevotella bivia	✓	✓	–	~	✓	–
Prevotella amnii	✓	✓	–	✓	✓	–
Prevotella timonensis	✓	✓	–	~	✓	–
Aerococcus christensenii	✓	✓	–	✓	✓	–
Streptococcus anginosus	✓	✓	–	~	✓	–
Sneathia sanguinegens	✓	✓	–	✓	✓	–
Megasphaera elsdenii	✓	✓	–	~	✓	–
Faecal-associated bacteria						
Bacteroides (genus)	✓	✓	–	~	✓	–
Faecalibacterium prausnitzii	✓	✓	–	~	✓	–
Roseburia faeces	–	✓	–	~	✓&–	–
Coriobacterium sp.	✓	✓	–	~	✓	–
Collinsella intestinalis	✓	✓	–	+	✓	–
Suspected oral origin						
Fusobacterium nucleatum	✓	✓	✓	~	✓	–
Streptococcus mitis	✓	✓	✓	~	✓	–
Streptococcus vestibularis	–	✓	✓	~	✓&–	–
Genuine reagent contaminants						
Acinetobacter baumanii[f]	–	✓	✓	~	–	~
Thiohalocapsa halophila	–	✓	✓	–	–	–
Propionibacterium acnes	–	✓	✓	–	–	–
Stenotrophomonas maltophilia	–	✓	✓	–	–	–
Bradyrhizobium japonicum	–	✓	✓	–	–	–
Melioribacter roseus	–	✓	✓	–	–	–
Pelomonas (genus)	–	✓	✓	–	–	–
Methylobacterium (genus)	–	✓	✓	–	–	–
Aquabacterium (genus)	–	✓	✓	–	–	–
Sediminibacterium (genus)	–	✓	✓	–	–	–
Desulfovibrio alkalitolerans	–	✓	✓	–	–	–
Delftia tsuruhatensis	–	✓	✓	–	–	–
Streptococcus pyogenes	–	✓	✓	~	–	–
Burkholderia multivorans	–	✓	✓	–	–	–
Caulobacter (genus)	–	✓	✓	–	–	–
Steroidobacter sp. JC2953	–	✓	✓	–	–	–
Afipia (genus)	–	✓	✓	–	–	–
Burkholderia silvatlantica	–	✓	✓	–	–	–
Lysinimicrobium mangrove	–	✓	✓	–	–	–
Bradyrhizobium elkanii	–	✓	✓	–	–	–
Achromobacter xylosoxidans	–	✓	✓	–	–	–
Corynebacterium tuberculostearicum	–	✓	✓	–	–	–
Rhodococcus fascians	✓	–	✓	~	✓	–
Sphingobium rhizovicinum	✓	–	✓	~	✓	–
Methylobacterium organophilum	✓	–	✓	~	✓	–
Deinococcus geothermalis[f]	✓	–	✓	✓	✓	✓

[a]Includes batch effects caused by different DNA isolation kits, PCR reagents and MiSeq run.
[b]See Figs. 1g, 2d for details.
[c]A tick '✓' indicates absence; '~' indicates detection (any percentage) in less than 20% of negative controls.
[d]Detection of signal in corresponding Qiagen and Mpbio DNA isolations. '✓&–' indicates that signals from these operational taxonomic units are sample-associated in most 16S runs, but reagent contaminants in others. See Supplementary Table 4 for details.
[e]See Table 1 and Supplementary Table 3. A '~' indicates some level of concordance was detected using a different 16S threshold.
[f]The presence or absence of verification should be interpreted with caution, as indicated by examples.

the same patient extracted using the two different kits (Supplementary Table 4). The majority of the most-prevalent bacterial groups had low kappa scores and there was a low correlation between the magnitude of the signals comparing the two DNA extraction methods (Extended Data Fig. 5b). Moreover, these signals also demonstrated notable batch effects using PCA (Extended Data Fig. 5a). Interestingly, four ecologically unexpected bacterial groups of high prevalence exhibited a fair level of concordance (*Rhodococcus fascians*, *Sphingobium rhizovicinum*, *Methylobacterium organophilum* and *D. geothermalis*). Further analysis demonstrated a temporal pattern of these signals (Fig. 1g). All placental samples were washed in sterile PBS to remove surface contamination, such as maternal blood, and the temporal pattern of these bacterial signals is consistent with them being derived from batches of this reagent. Some ecologically plausible species, such as *S. agalactiae* and *Listeria monocytogenes*, vaginal lactobacilli, vaginosis-associated bacteria, faecal bacteria and some bacteria of probable oral origin had modest to high kappa scores, indicating that they were sample-associated signals. In contrast to the laboratory contaminants, the signals for these bacterial groups correlated when comparing the two DNA extraction methods (Fig. 2a) and were not associated with batch effects identifiable using PCA. Sample-associated signals (non-reagent contaminants) of a few species not typically associated with a vaginal or rectal habitat but with the oral habitat were detected, such as *Streptococcus mitis*, *Streptococcus vestibularis* and *Fusobacterium nucleatum*. However, it was only a very small minority of samples that exhibited these signals (below that of *S. agalactiae*) and none of these oral signals was identified by metagenomic analysis of pre-labour Caesarean section samples (cohort 1).

Delivery-associated signals

Vaginal organisms (lactobacilli and vaginosis-associated bacteria) were more abundant than *S. agalactiae* in cohort 2 (vaginal, intrapartum and pre-labour Caesarean section deliveries) but less abundant than *S. agalactiae* in cohort 1 (pre-labour Caesarean section deliveries only). Hence, we next examined the relationship between the mode of delivery and the 16S rRNA signal. Vaginal lactobacilli (*Lactobacillus iners*, *Lactobacillus crispatus*, *Lactobacillus gasseri* and *Lactobacillus jensenii*) were found more frequently and in higher numbers in vaginally delivered placentas than in placentas delivered via intrapartum or pre-labour Caesarean section (Fig. 2b), irrespective of the DNA isolation method (Extended Data Fig. 7a, b). Vaginosis-associated bacteria were found at approximately the same frequency in vaginal and intrapartum Caesarean section samples, but significantly less frequently in pre-labour Caesarean section samples (Fig. 2c). A heat map generated using the Spearman rho correlation coefficients of all abundant and relevant bacterial groups generated a cluster of vaginally associated bacteria, representative of vaginal community group IV[16], which reflects sample contamination during labour and delivery (Extended Data Fig. 8). The other clusters represented the contamination signatures of the two different DNA extraction kits and a fourth cluster reflected contamination associated with the date of collection of the placental biopsies (2012–2013).

Genuine signals and pregnancy outcome

The presence of *S. agalactiae* was analysed with respect to clinical outcome (SGA, pre-eclampsia, PTB) as it was the only organism that met all of the criteria of a genuine placenta-associated bacterial signal (Table 2). There was no association with SGA, pre-eclampsia or PTB (Fig. 3). Exploratory analysis of the 16S amplicon sequencing data of all sample-associated signals, including delivery-associated bacteria, showed that *S. mitis* and *F. nucleatum* were not associated with adverse pregnancy outcome (Supplementary Table 5). Of note, however, were the significant associations of the delivery-associated bacteria *L. iners* with pre-eclampsia and *Streptococcus anginosus* and the *Ureaplasma* genus with PTB (Fig. 3, Supplementary Table 5 and Extended Data Fig. 9). In one placental sample from a preterm birth, a strong

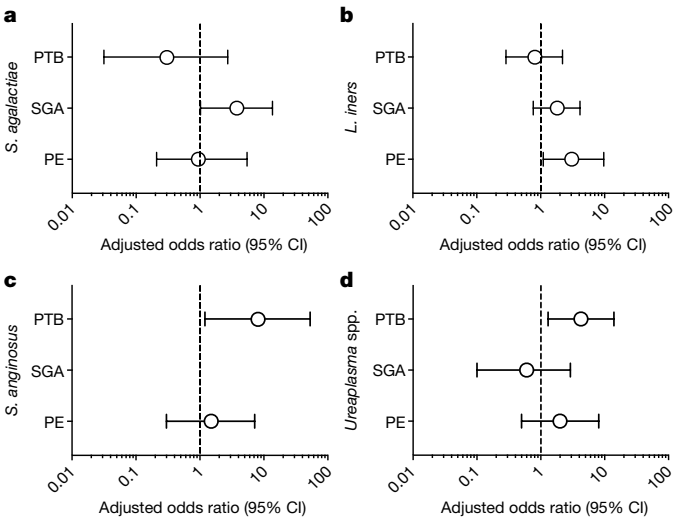


Fig. 3 | Bacterial signals and adverse pregnancy outcome. a–d, Adjusted odds ratios for the association of *S. agalactiae* (**a**), *L. iners* (**b**), *S. anginosus* (**c**) and *Ureaplasma* spp. (**d**) with PTB, SGA and pre-eclampsia (PE). Pre-eclampsia and SGA both had 100 matched cases and controls. The PTB analysis included 56 preterm cases and 136 unmatched controls (all vaginally delivered). Odds ratios were adjusted for clinical characteristics by logistic regression. The odds ratio and its confidence interval (CI) cannot be calculated for *S. anginosus* and SGA because one of the discordant values is zero. See Supplementary Table 5 for further details.

L. monocytogenes signal was found (7% and 52% of all reads with Mpbio and Qiagen, respectively).

Validating *Streptococcus agalactiae*

A nested PCR and quantitative PCR (qPCR) approach targeted towards the *sip* gene, which encodes the surface immunogenic protein (SIP) of *S. agalactiae*, was used to verify its presence in 276 placental samples for which a 16S sequencing result was available. In total, 7 out of 276 samples were positive using PCR–qPCR and all seven were also positive (more than 1%) by 16S analysis. A total of 14 samples were positive by 16S sequencing but not by PCR–qPCR, no sample was positive using PCR–qPCR and negative by 16S, and 255 samples were negative by both methods. This yielded a kappa statistic of 0.48, indicating moderate agreement and a P value of 9.7×10^{-21}. We conclude that the detection of *S. agalactiae* by 16S rRNA amplification was verified by two further independent methods (metagenomics and PCR–qPCR) and the level of agreement in both cases was well above what could be expected by chance. It remains to be determined why some samples were positive for *S. agalactiae* by 16S sequencing but negative by the PCR–qPCR method. Generally, the latter would be considered more sensitive, particularly in samples with a higher microbial biomass, owing to the complex amplification kinetics when a large number of diverse 16S template molecules are present. However, in the absence of other bacterial signals, it is possible that 16S sequencing is more sensitive for detecting very small numbers of *S. agalactiae*, as the genome of the organism has seven copies of the 16S rRNA gene, but only one copy of *sip*[17].

Discussion

We studied placental biopsies from a total of 537 women, including 318 cases of adverse pregnancy outcome and 219 controls, using multiple methods of DNA extraction and detection, and drew several important conclusions. First, we found that the biomass of bacterial sequences in DNA extracted from human placenta was extremely small. Second, the major source of bacterial DNA in the samples studied was contamination from laboratory reagents and equipment. Third, both metagenomics and 16S amplicon sequencing were capable of detecting a very low amount of a spiked-in signal. Fourth, samples of placental tissue

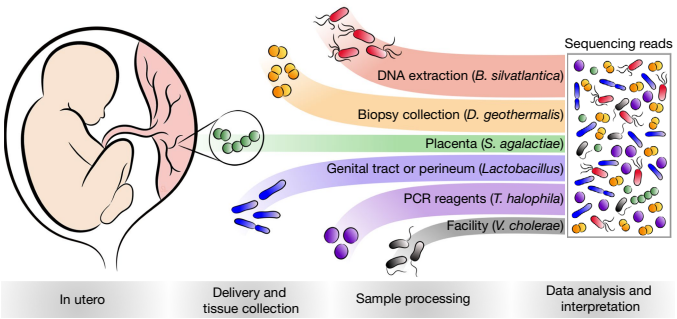


Fig. 4 | Sources of bacterial signals detected in human placental samples. Bacteria may sometimes be present in utero, such as *S. agalactiae*. Bacteria or bacterial DNA also frequently contaminate the placenta during labour and delivery (for example, *Lactobacillus*), during sample collection (for example, *D. geothermalis*), and during sample processing (for example, *B. silvatlantica* and *T. halophila*). Contamination may also occur during library preparation or sequencing from other projects carried out at the facility (for example, *V. cholerae* in the metagenomic sequencing).

become contaminated during the process of labour and delivery, even when they were dissected from within the placenta. Finally, the only organism for which there was strong evidence that it was present in the placenta before the onset of labour was *S. agalactiae*. It was not part of any batch effect, it was detected by three methods, there was a statistically significant level of agreement between 16S amplicon sequencing and both metagenomics ($P = 1.5 \times 10^{-8}$) and a targeted PCR–qPCR assay ($P = 9.7 \times 10^{-21}$), none of 47 negative controls analysed by 16S sequencing was positive for *S. agalactiae*, and there was no association with mode of delivery (Extended Data Fig. 7). However, there was no significant association between the presence of the organism and pre-eclampsia, SGA or PTB. Exploratory analysis of other signals did demonstrate an association between PTB and the presence of *Ureaplasma* reads (>1%), consistent with previous studies[18], but this was probably the result of ascending uterine infection. We conclude that bacterial placental infection is not a major cause of placentally related complications of human pregnancy and that the human placenta does not have a resident microbiome.

The finding of *S. agalactiae* in the placenta before labour could be of considerable clinical importance. Perinatal transmission of *S. agalactiae* from the mother's genital tract can lead to fatal sepsis in the infant. It is estimated that routine screening of all pregnant women for the presence of *S. agalactiae* and targeted use of antibiotics prevents 200 neonatal deaths per year in the United States[19]. Our findings identify an alternative route for perinatal acquisition of *S. agalactiae*. Further studies will be required to determine the association between the presence of the organism in the placenta and fetal or neonatal disease. However, if such a link was identified, rapid testing of the placenta for the presence of *S. agalactiae* might allow targeting of neonatal investigation and treatment. Our work also sheds light on the possible routes of fetal colonization. Although we see no evidence of a placental microbiome, the frequency of detection of vaginal bacteria in the placenta increased after intrapartum Caesarean section, suggesting ascending or haematogenous spread. Similarly, haematogenous spread as the result of transient bacteraemia could potentially explain the presence of the small number of sample-associated oral bacterial signals[14]. Such spread could lead to fetal colonization immediately before delivery.

We identified five different patterns of contamination (Fig. 4)—namely, contamination of the placenta with real bacteria during the process of labour and delivery (Fig. 2); contamination of the biopsy when it was washed with PBS; contamination of DNA during the extraction process; contamination of reagents used to amplify the DNA before sequencing; and contamination from the reagents or equipment used for sequencing. Using 16S rRNA amplicon sequencing, the positive control (*S. bongori*) accounted for more than half of the reads, indicating that the method is highly sensitive. However, when the method is applied to samples with little or no biomass, these sources of contamination can lead to apparent signals, hence it is crucial to use a method that allows differentiation between true bacterial signals and these sources of contamination (see Supplementary Information 1 for further technical discussion).

In conclusion, in a study of 537 placentas carefully collected, processed and analysed to detect real bacterial signals, we found no evidence to support the existence of a placental microbiome and no significant relationship between placental infection with bacteria and the risk of pre-eclampsia, SGA and preterm birth. However, we identified an important pathogen, *S. agalactiae*, in the placenta of approximately 5% of women before the onset of labour.

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/s41586-019-1451-5.

Received: 16 January 2019; Accepted: 28 June 2019;
Published online 31 July 2019.

1. Brosens, I., Pijnenborg, R., Vercruysse, L. & Romero, R. The "Great Obstetrical Syndromes" are associated with disorders of deep placentation. *Am. J. Obstet. Gynecol.* **204**, 193–201 (2011).
2. Aagaard, K. et al. The placenta harbors a unique microbiome. *Sci. Transl. Med.* **6**, 237ra65 (2014).
3. Antony, K. M. et al. The preterm placental microbiome varies in association with excess maternal gestational weight gain. *Am. J. Obstet. Gynecol.* **212**, 653.e1–653.e16 (2015).
4. Collado, M. C., Rautava, S., Aakko, J., Isolauri, E. & Salminen, S. Human gut colonisation may be initiated in utero by distinct microbial communities in the placenta and amniotic fluid. *Sci. Rep.* **6**, 23129 (2016).
5. Perez-Muñoz, M. E., Arrieta, M. C., Ramer-Tait, A. E. & Walter, J. A critical assessment of the "sterile womb" and "in utero colonization" hypotheses: implications for research on the pioneer infant microbiome. *Microbiome* **5**, 48 (2017).
6. Salter, S. J. et al. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol.* **12**, 87 (2014).
7. Jervis-Bardy, J. et al. Deriving accurate microbiota profiles from human samples with low bacterial content through post-sequencing processing of Illumina MiSeq data. *Microbiome* **3**, 19 (2015).
8. de Goffau, M. C. et al. Recognizing the reagent microbiome. *Nat. Microbiol.* **3**, 851–853 (2018).
9. Lauder, A. P. et al. Comparison of placenta samples with contamination controls does not provide evidence for a distinct placenta microbiota. *Microbiome* **4**, 29 (2016).
10. Leiby, J. S. et al. Lack of detection of a human placenta microbiome in samples from preterm and term deliveries. *Microbiome* **6**, 196 (2018).
11. Theis, K. R. et al. Does the human placenta delivered at term have a microbiota? Results of cultivation, quantitative real-time PCR, 16S rRNA gene sequencing, and metagenomics. *Am. J. Obstet. Gynecol.* **220**, 267.e1–267.e39 (2019).
12. Leon, L. J. et al. Enrichment of clinically relevant organisms in spontaneous preterm delivered placenta and reagent contamination across all clinical groups in a large UK pregnancy cohort. *Appl. Environ. Microbiol.* **84**, e00483-e18 (2018).
13. Sovio, U., White, I. R., Dacey, A., Pasupathy, D. & Smith, G. C. S. Screening for fetal growth restriction with universal third trimester ultrasonography in nulliparous women in the Pregnancy Outcome Prediction (POP) study: a prospective cohort study. *Lancet* **386**, 2089–2097 (2015).
14. Hornef, M. & Penders, J. Does a prenatal bacterial microbiota exist? *Mucosal Immunol.* **10**, 598–601 (2017).
15. Leong, H. N. et al. The prevalence of chromosomally integrated human herpesvirus 6 genomes in the blood of UK blood donors. *J. Med. Virol.* **79**, 45–51 (2007).
16. Ravel, J. et al. Vaginal microbiome of reproductive-age women. *Proc. Natl Acad. Sci. USA* **108** (suppl. 1), 4680–4687 (2011).
17. Glaser, P. et al. Genome sequence of *Streptococcus agalactiae*, a pathogen causing invasive neonatal disease. *Mol. Microbiol.* **46**, 1499–1513 (2002).
18. Abele-Horn, M., Scholz, M., Wolff, C. & Kolben, M. High-density vaginal *Ureaplasma urealyticum* colonization as a risk factor for chorioamnionitis and preterm delivery. *Acta Obstet. Gynecol. Scand.* **79**, 973–978 (2000).
19. Schrag, S. J. et al. Group B streptococcal disease in the era of intrapartum antibiotic prophylaxis. *N. Engl. J. Med.* **342**, 15–20 (2000).

METHODS

Ethics. This study is in compliance with all relevant ethical regulations. The Pregnancy Outcome Prediction study (POPs) was approved by the Cambridgeshire 2 Research Ethics Committee (reference number 07/H0308/163). The study and the characteristics of the eligible and participating women have been previously described in detail[13,20]. In brief, 4,212 nulliparous women with a singleton pregnancy were followed through from their first ultrasound scan to delivery. At the time of delivery, placental samples were obtained using a standardized protocol by a team of trained technicians, in which most samples were obtained within 3 h of delivery (interquartile range: 0.3–8.4 h). All participants gave written informed consent for the study and for subsequent analysis of their samples.

Patient selection. For cohort 1, cases of SGA (≤fifth percentile based on customized birth weight[21]; $n = 20$) or pre-eclampsia (according to the 2013 ACOG (The American College of Obstetricians and Gynaecologists) Guidelines[22]; $n = 20$) were matched one-to-one with healthy controls ($n = 40$). Only deliveries by pre-labour Caesarean section were included in this cohort. The cases and controls were matched as closely as possible for maternal body mass index, maternal age, gestational age, sample collection time, maternal smoking, and fetal sex. Clinical characteristics are presented in Supplementary Table 1.

For cohort 2, cases of SGA (≤fifth customized birth weight percentile[21]; $n = 100$) or pre-eclampsia (2013 ACOG guidelines[22]; $n = 100$) were selected. The cases were matched one-to-one with healthy controls ($n = 198$, two controls were used twice). All deliveries were at term (≥37 weeks gestation). The same matching criteria as in the first cohort were used with the addition of an absolute match for mode of delivery. Placentas from 100 preterm births (<37 weeks gestation) deliveries were also included in the study (clinical characteristics in Supplementary Table 2). Flow charts describing the two cohorts and subsequent sample-processing and analysis steps are presented in Extended Data Fig. 1.

Placenta collection. Placentas were collected after delivery and the procedure has previously been described in detail[20]. We confined our sampling to the placental terminal villi (fetal tissue). We chose this as the villi are the site of exchange, across the vasculosyncytial membrane, between the fetus and mother. This location is the closest interface between the fetus with the mother's blood and tissues. If the placenta was colonized, one would expect bacteria to ascend the genital tract (local infiltration) or to come from the mother's blood (haematogenous). Hence, we believe that this would be the most plausible site for bacteria to be found. Villous tissue was obtained from four separate lobules of the placenta after trimming to remove adhering decidua from the basal plate. The tissue in the selected areas had no visible damage, haematomas, or infarctions. To remove maternal blood, the selected tissue samples were rinsed in chilled sterile PBS (Oxoid Phosphate Buffered Saline Tablets, Dulbecco A; Thermo Fisher Scientific) dissolved in ultrapure water (ELGA Purelab Classic 18MΩ.cm). After initial collection, all placental samples were frozen in liquid nitrogen and stored at −80 °C until further processing. For DNA isolation, approximately 25 mg of villous tissue (combined weight obtained from fragments of all four biopsy collection points) was cut from the stored tissue. To reduce the risk of environmental contamination of the samples, the entire experimental procedure was carried out in a class 2 biological safety cabinet (tissue cutting, DNA isolation, setting up PCR reactions). The tissue was cut with single-use sterile forceps and scalpel. Each matched case–control pair was processed in parallel on the same day for each step of the entire experimental procedure (tissue cutting, DNA isolation, setting up PCR reactions). Also, the same lot of laboratory reagents was used for each pair. For each lot of laboratory reagents, negative controls were included (described in detail below).

DNA isolation from cohort 1. DNA was isolated from placental tissue with the Qiagen Qiaamp DNA mini kit (51304; Qiagen) according to the manufacturer's instructions with the addition of a freeze–thaw cycle after the overnight tissue lysis. Before DNA isolation, intact *S. bongori* was added to the placental tissue (1,100 CFUs, described in detail below). The placental tissue with added *S. bongori* was lysed in a proteinase-K-based solution (100 µl buffer ATL (Qiagen), 80 µl of *S. bongori*, 20 µl proteinase K) overnight (18 h at 56 °C) and thereafter freeze–thawed once. After the thawed samples were brought to room temperature, RNA was removed with the addition of 4 µl RNase A (Qiagen, 19101) and incubated at room temperature for 2 min. Spin-filtering and washing of the DNA was carried out according to the manufacturer's instructions. The DNA was eluted from the spin column with 200 µl buffer AE (Qiagen) after a 5 min incubation (the elution step was repeated once with another 200 µl buffer AE and 5 min incubation). To prevent accidental cross-contamination between samples, gloves were changed between handling each sample. Throughout the protocol (DNA extraction, primer aliquoting, 16S rRNA gene amplification and library preparation), nuclease-free plastics were used (unless supplied with kit): PCR clean 2.0 and 1.5 ml DNA LoBind Tubes (Eppendorf), and nuclease-free filter tips (TipONE sterile filter tips, STARLAB). For each box of DNA isolation kit used, extraction blanks were carried out. These DNA extraction blanks, or negative controls, contained only the reagents from each DNA isolation kit (no added biological material) and were subjected to the complete DNA extraction procedure: tissue homogenization, matrix binding, spin-filtering, washing, and elution of nucleic acids. The negative controls were subjected to the entire analysis protocol alongside the placental samples: DNA isolation, 16S rRNA gene PCR amplification, sequencing and data analysis.

Positive control. As a positive control, a known amount of intact *S. bongori* (strain NCTC-12419) was added to each of the placental tissue samples in cohort 1 before DNA isolation ($n = 80$). *S. bongori* was incubated with shaking overnight at 37 °C in LB broth. When the OD_{600} reached 0.9 (approximately equivalent to 7.2×10^8 bacteria per ml, measured with a Ultrospec 10 Cell Density Meter, GE Healthcare) the culture was chilled on ice. To minimize bacterial growth outside of the shaking incubator, all cultures and dilutions were kept on ice. To increase the proportion of live bacteria added as positive controls, 1 ml of the *S. bongori* suspension was diluted in 14 ml fresh LB broth (OD_{600} was 0.06) and incubated with shaking (1.5 h at 37 °C; OD_{600} was 0.8). The *S. bongori* culture was then serially diluted to an estimated concentration of 1,000 *S. bongori* per 80 µl, which was used to spike the placental samples. To determine the actual number of CFUs added to the placental samples, the *S. bongori* suspension was further diluted and aliquots cultured on LB plates overnight (37 °C). The number of colonies was counted. On the basis of three plates with distinct individual colonies (between 29 and 205 colonies per plate), the number of *S. bongori* added to each placental tissue sample was calculated to be 1,100 CFUs.

DNA isolation from cohort 2. DNA was isolated twice from each placenta using two different extraction kits. The DNA isolations were carried out in accordance with respective manufacturer's instructions, with the addition of two extra washes in the MP Biomedical kit.

For the Qiagen Qiaamp DNA mini kit (Qiagen, 51304), the placental tissue was digested in a proteinase-K-based solution (100 µl buffer ATL, 80 µl PBS, 20 µl proteinase K) for at least 3 h. Then, 4 µl of RNase A (Qiagen, 19101) was added to the tissue lysate and incubated at room temperature for 2 min. Spin-filtering and washing of the DNA was carried out according to the manufacturer's instructions. The DNA was eluted from the spin column with 200 µl buffer AE after a 5 min incubation (the elution step was repeated once with another 200 µl buffer AE and 5 min incubation).

For the MP Biomedical Fast DNA Spin kit (MP Biomedical, 116540600), the placental tissue was homogenized in 1.0 ml of CLS-TC solution by bead-beating (Lysing Matrix A tubes, 40 s, speed 6.0 on a FastPrep-24, MP Biomedical). After spinning the samples, equal volumes of the supernatant were combined with Binding Matrix. The mixture was transferred to a spin filter, after spin filtering the DNA was washed three times with SEWS-M. The DNA was eluted by re-suspending the Binding Matrix in 100 µl DES buffer, incubating the tubes at 55 °C for 5 min before recovering the DNA by centrifugation.

The same measures to prevent contamination of the samples as described in the cohort 1 DNA isolation section were taken. Extraction blanks were generated for each box/lot of both DNA isolation kits in a similar manner as was done for cohort 1. DNA concentrations were determined by Nanodrop Lite (Thermo Fisher Scientific).

Metagenomic sequencing. Sample processing for the metagenomics analysis was performed exactly as previously described[23]. In brief, the NEB Ultra II custom kit (New England Biolabs) was used for library generation, and samples were then sequenced on the Illumina HiSeq X Ten platform (150 base pairs, paired end) in 10 runs (flowcells) of 8 samples (lanes) each. The sequencing coverage was designed to generate more than 30-fold coverage of the human chromosomal DNA in each sample.

16S rRNA gene amplification. For detection of the bacterial 16S rRNA gene, PCR amplification of the V1–V2 region was performed using V1 primers with four degenerate positions to optimize coverage as previously recommended[24]. The V1–V2 amplicon is relatively short (~260 bp) and, with paired-end reads, almost all of the amplified product is sequenced on both strands and thus at higher accuracy. This is not the case with the longer V1–V3 amplicon. This region has also been used in other studies of the placental microbiome[10]. The following barcoded primers were used forward-27: 5′-AATGATACGGCGACCACCGAGATCTACAC*nnnnnnnnnnnn*ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNAGMGTTYGATYMTGGCTCAG-3′ and reverse-338: 5′-CAAGCAGAAGACGGCATACGAGAT *nnnnnnnnnnnn* GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTNNNNG CTGCCTCCCGTAGGAGT. The *n*-string represents unique 12-mer barcodes used for each sample studied and distinct indexes were used at both the 5′ and 3′ ends of the amplicons. The primers were purchased from Eurofins Genomics. Before aliquoting, the cabinet and pipettes were cleaned with DNA AWAY Surface Decontaminant. The primers were diluted in Tris-EDTA buffer (Sigma-Aldrich) in PCR clean nuclease-free DNA LoBind Tubes (Eppendorf) with nuclease-free filter tips (TipONE sterile filter tips, STARLAB). The PCR amplification was carried out in quadruplicate reactions for each sample on a SureCycler 8800 Thermal Cycler (Agilent Technologies) with high-fidelity Q5 polymerase (M0491L; New

England Biolabs), dNTP solution mix (N0447L, New England Biolabs), and UltraPure DNase/RNase-Free Water (Thermo Fisher Scientific) in 0.2 ml PCR strips (STARLAB). Amplification was performed with 500 ng DNA per reaction, and the final primer concentration was 0.5 μM. The PCR amplification profile was an initial step of 98 °C for 2 min followed by 10 cycles of touch-down (68 to 59 °C; 30 s), and 72 °C (90 s), followed by 30 cycles of 98 °C (30 s), 59 °C (30 s), and 72 °C (90 s). After completion of cycling, the reactions were incubated for 5 min at 72 °C. After completion of the PCR, the four replicates of each sample were pooled, cleaned up with AMPure XP beads (A63881; Beckman Coulter) and eluted in Tris-EDTA buffer (Sigma-Aldrich). DNA concentration was determined by Qubit Fluorometric Quantitation (Q32854; Invitrogen). Equimolar pools of the PCR amplicons were run on 1% agarose/TBE gels and ethidium bromide used to visualize the DNA. The DNA bands were excised and cleaned up with a Wizard SV Gel and PCR Clean-Up System (Promega UK). The equimolar pools were sequenced on the Illumina MiSeq platform using paired-end 250 cycle MiSeq Reagent Kit V2 (Illumina).

Bioinformatic analysis of metagenomics data. Bioinformatic analysis first required removal of human reads followed by identification of the species of non-human reads. KneadData (http://huttenhower.sph.harvard.edu/kneaddata) is a tool designed to perform quality control on metagenomic sequencing data, especially data from microbiome experiments, and we used this to remove the human reads. Forward and reverse reads from each sample were filtered using KneadData (v.0.6.1) with the following trimmomatic options: HEADCROP9, SLIDINGWINDOW:4:20, MINLEN: 100. A custom Kraken[25] reference database (v.0.10.6) was built, using metagm_build_kraken_db and -max_db_size 30, to detect any bacterial, viral and potential non-human eukaryotic signals. This custom Kraken reference database included both the default bacterial and viral libraries, and an accessions.txt file was supplied (via -ids_file) containing a diverse array of organisms chosen from all sequenced forms of eukaryotic life (see Supplementary Table 6 for accession numbers). This wide array was chosen to both detect potentially relevant unknown organisms, but also to identify additional human reads that had not been mapped to the human reference genome. In the metagenomic data, various non-human eukaryotic signals were identified by Kraken in every placental sample at a similar percentage, and were mostly assigned to *Pan paniscus* (Supplementary Table 6). As a verification, reads mapping to eukaryotic species were extracted (Supplementary Information 1) and contigs were assembled. These were analysed using BLASTN and were indeed identified as human. This indicates that these (often lower quality or repetitive) eukaryotic reads are in fact human reads that were not removed by mapping against the human reference genome. An exception to this was that in 17 samples an elevated number of reads were assigned to *Danio rerio* (zebrafish) and *Sarcophilus harrisii* (Tasmanian devil), both of which had been sequenced on the Sanger Institute pipeline. Kraken was run using the metagm_run_kraken option. All human-derived signals (eukaryotic non-fungal reads found in every placental sample at a similar percentage) were removed before further analysis. See Source Data of Fig. 1a–c for abundance information. The origins of *Streptococcus pneumonia* and *Vibrio cholerae* reads were analysed by extracting their respective reads as identified by the Kraken using custom scripts (Supplementary Information 1), performing an assembly on these reads using Spades (v.3.11.0)[26] and by using BLAST (blastn, database: others)[27] to find the closest match. The first step of the strain level analysis of *E. coli* reads to find the closest *E. coli* reference genome match was identical to the steps described above. Subsequently, *E. coli* reads were mapped against *E. coli* WG5 (GenBank: CP02409.1) using BWA (v.0.7.17-r1188)[28] and visualized using Artemis (v.16.0.0)[29]. *E. coli* reads were both analysed per sample and by combining all *E. coli* reads from all samples together.

Bioinformatic analysis of 16S rRNA gene amplicon data. To analyse all 14 16S rRNA amplicon data together using the MOTHUR (v.1.40.5) MiSeq SOP[30] and the Oligotyping (v.2.1) pipeline[31], the data from each individual run were initially individually processed in the MOTHUR pipeline as described below. All of the reads need to be aligned together as a requirement of the Oligotyping pipeline so after the most memory intensive-filtering steps had been performed, they were combined and processed again. Modifications to the MOTHUR MiSeq SOP are as follows: the 'make.contigs' command was used with no extra parameters on each individual run. The assembled contigs were taken out from the MOTHUR pipeline and the four poly NNNNs present in the adaptor/primer sequences were removed using the '-trim_left 4' and '-trim_right 4' parameters in the PRINSEQ-lite (v.0.20.3) program[32]. The PRINSEQ trimmed sequences were used for the first 'screen.seqs' command to remove ambiguous sequences and sequences containing homopolymers longer than 6 bp. In addition, any sequences longer than 450 bp or shorter than 200 bp were removed. Unique reads ('unique.seqs') were aligned ('align.seqs') using the Silva bacterial database 'silva.nr_v123.align'[33] with flip parameter set to true. Any sequences outside the expected alignment coordinates ('start=1046', 'end=6421') were removed. The correctly aligned sequences were subsequently filtered ('filter.seqs') with 'vertical=T' and 'trump='. The filtered sequences were de-noised by allowing three mismatches in the "pre.clustering" step and chimaeras were removed using Uchime with the dereplicate option set to 'true'. The chimaera-free sequences were classified using the Silva reference database 'silva.nr_v123.align' and the Silva taxonomy database 'silva.nr_v123.tax' and a cut-off value of 80%. Chloroplast, mitochondria, unknown, archaea, and eukaryota sequences were removed. All reads from each sample were subsequently renamed, placing the sample name of each read in front of the read name. The 'deunique.seqs' command, which creates a redundant fasta file from a fasta and name file, was performed before concatenating all of the data of all 14 16S runs together using the 'merge.files' command, which was done on both the fasta and the group files. The 'unique.seqs' command was again used before again aligning all reads as described previously before finishing the MOTHUR pipeline with the 'deunique.seqs' command.

Oligotyping and species identification. After the MOTHUR pipeline, the redundant fasta file, which now only contains high-quality aligned fasta reads, was subsequently used for oligotyping using the unsupervised minimum entropy decomposition (MED) for sensitive partitioning of high-throughput marker gene sequences[31]. A minimum substantive abundance of an oligotype (-M) was defined at 1,000 reads and a maximum variation allowed (-V) was set at 3 using the command line 'decompose 14runs.fasta -M 1000 -V 3 -g –t'. The node representative sequence of each oligotype (OTP) was used for species profiling using the ARB program (v.5.5-org-9167)[34]. For ARB analysis, we used a customized version of the SILVA SSU Ref database (NR99, release 123) that was generated by removing uncultured taxa. Oligotype abundances are provided in Supplementary Information 2 and additional metadata, for example, contamination identification via PCA (Extended Data Fig. 3), is provided in the Source Data.

Sensitivity analysis. To compare 16S rRNA amplicon sequencing and metagenomics sensitivity, the *S. bongori* signals (positive control) spiked into cohort 1 were analysed (Extended Data Fig. 2a, b). In 16S rRNA amplicon sequencing analysis 1,100 CFUs of *S. bongori* resulted in an average of 33,000 *S. bongori* reads (~54%). Thus, the remaining bacterial signal (reagent contamination background plus other signals) contributes the remaining 46% of the reads. This is approximately equivalent to another 937 *S. bongori* CFUs (1,100/(54/46)). Thus, if there are 937 bacteria in the sample (everything except the spike), this should produce a signal of 100% when there are no spiked-in bacteria present. Thus, the sensitivity of this assay in cohort 2, which did not contain a spike, is 0.106% of sequencing reads per CFUs (100%/937 CFUs). However, although an average of 54% *S. bongori* reads were detected in all spiked samples, it can be reasoned that samples with the highest *S. bongori* percentages only have reagent contamination DNA to compete with during the PCR step and not any other sample-associated signals. *S. bongori* percentages in the top 20th percentile on average account for 71% of all reads, which would correspond to a sensitivity limit of ~0.2% of reads per CFU (100/(1,100/(71/29)). However, a threshold of 1%, as previously used[9], can be considered a more reliable cut-off for determining whether a signal should be considered biologically relevant. A threshold of 1% would be indicative of multiple replication events (more than 2) and thus metabolic activity or repeated invasion of the tissue by the respective organism. In addition, a 1% threshold for the 16S rRNA data is comparable with the sensitivity of metagenomics as on average 180 *S. bongori* read pairs were detected with metagenomics (Extended Data Fig. 2a). In contrast to 16S analysis, the *S. bongori* spike has no meaningful effect on quantification in metagenomics as microorganisms only represent a very small fraction of the total amount of reads (the vast majority of reads are human). Hence, 6 CFUs are required on average per metagenomics read pair and 6 CFUs would result in a signal of approximately 1% of 16S amplicon reads in cohort 2 using the Qiagen kit.

Nested PCR. We developed a nested PCR assay to sensitively detect the *S. agalactiae sip* gene. In total, 276 placental DNA samples (isolated with the Qiagen kit as described above) were used of which 226 had no (0%) *S. agalactiae* reads detected by 16S rRNA gene sequencing, while *S. agalactiae* reads were detected in 50 samples (range 0.002–63.37% of 16S rRNA reads). The first-round PCR was performed using the DreamTaq PCR Master Mix (2×) (K1071; Thermo Fisher Scientific) and the following primers for the *sip* gene at a final concentration of 0.5 μM: forward 5′-TGAAAATGAATAAAAAGGTACTATTGACAT-3′ and reverse 5′-AAGCTGGCGCAGAAGAATA-3′. Amplification was performed in 50-μl aliquots and using 500 ng of placental DNA per reaction. Genomic *S. agalactiae* DNA (ATCC BAA-611DQ) was used as positive control at 20 or 2 copies per reaction. One reaction was set up with water instead of gDNA as negative control. The PCR amplification profile had an initial step of 95 °C for 3 min followed by 15 cycles of 95 °C (30 s), 48 °C (30 s), and 72 °C (60 s). After completion of cycling, the reactions were incubated for 3 min at 72 °C. The second-round qPCR was performed using the TaqMan Multiplex Master Mix (4461882; Thermo Fisher Scientific) and two TaqMan Assays (Thermo Fisher Scientific): Ba04646276_s1 (Gene Symbol: SIP; Dye Label, Assay Concentration: FAM-MGB, 20×) at a final 1× concentration; RNase P TaqMan assay (ABY dye/QSY probe Thermo Fisher Scientific 4485714) at a final 0.5× concentration, added as a positive control for the

human DNA. In each well, 6 μl of the first-round PCR (or water in the no template control/blank wells) was used as the reaction substrate in a total volume of 15 μl. The PCR amplification profile had an initial step of 95 °C for 20 s followed by 40 cycles of 95 °C (5 s) and 60 °C (20 s).

Statistics. The inter-rater agreement kappa scores[35] and *P* values were computed by DAG_Stat[36]. Comparison of cases and controls was performed using multivariable logistic regression, with conditional logistic regression employed for paired comparisons, using Stata v.15.1 (Statacorp). Other statistical calculations were performed in GraphPad Prism 7 (GraphPad Software). PCAs were performed with the prcomp function from the R package in RStudio (v.0.99.902) with all settings, where applicable, set to 'true'. As the effect size was not known in advance, we performed power calculations with varying prevalence and effect sizes (odds ratio) for 100 case–control pairs (pre-eclampsia and growth restriction) used in the 16S rRNA amplicon sequencing study. These showed that a 5% prevalence in controls and OR = 5 gives 82% power to detect the signal at significance level 0.05. The bioinformatic analysis and the setting of the minimum detection thresholds were performed in a blinded fashion in respect to adverse pregnancy outcome status. All reported *P* values are two-sided except for concordance calculations, as indicated. The experiments were not randomized, and investigators were not blinded to allocation during experiments and outcome assessment unless described otherwise.

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

Data availability

The 16S rRNA gene sequencing datasets generated and analysed in this study are publicly available under European Nucleotide Archive (ENA) accession number ERP109246. The metagenomics datasets, which primarily contain human sequences, are available with managed access in the European Genome-phenome Archive (EGA) accession number EGAD00001004198.

20. Pasupathy, D. et al. Study protocol. A prospective cohort study of unselected primiparous women: the pregnancy outcome prediction study. *BMC Pregnancy Childbirth* **8**, 51 (2008).
21. Gardosi, J., Mongelli, M., Wilcox, M. & Chang, A. An adjustable fetal weight standard. *Ultrasound Obstet. Gynecol.* **6**, 168–174 (1995).
22. American College of Obstetricians and Gynecologists & Task Force on Hypertension in Pregnancy. Report of the American College of Obstetricians and Gynecologists' Task Force on Hypertension in Pregnancy. *Obstet. Gynecol.* **122**, 1122–1131 (2013).
23. Lager, S. et al. Detecting eukaryotic microbiota with single-cell sensitivity in human tissue. *Microbiome* **6**, 151 (2018).
24. Walker, A. W. et al. 16S rRNA gene-based profiling of the human infant gut microbiota is strongly influenced by sample processing and PCR primer choice. *Microbiome* **3**, 26 (2015).
25. Wood, D. E. & Salzberg, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* **15**, R46 (2014).
26. Nurk, S. et al. Assembling single-cell genomes and mini-metagenomes from chimeric MDA products. *J. Comput. Biol.* **20**, 714–737 (2013).
27. Johnson, M. et al. NCBI BLAST: a better web interface. *Nucleic Acids Res.* **36**, W5–W9 (2008).
28. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
29. Carver, T., Harris, S. R., Berriman, M., Parkhill, J. & McQuillan, J. A. Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. *Bioinformatics* **28**, 464–469 (2012).
30. Kozich, J. J., Westcott, S. L., Baxter, N. T., Highlander, S. K. & Schloss, P. D. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. *Appl. Environ. Microbiol.* **79**, 5112–5120 (2013).
31. Eren, A. M. et al. Oligotyping: differentiating between closely related microbial taxa using 16S rRNA gene data. *Methods Ecol. Evol.* **4**, 1111–1119 (2013).
32. Schmieder, R. & Edwards, R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* **27**, 863–864 (2011).
33. Quast, C. et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590–D596 (2013).
34. Ludwig, W. et al. ARB: a software environment for sequence data. *Nucleic Acids Res.* **32**, 1363–1371 (2004).
35. Viera, A. J. & Garrett, J. M. Understanding interobserver agreement: the kappa statistic. *Fam. Med.* **37**, 360–363 (2005).
36. Mackinnon, A. A spreadsheet for the calculation of comprehensive statistics for the assessment of diagnostic tests and inter-rater agreement. *Comput. Biol. Med.* **30**, 127–134 (2000).

Acknowledgements The work was supported by the Medical Research Council (UK; MR/K021133/1) and the National Institute for Health Research (NIHR) Cambridge Biomedical Research Centre (Women's Health theme). We thank L. Bibby, S. Ranawaka, K. Holmes, J. Gill, R. Millar and L. Sánchez Busó for technical assistance during the study. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care.

Author contributions G.C.S.S., D.S.C.-J., J.P. and S.J.P. conceived the experiments. G.C.S.S., D.S.C.-J., J.P., S.J.P. and S.L. designed the experiments. S.L. and M.C.d.G. optimized the experimental approach. S.L. and F.G. performed the experiments. M.C.d.G. analysed all of the sequencing data. U.S. matched cases and controls, performed statistical analyses and provided logistical support for patient and sample metadata. E.C. managed sample collection and processing and the biobank in which all sample were stored. All authors contributed in writing the manuscript and approved the final version.

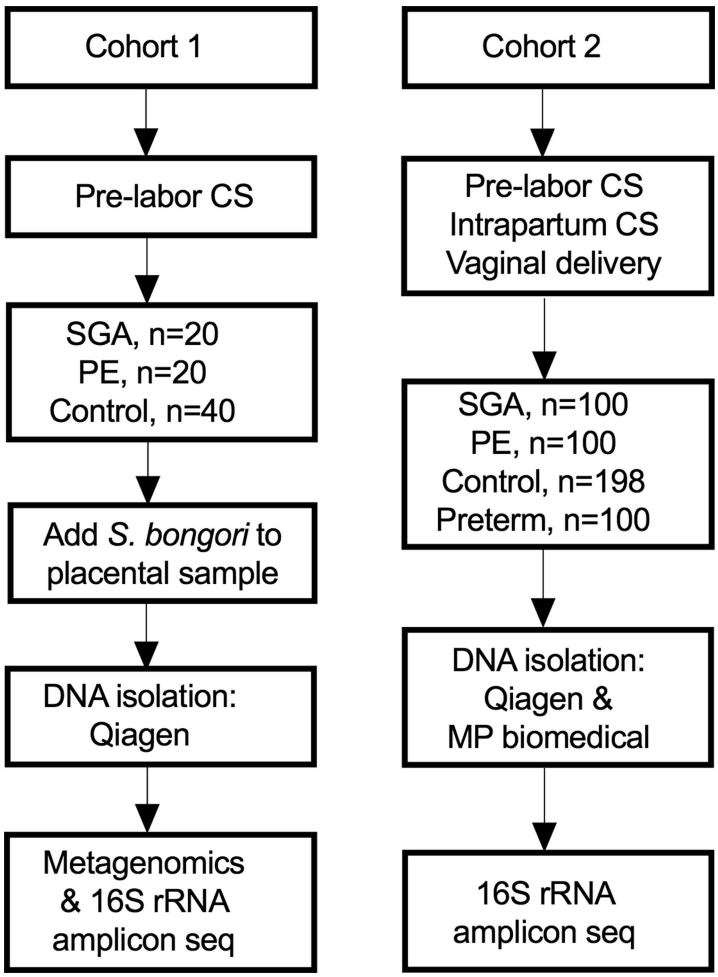
Competing interests J.P. reports grants from Pfizer, personal fees from Next Gen Diagnostics, outside the submitted work; S.J.P. reports personal fees from Specific, personal fees from Next Gen Diagnostics, outside the submitted work; D.S.C.-J. reports grants from GlaxoSmithKline Research and Development, outside the submitted work and non-financial support from Roche Diagnostics, outside the submitted work; G.C.S.S. reports grants and personal fees from GlaxoSmithKline Research and Development, personal fees and non-financial support from Roche Diagnostics, outside the submitted work; D.S.C.-J. and G.C.S.S. report grants from Sera Prognostics, non-financial support from Illumina, outside the submitted work. M.C.d.G., S.L., U.S., F.G. and E.C. have nothing to disclose.

Additional information

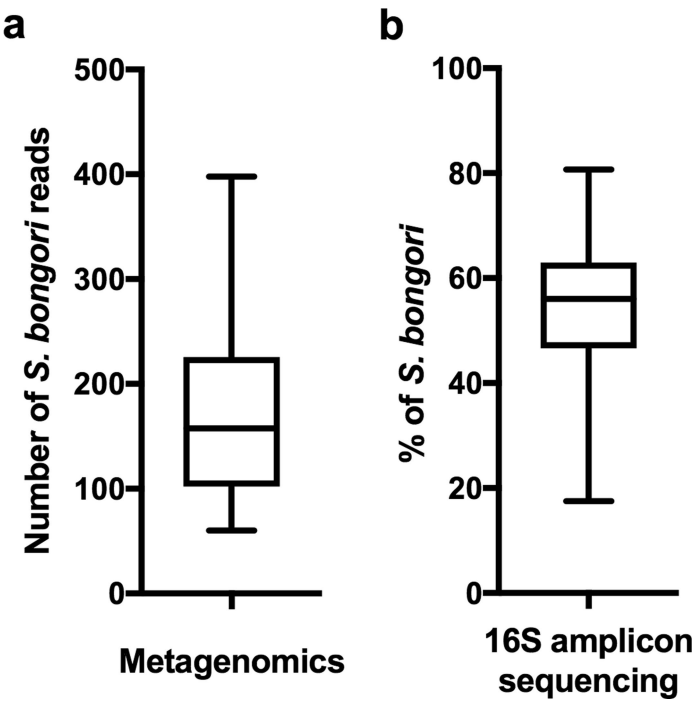
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-019-1451-5.

Correspondence and requests for materials should be addressed to J.P. or G.C.S.S.

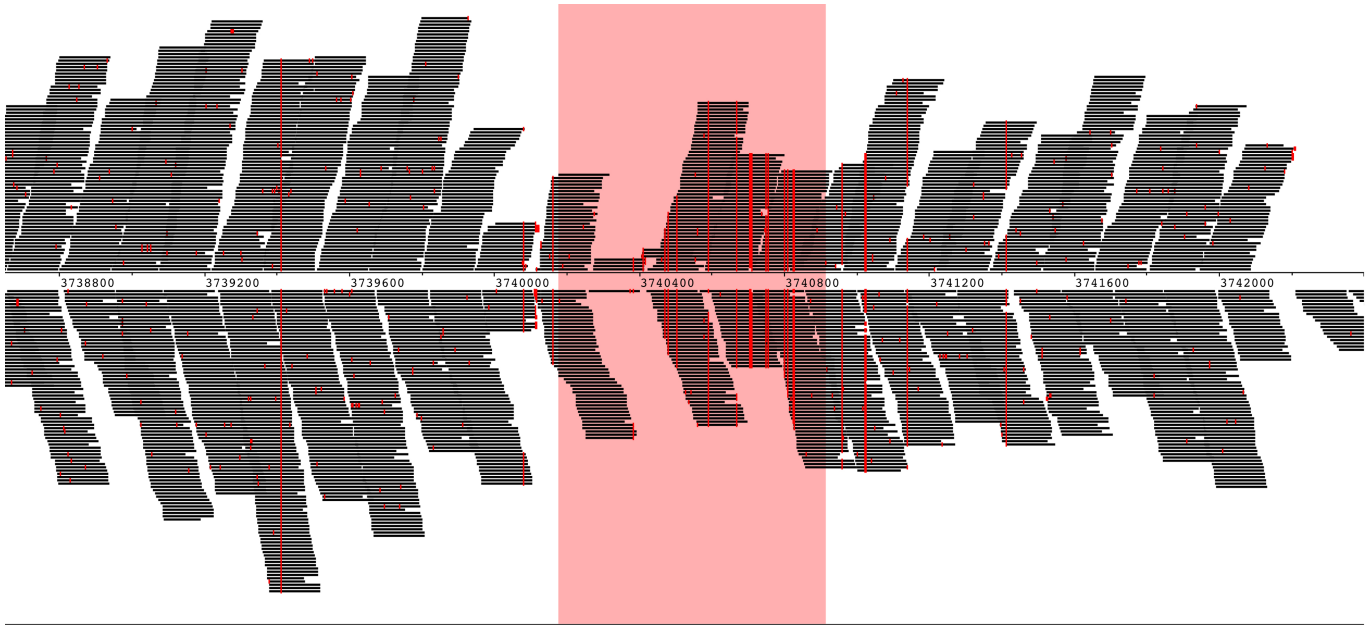
Peer review information *Nature* thanks David N. Fredricks and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.



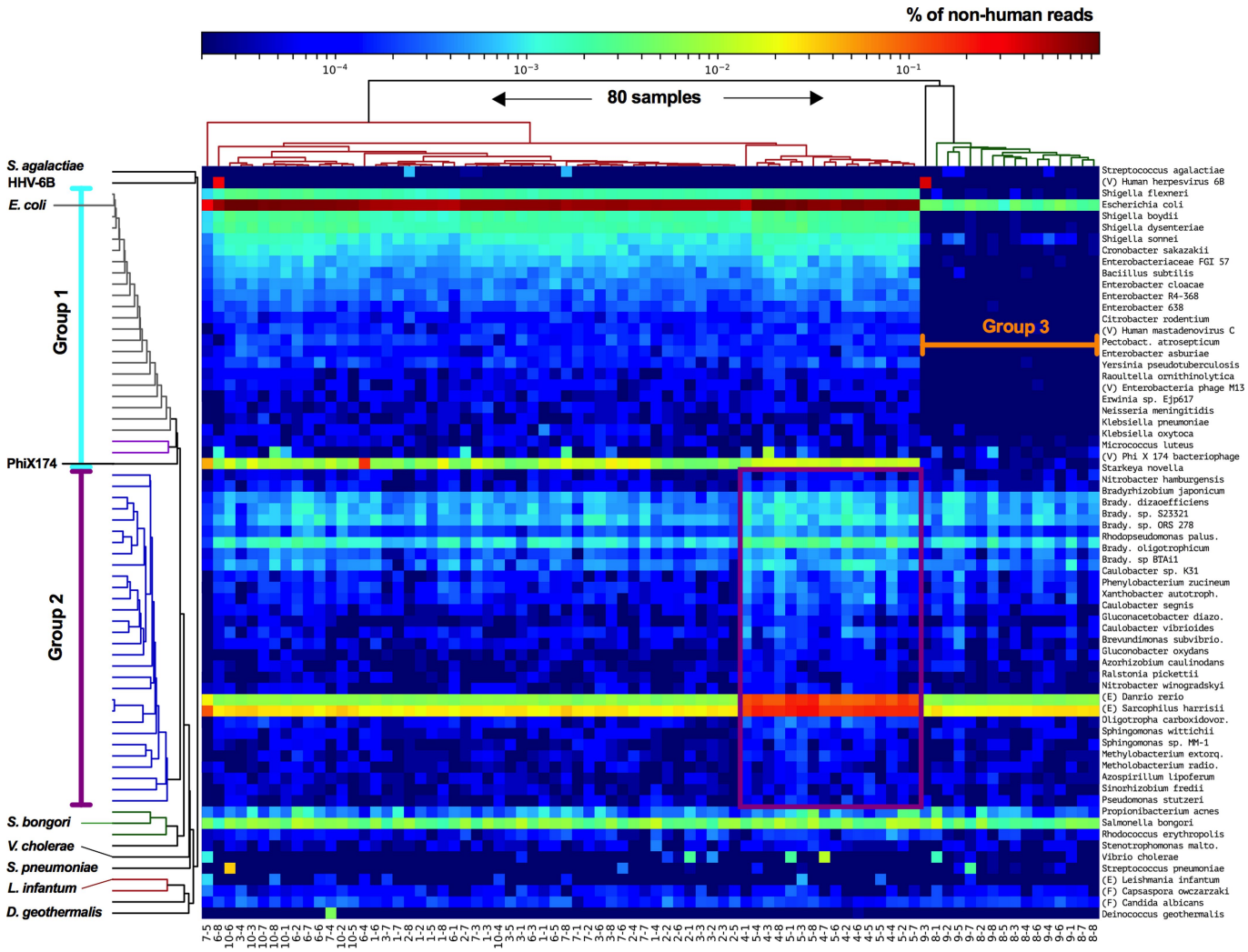
Extended Data Fig. 1 | Two cohorts of placental samples were analysed. Cohort 1 ($n = 80$) contained only samples from pre-labour Caesarean section (CS) deliveries and *S. bongori* was added to the samples before DNA isolation as a positive control. Samples in cohort 1 were analysed by both metagenomics and16S rRNA amplicon sequencing. Cohort 2 ($n = 498$) contained placental samples from Caesarean section and vaginal deliveries. DNA was isolated twice from each placental sample with two different DNA extraction kits. Samples were analysed by 16S rRNA amplicon sequencing. Pre-eclampsia (PE) was defined using The American College of Obstetricians and Gynaecologists (ACOG) 2013 definition. Small for gestational age (SGA) was defined as a birth weight less than the fifth percentile using a customized reference. Preterm denotes birth before 37 weeks gestation.



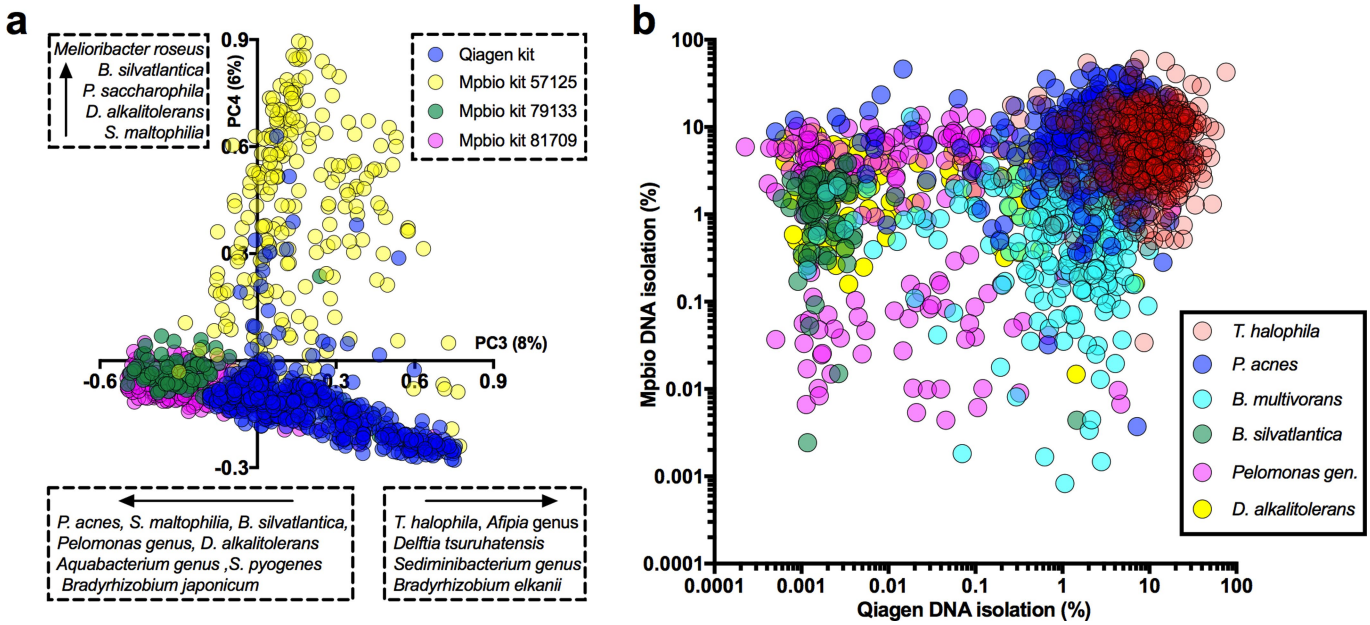
Extended Data Fig. 2 | Positive control experiment comparison between metagenomics and 16S amplicon sequencing. a, **b**, Adding approximately 1,100 CFUs of *S. bongori* to the placental tissue before DNA isolation resulted in an average of 180 reads (s.d. 90 reads) by metagenomic sequencing ($n = 80$) (**a**) or on average of 54% of all 16S rRNA amplicon sequencing reads (approximately 33,000 reads) being identified as *S. bongori* (s.d. 13%; $n = 79$) (**b**). Box represents the interquartile range; whiskers represent the maximum and minimum values; centre lines denote the median.



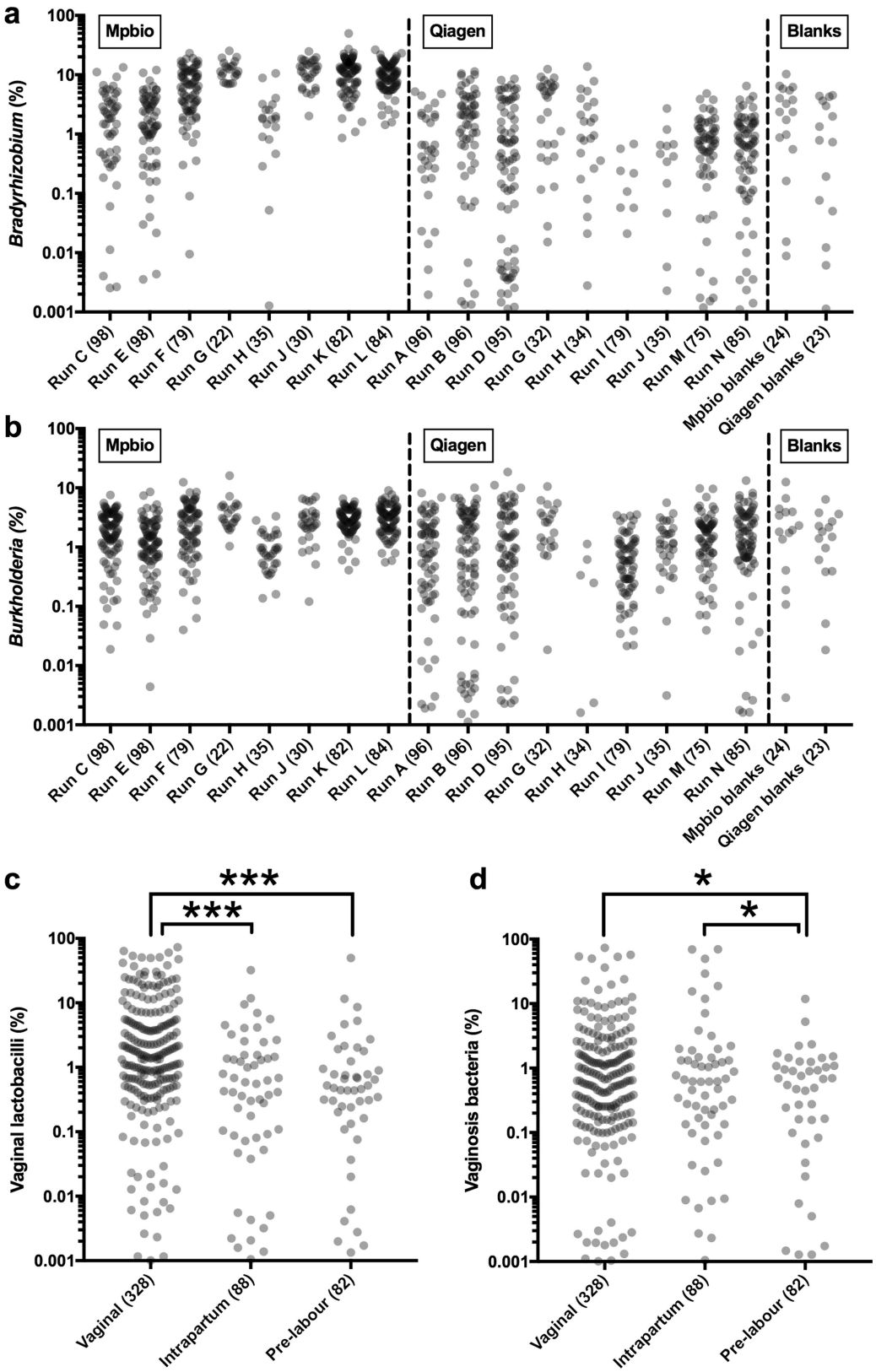
Extended Data Fig. 3 | Strain analysis of *E. coli* reads found by metagenomics. All reads identified in all 80 samples by Kraken[25] as *E. coli* were extracted and mapped together against the closest *E. coli* reference genome (GenBank: CP02409.1). Single nucleotide polymorphisms, shown in red, were consistent for all samples across the genome. Single nucleotide polymorphisms were rare, except in the fimbrial chaperone protein gene (*EcpD*) indicated in light red. Sequence differences that appear as short sporadic red lines represent sequencing errors. Strain variation would have resulted in dashed vertical lines.



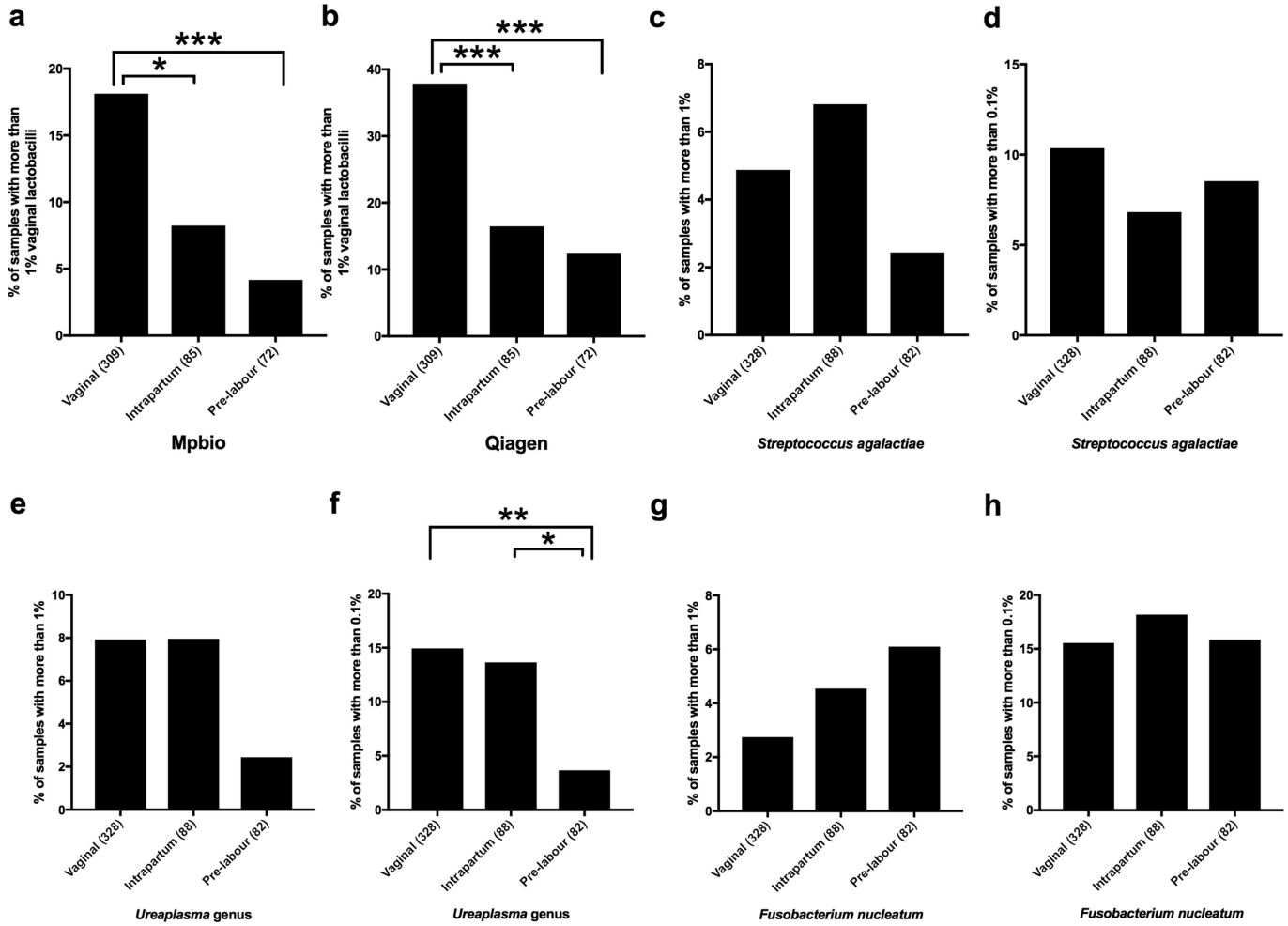
Extended Data Fig. 4 | Detailed heat map metagenomic data. Heat map showing the abundance of all non-human reads as detected by metagenomics. Human reads remaining after filtering (89.8%; s.d. 1.5%) are not shown for scaling purposes. Most taxa (shown on the right) are found in higher abundance within groups 1 and/or 2 (indicated on the left with light blue and purple, respectively). The purple box highlights the samples and species associated with group 2. The lane ID of each sample is represented by the first number (*x* axis). All samples from lanes 4 and 5 form group 2, and all samples from lanes 8 and 9 form group 3 (see Fig. 1a, b).



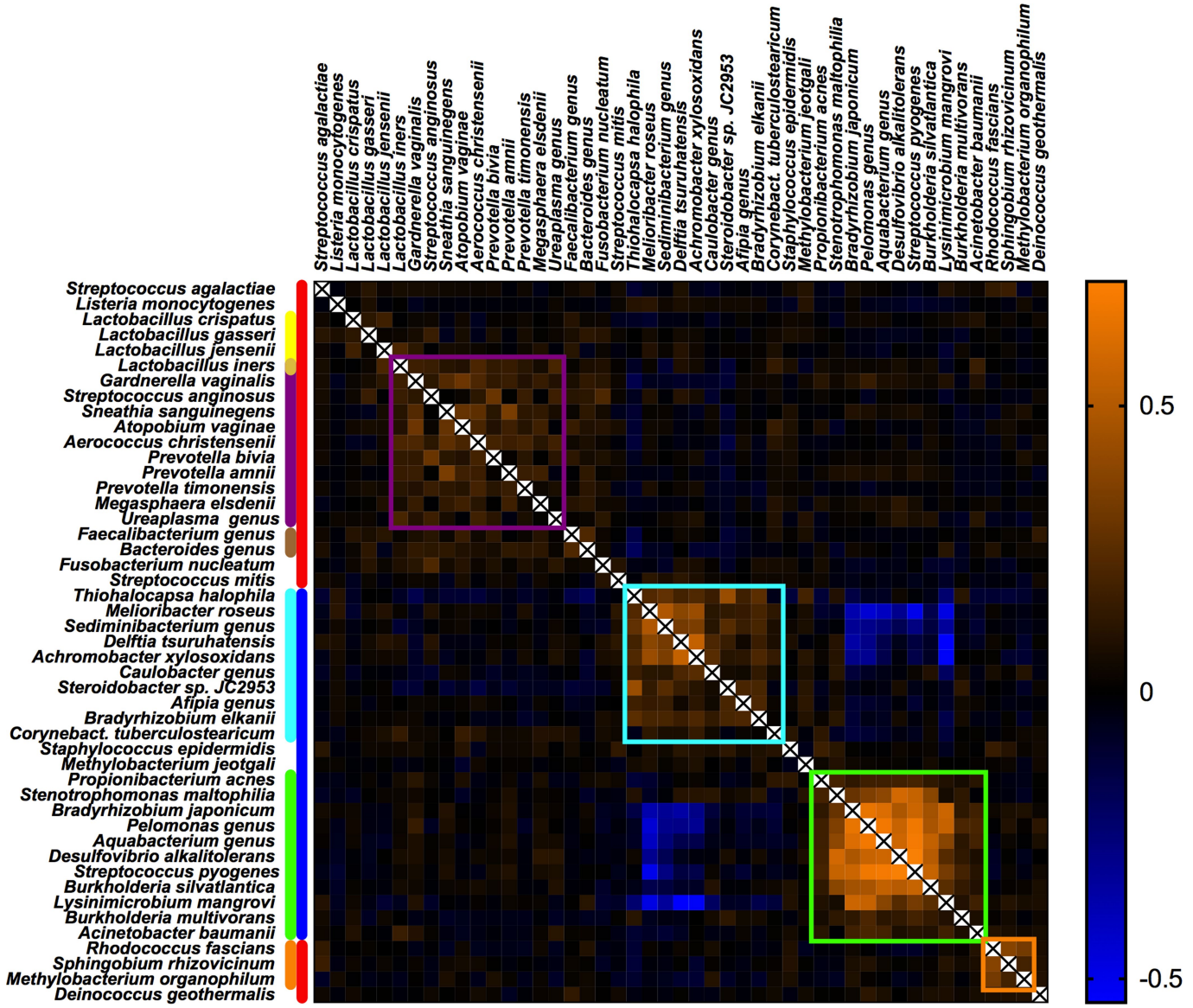
Extended Data Fig. 5 | Species associated with batch effects visualized by PCA also do not show signal reproducibility. a, PCAs of selections of samples from cohort 2 (16S), or of all cohort 2 samples as shown here, allows for the identification of batch effects and allows for the identification of contaminating species associated with the use of specific DNA isolation methods, kits and/or other reagents. An analysis of all samples shows that principal components 3 (*x* axis) and 4 (*y* axis) are strongly correlated with the use of Qiagen or specific Mpbio DNA isolation kits. **b**, Examples of bacteria detected in high abundance and frequency when processed with the Qiagen (*x* axis) and/or Mpbio (*y* axis) DNA isolation kits. Patterns that lack positive correlation (compare with Fig. 2a) demonstrate that signals are not sample- but batch-associated.



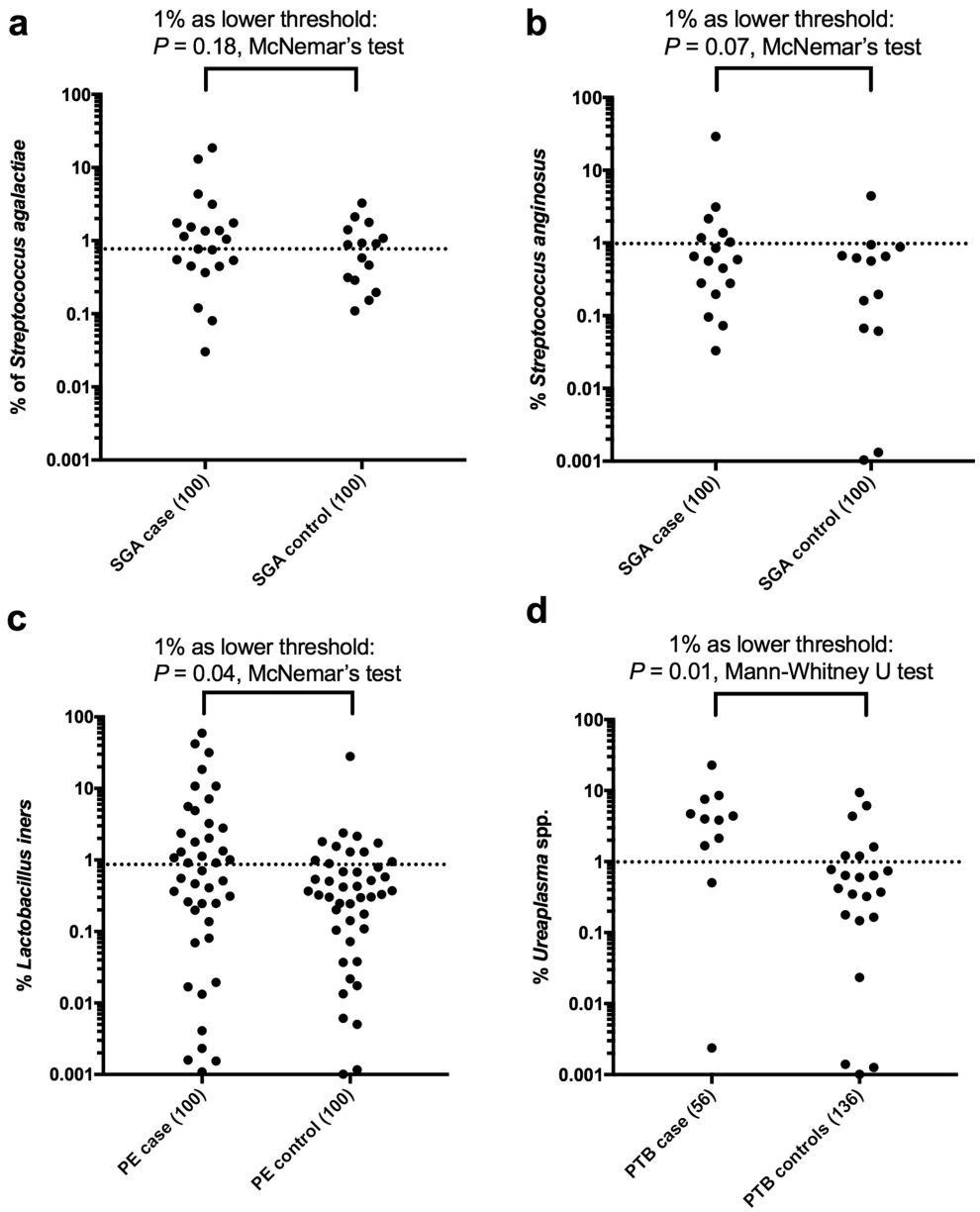
Extended Data Fig. 6 | Scatterplot representations of the abundance of *Bradyrhizobium*, *Burkholderia*, vaginal lactobacilli and vaginosis bacteria during 16S amplicon sequencing. a, **b**, The abundance of *Bradyrhizobium* (**a**) or *Burkholderia* (**b**) with respect to sequencing run batch effects during 16S amplicon sequencing. Numbers in parentheses indicate the number of samples sequenced in a given run. Values of zero are not shown on the logarithmic axis. **c**, **d**, The abundance of vaginal lactobacilli (**c**) and vaginosis bacteria (**d**) with respect to the mode of delivery during 16S amplicon sequencing. $^{*}P < 0.05$, $^{***}P < 0.001$, Mann–Whitney *U*-tests, where values below 1% are regarded as 0% (not biologically relevant).



Extended Data Fig. 7 | Mode of delivery and the detection of bacterial signals. a, **b**, The association of vaginal lactobacilli with the mode of delivery, as determined by the analysis of 466 samples by 16S amplicon sequencing that were successfully sequenced twice using the Mpbio (**a**) and Qiagen (**b**) DNA isolation methods. Comparisons of the Mpbio and Qiagen DNA isolation techniques highlight that the same patterns are observed in the associations with mode of delivery. Comparisons also show that the Qiagen DNA isolation was more sensitive, resulting in twice as many signals above the 1% threshold. **c**–**h**, The association of bacterial groups with mode of delivery. Analyses were performed using all 498 placental samples with the highest value of either DNA isolation method for each bacterial group per sample. **c**, **d**, *S. agalactiae* was not associated with the mode of delivery irrespective of whether a 1% threshold was used (the minimum percentage considered to be potentially ecologically relevant) (**c**) or a 0.1% threshold was used (the 16S detection limit, relevant for detecting traces of contamination during delivery) (**d**). **e**, **f**, The *Ureaplasma* genus was significantly associated with the mode of delivery using the 0.1% threshold, similar to Fig. 2c, which describes the combination of all vaginosis-associated bacteria. **g**, **h**, *F. nucleatum* was not associated with the mode of delivery, irrespective of whether a 1% (**g**) or 0.1% (**h**) threshold was used. $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$, Mann–Whitney *U*-tests.



Extended Data Fig. 8 | Heat map of Spearman's rho correlation coefficients of bacterial signals as found by 16S rRNA amplicon sequencing. Sample-associated signals (red bar), are typically identified by increased kappa scores, as shown in Supplementary Table 4. Reagent contaminants are indicated by a blue bar. Vaginosis-associated bacteria (purple bar) show positive correlations (purple square) with each other, *Lactobacillus iners* and faecal bacteria (brown bar). Lactobacilli (yellow bar) show limited positive correlation with faecal bacteria. Reagent contaminants mainly associated with the Qiagen (light blue) or the Mpbio (green) kit form distinct clusters. Species that are strongly associated with sample collection contamination in 2012–2013 are indicated in orange. For each species the highest value (percentage) found using either the Qiagen or the Mpbio DNA isolation kit, was used as input (using all 498 samples).



Extended Data Fig. 9 | Bacterial signals and adverse pregnancy outcome. a–d, Scatterplot representations of the non-significant associations of *S. agalactiae* with SGA (**a**), *S. anginosus* with SGA (**b**), and of the significant associations of *L. iners* with pre-eclampsia (**c**), and *Ureaplasma* with PTB (**d**). Samples with 0% signal are not shown on the logarithmic scale. Signals above 1% (dotted line) are regarded as positive for use in McNemar's test (**a–c**), and signals below 1% are considered as negative. The Mann–Whitney *U*-test was used for unpaired samples in **d**.

nature research

Corresponding author(s): Gordon C. S. Smith

Last updated by author(s): May 30, 2019

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	The only software used to collect data was the standard MiSeq and HiSeq (Illumina) sequencing machine software and the quantitative PCR machine software (QuantStudio 6 Flex system, ThermoFisher Scientific).
Data analysis	KneadData (v0.6.1), Kraken (v0.10.6), Mothur (v1.40.5), PRINSEQ-lite (v0.20.3), oligotyping (v2.1) , ARB (v5.5-org-9167), DAG_Stat, Stata (v15.1), R package RStudio (v0.99.902), Past3 (v3.14), Prism 7 (v7.0c), Spades (v3.11.0), BWA (v0.7.17-r1188), Artemis (v.16.0.0), BLASTN (https://blast.ncbi.nlm.nih.gov/Blast.cgi) and custom script was used to extract reads identified of a particular group of interest identified by Kraken (Supplemental Information).

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

The 16S rRNA gene sequencing datasets utilized in this study are publicly available under European Nucleotide Archive (ENA) accession no. ERP109246. The metagenomics data sets, which primarily contain human sequences, are available in the European Genome-phenome Archive (EGA) with managed access (EGAD00001004197).

Field-specific reporting

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

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

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

Life sciences study design

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

Sample size	A power calculation was performed during the planning phase of the Pregnancy Outcome Prediction (POP) study and it is described in Pasupathy et al (BMC Pregnancy and Childbirth 2008 PMID 19019223). In brief, the sensitivity of different models for a given screen positive rate was quantified by 95% confidence intervals. The calculations indicated that the study was likely to provide reasonably precise estimates of sensitivity for conditions with a 3% incidence, such as severe SGA. The use of a nested case-control design with a 1:1 matching of cases and controls on key maternal characteristics was also planned in advance in the context of very expensive or labor intensive methodologies (Pasupathy et al). For the 16S rRNA amplicon sequencing study we used 100 matched cases and controls for both pre-eclampsia and growth restriction (ie 200 samples in total). As the effect size was not known in advance we performed power calculations with varying prevalence and effect sizes (OR) for 100 case-control pairs. These showed that a 5% prevalence in controls and OR=5 gives 82% power to detect the signal at significance level 0.05.
Data exclusions	A total of 4512 women with a viable singleton pregnancy were recruited to the POP study. The only clinical exclusion criterion was multiple pregnancy.
Replication	Reproducibility of signals was confirmed by analyzing samples both by metagenomic and 16S rRNA amplicon analysis (cohort 1) and by analysing each sample from cohort 2 twice by 16S rRNA amplicon sequencing using 2 different DNA isolation methods. A large part of the manuscript is about proving the reproducibility of signals in order to show which signals are real and which ones are spurious
Randomization	The POP study is a prospective cohort study of nulliparous women attending the Rosie Hospital (Cambridge, UK) for their dating ultrasound scan. All eligible participants were included. For the purpose of the experimental projects described in this manuscript, participants were allocated into groups based on pregnancy outcome (details in Methods and Supplementary information). Outcome data were ascertained by review of each woman's paper case record by research midwives and by record linkage to clinical electronic databases. Paired cases and controls were always processed together and sequenced in the same run.
Blinding	All the aspects of the POP study were conducted blind: the results of the research ultrasound scans and the biochemical marker data were not revealed to the clinicians, patients and researchers performing the downstream experiments. Data were unblinded only at the statistical analysis stage. Specifically, all of the bioinformatic analysis of 16S rRNA amplicon data and the metagenomic data was performed in a blinded fashion. Reagent contamination recognition was also performed prior to unblinding. Finally, a statistical analysis plan was written prior to unblinding for the analysis of Streptococcus agalactiae, the only bacterial signal that passed all quality checks for being a genuine and possibly important. All other bacterial analyses (done for all the other bacteria) should be considered exploratory.

Reporting for specific materials, systems and methods

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

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	Samples were from the Pregnancy Outcome Prediction (POP) study. In the whole POP study population (n=4212), the median age, height and BMI (IQR) were 30.3 (26.8 to 33.4) years, 165 (161 to 169) cm, 24.1 (21.8 to 27.3) kg/m2, respectively, and 13% of the women were smokers at recruitment. Detailed characteristics of women whose samples were selected for sequencing in this study are given in Extended Data Tables 1 and 2. In brief, the median maternal age varied between 29.7 and 30.9 years between the groups of 100 cases or controls (Extended Data Table 2). The median height was similar (164-165 cm) between the groups. The median BMI was highest in the PE cases (25.7 kg/m2) and otherwise varied between 24.1 and 25.0 kg/m2 between the groups. The prevalence of smoking at booking varied the most; it was 28% in the SGA group and 7% among the controls of PE cases.
Recruitment	Samples were from the Pregnancy Outcome Prediction (POP) study. Nulliparous women with a viable singleton pregnancy who attended their dating ultrasound scan at the Rosie Hospital (Cambridge, UK) between 14 January 2008 and 31 July 2012 were eligible (n=8028), and 4512 (56%) of them provided an informed consent and were recruited. The recruited and non-recruited women were broadly comparable, although according to the hospital record data the women who were recruited were slightly older, more often of white ethnic origin and less likely to smoke. In addition, women were excluded because they delivered elsewhere (n=233) or withdrew their consent (n=67). The cohort of 4212 women used for the sample selection in the present study can be regarded as fairly well representative of the eligible population. See Sovio et al Lancet 2015 PMID 26360240 and Gaccioli et al Placenta 2017 PMCID PMC5701771 for a complete description.
Ethics oversight	The Pregnancy Outcome Prediction study was approved by the Cambridgeshire 2 Research Ethics Committee (reference number 07/H0308/163).

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