Strong pathogen competition in neonatal gut colonisation

Tommi Mäklin[1,*], Harry A. Thorpe[2], Anna K. Pöntinen[2,3], Rebecca A. Gladstone[2], Yan Shao[4], Maiju Pesonen[2], Alan McNally[5], Pål J. Johnsen[6], Ørjan Samuelsen[3,6], Trevor D. Lawley[4], Antti Honkela[1], Jukka Corander[2,4,7,*]

1 Helsinki Institute for Information Technology HIIT, Department of Computer Science, University of Helsinki, Helsinki, Finland

2 Department of Biostatistics, University of Oslo, Oslo, Norway

3 Norwegian National Advisory Unit on Detection of Antimicrobial Resistance, Department of Microbiology and Infection Control, University Hospital of North Norway, Tromsø, Norway

4 Parasites and Microbes, Wellcome Sanger Institute, Hinxton, Cambridgeshire, UK

5 Institute of Microbiology and Infection, University of Birmingham, Birmingham, UK

6 Department of Pharmacy, Faculty of Health Sciences, UiT The Arctic University of Norway, Tromsø, Norway

7 Helsinki Institute for Information Technology HIIT, Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland

[*] Corresponding author, tommi.maklin@helsinki.fi, jukka.corander@medisin.uio.no

Abstract

Bacterial pathogen species and their strains that colonise the human gut are generally understood to compete against both each other and the commensal species colonising this ecosystem. However, currently we are lacking a population-wide quantification of strain-level colonisation dynamics for many common bacterial pathogens and the relationship of colonisation potential to prevalence in disease is unknown. In addition, it is unclear how ecological factors might be modulating the dynamics. Here, using a combination of latest high-resolution metagenomics and strain-level genomic epidemiology methods leveraging large genomic reference libraries of key pathogens, we performed a characterisation of the competition (co-exclusion and co-presence) and colonisation dynamics for a longitudinal cohort of neonatal gut microbiomes. We found a strong inter- and intra-species competition dynamic in the gut colonisation process, but also a number of synergistic relationships among several species belonging to genus *Klebsiella*, which includes the prominent human pathogen *Klebsiella pneumoniae*. Additionally, we find no evidence of preferential colonisation by hospital-adapted pathogen lineages in either vaginal or caesarean section birth groups. Our analysis also enables the first unbiased assessment of the strain-level colonisation potential of extra-intestinal pathogenic *Escherichia*

coli (ExPEC) in comparison with their potential to cause bloodstream infections. We determined that the established common ExPEC clones ST73 and ST95 are overall significantly more pathogenic than the more recent, globally circulating multi-drug resistant clone ST131, where only a single subclone (ST131-C2) exhibited excess pathogenic potential. Our study highlights the importance of systematic surveillance of bacterial gut pathogens, not only from disease but also from carriage state, to better inform therapies and preventive medicine in the future.

Introduction

Human gut bacteria are generally considered commensal organisms but some of them harbour considerable potential to cause either mild or severe infections outside the gut. One of the most prominent examples is extra-intestinal pathogenic *Escherichia coli*, which is the predominant facultative anaerobe in the large intestine [1]. Work on *E. coli* going back several decades suggests strong intra-species competition in healthy colonisation based on serotypic variation, or the lack of thereof [2]. Multi-locus enzyme electrophoresis typing studies done on longitudinal collections of stool confirmed these conclusions in the early 1980s [3], [4].

Recent "bottom-up" experimental studies further support that inter-species competition plays a key role in shaping bacterial gut communities [5], [6]. However, despite considerable research effort over the years on this topic, systematic population-wide characterisation of the colonisation potential and competition dynamics simultaneously across intra- and inter-species levels is still lacking. Our current study aims to address the need to assess these aspects for several of the major human gut species with pathogenic potential.

In animal models the role of intra-species competition in colonisation has been widely studied, producing certain interesting results related to the species that also colonise the human gut microbiome. For example, colonisation by *E. coli* has been shown to reduce abundance of *Salmonella typhimurium* in the mouse gut through competition for iron [7], and *Klebsiella michiganensis* was recently demonstrated to prevent mouse gut colonisation by *E. coli* in a particular ecological setting [8]. Other examples of intra-species competition include varying colonisation abilities between *E. coli* strains in

gnotobiotic mice [9], and subpopulations of *Salmonella typhimurium* within intestinal tissues mediating colonisation resistance against systemic strains through nutrient competition [10].

In the human microbiome, a recent longitudinal study [11] tracked the competition among *E. coli* strains in particular and the microbiome composition in general over multiple years in a single patient suffering from Crohn's disease, consequently having a dysbiotic microbiome and highly dominant abundance of *E. coli* [11]. Several commonly found *E. coli* clones were seen taking the role of the dominant strain over time, but return to a previously identified strain was never observed [11]. In another study [12], the diversity of *E. coli* colonising the gut prior, during, and after an ecological disruption was highlighted in an entero-toxigenic *E. coli* challenge, but for a small number of test subjects, which makes it hard to draw more general conclusions about the colonisation potential [12]. Characterising the diversity of colonising pathogenic bacteria in the gut is relevant per se, but also interesting in relation to polymicrobial infections which have not been widely studied to date. In particular in the urinary tract infection context, it has been shown that multiple pathogen species can form a complex network with both negative and positive interactions [13].

In this study we address the colonisation potential and competition dynamics questions simultaneously across intra- and inter-species levels. This was done by performing strain identification and genome assembly for a set of species with pathogenic potential using metagenomes collected from a longitudinal cohort of neonatal gut microbiome samples [14]. We chose the particular neonatal cohort for multiple reasons. Firstly, because of its longitudinal aspect which enabled us to investigate dynamics of the competition. Secondly, due to its high sequencing depth and state-of-the-art short-read technology used, which enabled us to assemble the genomes of the colonising strains in sufficient detail for genomic epidemiological analysis. Thirdly, the DNA obtained from a sufficiently large number of faecal samples via the culture-independent approach used in the study provided an opportunity for us to screen for the pathogens in an unbiased and robust manner. Finally, the neonatal cohort was ideal for the study since it enabled us to probe the competition dynamics between the pathogen strains and species when they colonise a niche that is approximately empty, unlike the

situation later in life. In addition, we screened a large body of microbiome to identify other published studies that would be suitable for our purposes, but only the chosen study [14] was deemed appropriate, satisfying all these desiderata.

In the selected study cohort, stool samples were collected from the neonates at 4, 7, and 21 days after birth, later during the infancy period for a subcohort, and also from the mothers of a subcohort during their stay in the maternity unit. The neonates participating in the study represented two delivery cohorts: one for the vaginally born, and the other for caesarean section birth, the latter of which was shown to be associated with a massive shift in the general gut microbiome composition in the original publication [14]. Access to this longitudinal data from metagenomic sequencing at an unprecedented depth (median depth ~20 million paired-end 150bp reads), combined with recent analytical advances in genomic epidemiology of mixed samples [15], [16] and vastly improved availability of large and high-quality genome libraries from studies of key human pathogens based on whole-genome sequencing of single bacterial isolates, enabled us to interrogate the colonisation and competition processes at the level of assembled genomes of pathogen strains.

Results

Lineage-level analysis of neonatal gut microbiome data

We analysed 1679 sets of paired-end short reads from gut metagenomes based on stool samples collected from the two delivery cohorts vaginally and caesarean section born babies at several time points (Table 1). Our study extends the previous analysis [14] by providing lineage-level characterization and subsequent genome assembly from these samples for several important pathogen species (Table 2) as well as a more detailed exploration of the diversity within the *Klebsiella* genus. In the remaining text we refer to a “lineage” and a “strain” interchangeably, as is commonplace in the genomic epidemiology literature.

Timepoint (days after birth)	Number of babies sampled	
	C-section	Vaginal delivery
4	153	157
6	0	2
7	252	280
8	1	1
9	1	1
10	3	0
11	0	4
12	4	4
13	1	1
14	2	3
17	0	1
18	1	0
21	178	147

Table 1 Number of babies sampled at each timepoint during the first month of life.

In both the lineage-level characterization and the *Klebsiella* species analysis we applied the recent mSWEEP method for estimating the relative abundances of predefined lineages, or species, from metagenomic short-read sequencing data [15] and the mGEMS method for assigning the metagenomic reads to bins that correspond to a single lineage [16]. The use of both methods was enabled by a bespoke set of reference sequences (described in the Methods section) covering the within-species variation in the chosen target species (Table 2). Our intended focus was on pathogenic Enterobacteriaceae and Enterococci, which have been widely considered in colonisation experiments in animal models and have large representative

collections of high-quality reference genomes available. However, we did also screen for the presence of multiple other pathogen species but found only sporadic cases of colonisation (Table 2), which prevents drawing any robust conclusions regarding them and they are consequently excluded from further consideration. The analysis pipeline is described in more detail in the Methods section and in Supplementary Figure 1. Comprehensive sets of results from the analyses will be presented in the following sections.

Full name	**Abbreviation**	**Reliably identified colonisations**
Acinetobacter baumannii	A. bau	0
Escherichia coli	E. col	1124
Enterococcus faecalis	E. fcs	690
Enterococcus faecium	E. fcm	38
Klebsiella aerogenes	K. aer	12
Klebsiella grimontii	K. gri	100
Klebsiella huaxiensis	K. hua	1
Klebsiella michiganensis	K. mic	86
Klebsiella ornithinolytica	K. orn	9
Klebsiella oxytoca	K. oxy	27
Klebsiella pasteurii	K. pas	0
Klebsiella planticola	K. pla	4
Klebsiella pneumoniae	K. pne	186
Klebsiella quasipneumoniae subsp. quasipneumoniae	K. qpq	0
Klebsiella quasipneumoniae subsp. similipneumoniae	K. qps	1
Klebsiella spallanzanii	K. spa	0

Klebsiella variicola	K. var	23
Pseudomonas aeruginosa	P. aer	6
Staphylococcus aureus	S. aur	76
Streptococcus pneumoniae	S. pne	0

Table 2 List of the target pathogen species, abbreviations of their names used in the manuscript, and the numbers of times lineages of each species were identified reliably in the metagenomics data using the mGEMS pipeline.

The study cohort in the previously published study [14] consisted of 596 healthy babies from full-term pregnancies, delivered either via vaginal birth (314 babies) or caesarean section (282 babies). All 596 babies provided faecal samples at least once during their first month of life (see Table 1 for details), and 302 babies were resampled once between ages 7-11 months (142 caesarean section and 160 vaginal deliveries). Faecal samples were also available from 175 mothers (65 from the caesarean section cohort and 110 the vaginal delivery cohort), collected before or after delivery in the maternity unit and paired with 178 babies. Since the 175 samples from mothers were not available in a systematic fashion in the original study, we excluded a comparative analysis of the mother vs infant pairs of strains. Further, the isolate genome data (805 samples) from the original study were not included in the analysis of the within-sample variation to avoid biasing our conclusions. This is due to the fact that isolates were cultured in a highly uneven manner, such that a majority of them represented *Enterococcus* (451 out of 805), followed by *Klebsiella* spp. (235/805), while a very small fraction was available for the *Escherichia* spp (41/805). As *E. coli* was found to be the most dominant species in many samples by the metagenomic analysis, the isolate data would not have covered samples systematically. Although more isolates were available for *Enterococci*, it was necessary to treat all species in the same manner in our analysis.

Competition and synergistic relationships drive Enterobacteriaceae and *Enterococcus* colonisation

We first investigated the inter-species competition dynamics between various Enterobacteriaceae and two *Enterococcus* species, *Enterococcus faecalis* and

Enterococcus faecium. Using the species-level relative abundance estimates to estimate correlations between the target species, we identified statistically significant ($p < 0.05$, permutation test) antagonistic relationships between *E. coli* and *Klebsiella grimontii, Klebsiella michiganensis,* and *Klebsiella pneumoniae*, and similarly between *E. coli* and *E. faecalis* (Figure 1). The existence of this relationship is also suggested by the markedly more frequent absence of colonisation by *Klebsiella* species in the vaginally delivered cohort (Figure 2), where *E. coli* is in contrast more common, and has been previously verified for the *E. coli - K. michiganensis* pair in a mouse gut model [8]. A significant negative correlation was also found between *E. coli* and *Staphylococcus aureus;* however, since the latter species was only present in a limited number of samples and is known to be a common skin coloniser in the groin which could have led to contamination of the samples, no further analysis is conducted on the identified *S. aureus* lineages. The numbers of genomes reliably identified (meaning lineages that can be binned with mGEMS and pass an additional quality check as described in the Methods section) at the lineage-level are available in Table 2, providing an indication of the cardinality of our correlation analyses. Of note, large absolute values of a correlation coefficient in metagenomic sequencing of healthy microbiomes would not be expected because the typical (to this setting) low-abundance samples tend to dilute the correlation by signalling a random level association between particular taxa.

Within the *Klebsiella* genus we discovered that several species from the genus had a synergistic relationship with no statistically significant negative correlations observed in either cohort (Figure 1). Although some of the relationships were retained in both cohorts (*K. grimontii* with *K. michiganensis, Klebsiella oxytoca,* and *Klebsiella pasteurii*), notable differences between the cohorts were observed for the other *Klebsiella* species (Figure 1). Some of these differences are likely explained by the higher prevalence of *Klebsiella* in the caesarean section delivery cohort (Figure 2) but for species like *K. pneumoniae* that were commonly found in both cohorts, these observations may be indicative of more complex relationships arising from the different environments.

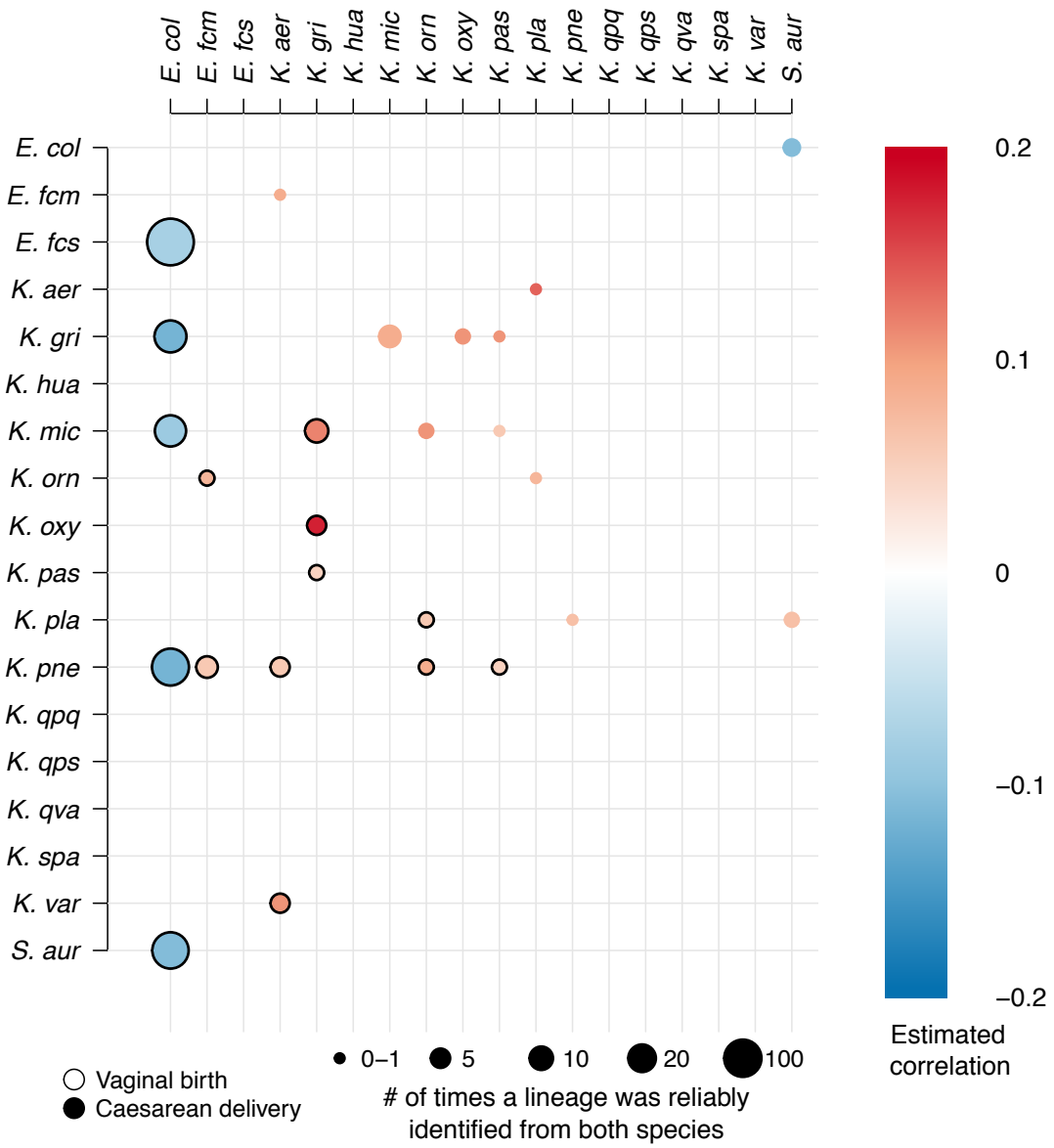

Figure 1 Correlations between the identified priority pathogens. The figure shows statistically significant ($p < 0.05$, permutation test) positive and negative correlations for the focus pathogen species. Correlations shown with a black border around the circle are for the vaginal delivery cohort (below diagonal), and correlations without a border for the caesarean section delivery cohort (above diagonal). Darker shades of red and blue represent stronger positive and negative correlation, while areas of the circles are proportional to the number of samples where the correlated pair was reliably identified. The correlation values were estimated from the relative abundance estimates.

Mode of birth affects the gut microbiome composition

Comparing the overall differences in species distribution between the caesarean section and the vaginal delivery cohorts using mSWEEP confirms the results presented in the original study [14]. Namely, *E. coli* is considerably more often found in the vaginal delivery cohort (Figure 2), while the *Klebsiella* species, *E. faecalis* and *E. faecium* are more common in the C section cohort (Figure 2). Table 2 shows the overall numbers of times a lineage from each of the species was identified in both cohorts, including screened species that were not identified in any of the samples.

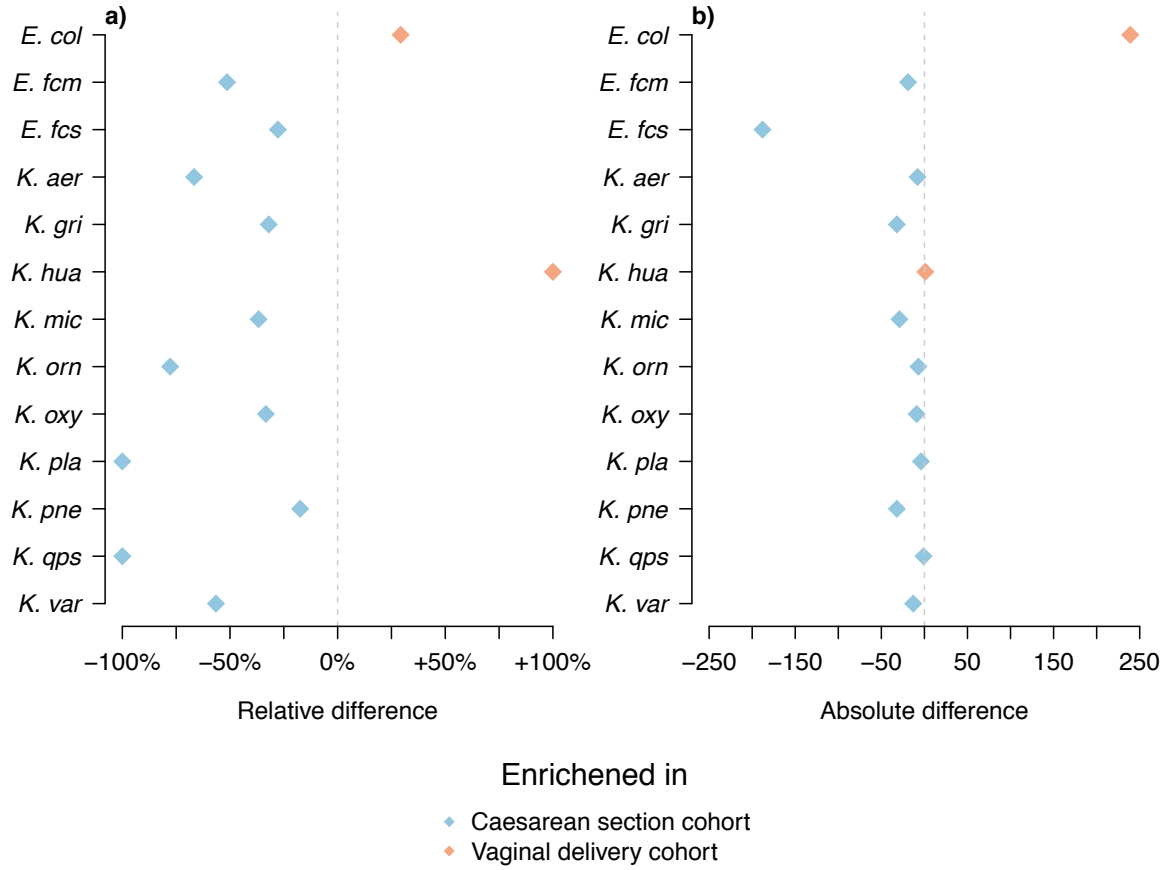

Figure 2 Differences in pathogen loads between cohorts. The figure shows differences in the number of reliably identified pathogens in each cohort. Panel **a)** shows the relative differences between the cohorts, and panel **b)** the absolute differences. Pathogens which are more common in the vaginally delivered cohort are coloured in orange and those more common in the caesarean section cohort are coloured in blue.

E. coli lineages rarely coexist with *Klebsiella* species or each other

Next, we looked in more detail into the *E. coli* lineage composition and coexistence by analysing co-occurrences of *E. coli* multilocus sequence types (STs) with each other and *Klebsiella* species. We found little overlap, with the majority of the cases containing just one *E. coli* ST or *Klebsiella* species (Figure 3). When coexistence was observed we did not find it happening in a systematic way, with most identified coexisting pairs or triplets observed just a few times depending on the overall prevalence of the particular types in the data set. Notable exceptions occurred in the case of the *K. michiganensis* - *K. grimontii* pair and the *K. grimontii* - *K. pneumoniae* pair, which were found together in the caesarean section delivered cohort a total of six times (out of total 596 samples collected from the c-section delivered babies during the first month of life) each and in the *K. michiganensis* - *K. grimontii* case were also established as synergistic in the correlation analysis (Figure 1).

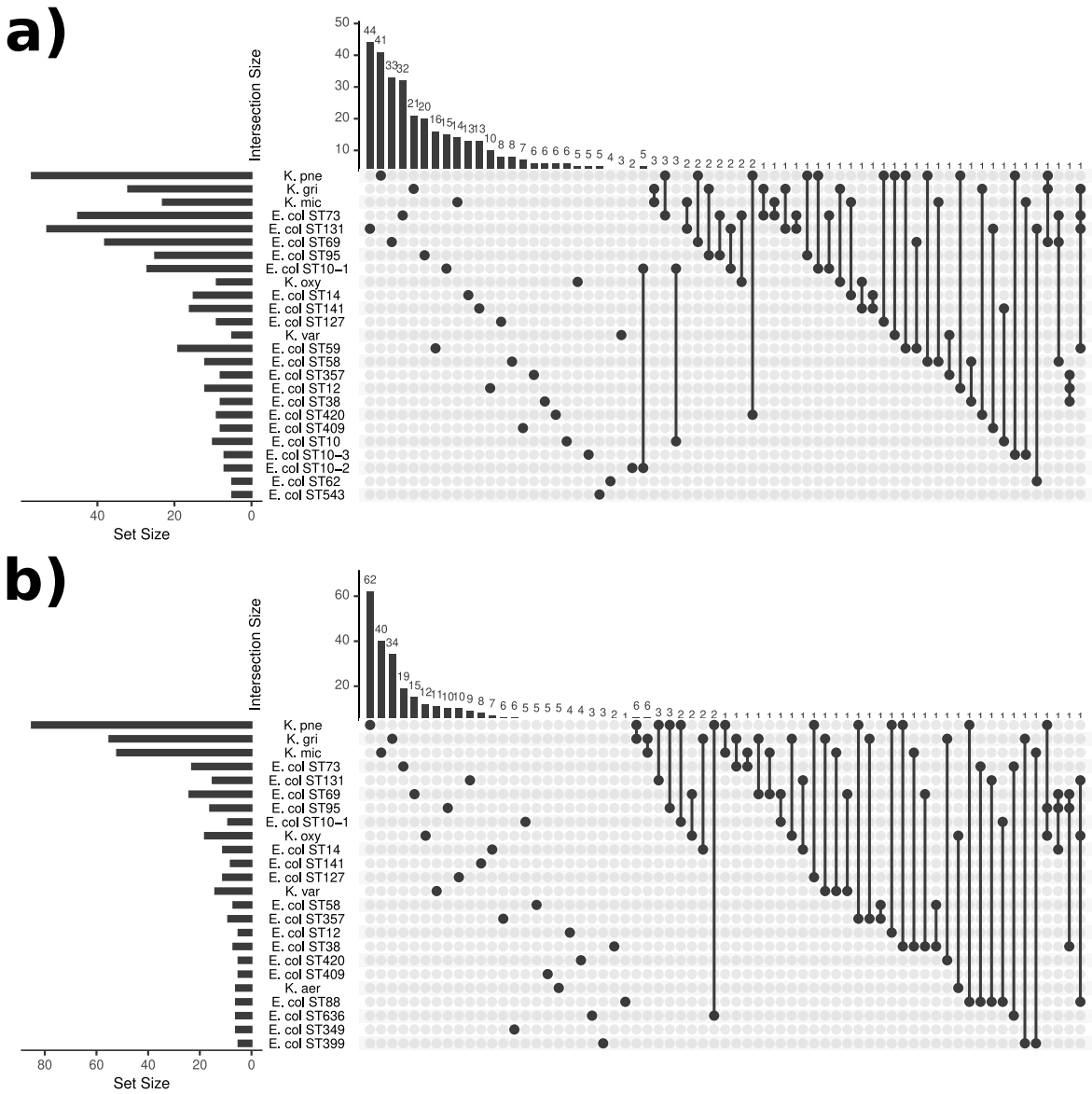

Figure 3 UpSet plot showing coexistence of *E. coli* lineages with *Klebsiella* species. The plot displays the number of times the *E. coli* lineages and various *Klebsiella* species were found either alone (single dots) or together in a sample (connected dots), and at least five times. Data are shown in panel **a)** for the vaginal delivery cohort, and in panel **b)** for the caesarean section delivery cohort. Set size (bottom-left subpanels) refers to the number of times a taxonomic unit was found in total, while intersection size (top subpanels) refers to the number of times a taxonomic unit was found alone or coexisting with other unit(s).

Neonatal gut colonisation of *E. coli* adheres to the first-come, first-served principle

A more detailed analysis was carried out to scrutinise the possible variation across the time points the *E. coli* lineages appear to colonise the neonatal gut and to what degree they are inherited from the mother. In addition, we examined whether the lineages that are successful within the first 21 days persist in the infancy period sampling 4-12 months later. Examining the colonisation-time trajectories for each individual, we found the colonisation typically to already haven taken place at the very first sampling time at 4 days in the vaginal delivery cohort (Figure 4a; 88 out of total 314 in the cohort were colonised at 4 days, 184 of whom were detected to carry *E. coli* at some point in the first 21 days). In the subsequent sampling point at 7 days after birth nearly all of the infants who had detectable amounts of *E. coli* at any time point were already colonised earlier (150 out of 184; Figure 4 and Supplementary Figure 2a) and 66 of the infants were carrying the lineage that was also observed at 4 days. In the remaining 22 individuals out of the total 88 carrying *E. coli* at 4 days, in 3 cases the lineage identified at 4 days had been replaced by another lineage and in 19 cases no *E. coli* was present in detectable amounts at 7 days.

Conversely, in the caesarean section delivery cohort there were markedly fewer *E. coli* found overall in the early time points (21 carried *E. coli* at 4 days out of total 282 in the cohort, of whom 97 had detectable amounts of *E. coli* in the first 21 days), with some signs of the initial colonisation happening slightly later at 7 days (59 out of 97; Figure 4 and Supplementary Figure 2b). Carriage of the same lineage persisted from 4 to 7 days in 11 infants out of the 21 who carried *E. coli* at 4 days, and in 23 infants from 4 or 7 days to 21 days.

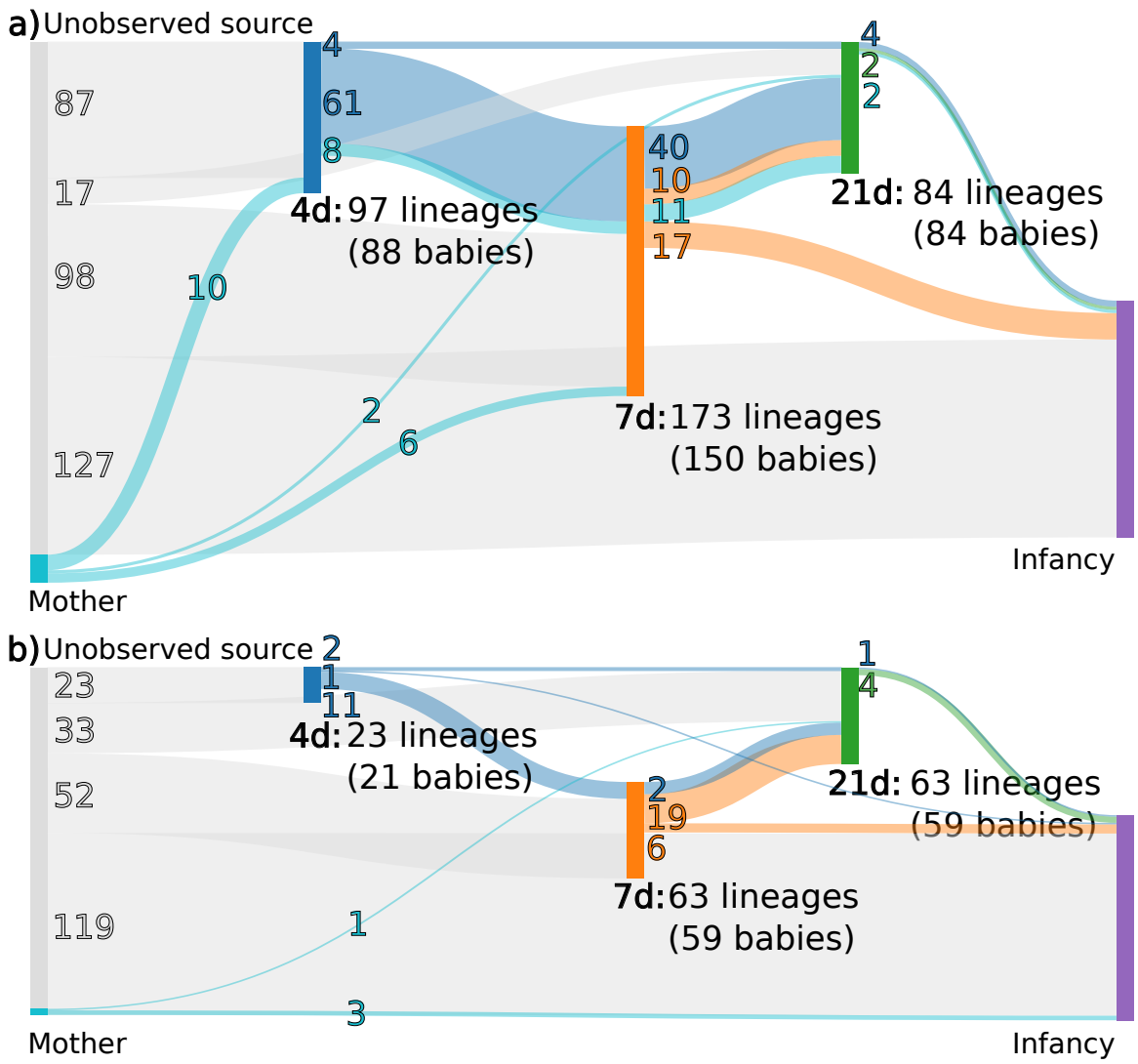

Figure 4 Summary figure of *E. coli* gut colonisation. The plot displays the numbers of *E. coli* lineages found at each of the five major sampling points (4, 7, and 21 days after birth, in the infancy period, and from the mothers in the maternity unit). The flows indicate the (possible) numbers of lineages that were transmitted between the sampling points, with the colours corresponding to the first time point the transmitted lineages were detected in. Flows that skip a sampling point (such as from 4 days to 21 days) indicate that a lineage returned to detectable levels after the skipped sampling point. Panel **a)** displays the vaginal delivery cohort, and panel **b)** the caesarean section cohort.

Comparing lineages identified in the mothers to those identified in the infants either at the time point of 4 or the 7 days in either cohort revealed 16 infants who shared the same lineage with their mothers (a total of 43 mothers out of the 175 in the full study cohort had detectable amounts of *E. coli*),

indicating potential transmission. All of the 16 potential transmissions at 4 or 7 days happened in the vaginal delivery cohort. Transmission in the caesarean section cohort was only observed in the later time points (4 observed transmissions; Figure 4 and Supplementary Figure 2b). Based on these results the majority of the initial colonisations in both cohorts appear to have been obtained from an unsampled source rather than transmitted directly from the mother, although we cannot exclude transmission of *E. coli* low-abundance (below the detection limit) lineages in the gut or from other body sites known to sometimes harbour *E. coli*, such as the vagina [17].

Examining the overall course from the first days to the final samplings at 21 days and in the infancy period revealed that, in both cohorts, the *E. coli* lineage that initially colonised the gut at the 4 or 7 days time point either persisted into the final day 21 sampling point or vanished completely (65 colonisations out of total 84 in the 21 day sampling point were already detected at 4 or 7 days in the vaginal delivery cohort; 23 out of 63 in caesarean delivery). Transitions to another lineage within the first 21 days of life were uncommon; only 4 such transitions were observed. When transitions happened, they primarily occurred between the last time point in the first 21 days time series, where *E. coli* was detected, and the infancy period sampling, where a longer period of time had passed (27 such transitions; Supplementary Figure 2).

Finally, investigation of the samples which were observed to contain multiple *E. coli* lineages (75 in the vaginal delivery cohort and 47 in caesarean, including samples from the mothers) showed that co-colonisation seldom occurred within the first 21 days with only 37 infants being co-colonized at any time point (27 in the vaginal delivery cohort and 10 in the caesarean). In the infancy period the numbers of co-colonised infants increased to 66 (35 vaginal, 31 caesarean). Co-colonisation was also somewhat more frequent in the established microbiomes of the mothers (17 mothers were co-colonised out of the 76 mothers identified carrying *E. coli*). Similar analyses were carried out for the *Klebsiella* species (Supplementary Figure 3) but they were often detectable only in a single time point despite them being commonly found in the caesarean section cohort (Figure 2), hindering further efforts to characterise *Klebsiella* persistence. Taken together with the findings from the individual colonisation-time trajectories, these results indicate a substantial

competitive advantage for the first strain to colonise the gut which lasts at least through the neonatal period (<1 months of age).

Transitions from carriage of one lineage to another rarely occur during the first weeks of life

We further detailed the dynamics of transition from carriage of one *E. coli* lineage to carriage of another by constructing an event (transmission or persistence) matrix for the lineages that were observed at least twice across all the sets of samples. The event matrix contains the numbers of times a transition from carriage of a specific lineage (rows) to carriage of another lineage or persistence of the same lineage (columns) was observed. The samples from the first 21 days (Figure 5) showed a strong preference for persistence of the first lineage to colonise the gut, with most of the events occupying the diagonal (persistence of the same lineage between two subsequent time points; 48 cases out of total 51 in the vaginal delivery cohort and 16 out of 17 in the caesarean section cohort). Including the infancy period (Supplementary Figure 4) results in slightly more variability, especially in the vaginal delivery cohort (Supplementary Figure 4 panel a), however, the observed events still mostly remain on the diagonal.

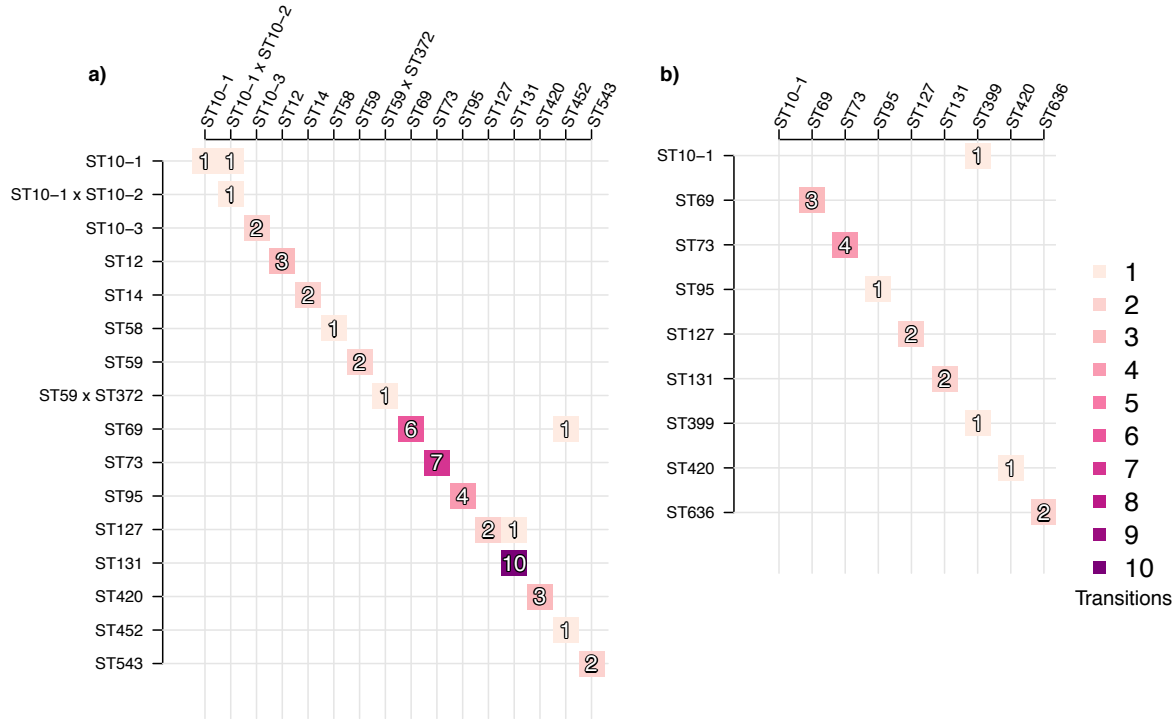

Figure 5 Event matrix displaying colonisation identities with respect to *E. coli* lineages between subsequent time points. The figure shows events corresponding to either transition from one *E. coli lineage* (rows) to another *E. coli* lineage (columns) or persistence of the same lineage (diagonal). Panel **a)** shows events for the vaginal delivery cohort with samples from the infancy period excluded, and panel **b)** shows the caesarean section delivery cohort with infancy period excluded. Darker shades of purple denote more common events, the count of which is also indicated by the number contained within the shaded boxes. Lineages shown were observed at least twice across the whole set of samples.

Colonisation potential vs invasiveness of *E. coli* sequence types

We determined the relative invasiveness of *E. coli* lineages by using odds ratios (ORs) for the frequency of each lineage in neonatal gut colonisation compared against the frequency of the same lineages in two systematic disease cohorts of *E. coli* from bloodstream infections: the NORM collection from Norway [18], and the BSAC collection from the UK [19] (see Methods for more details). These two cohorts represent the largest, and most importantly systematic and unbiased, genomic studies of *E. coli* bloodstream infections from numerous hospital districts in each study country (Norway/UK), thus

providing the necessary resolution for comparing frequencies in colonisation and disease.

This analysis resulted in confidence intervals for the ORs, which were labelled as either invasive (statistically significant confidence interval with a lower bound higher than 1), commensal (significant with an upper bound less than 1), or intermediate (confidence interval overlapping 1 or not significantly different from 1). In our results (Figure 6), the interpretations were not influenced by the choice of the disease cohort, with the exception of the intervals obtained for the multi-drug resistant sequence type ST131 (Figure 6, Supplementary Figure 5). The invasiveness of ST131 is overestimated when using the BSAC (UK) data, which was sampled during 2001-2012 when the prevalence of ST131 changed markedly.

We observed that the prevalence in carriage in our neonatal data and carriage in the NORM (Norway) and BSAC (UK) disease collections was similar for ST69 and ST131 but differed for ST73 and ST95, which were significantly overrepresented among the disease isolates from the NORM and BSAC collections (Figure 6a). In addition to these four nosocomially associated STs, numerous other lineages were observed to have ORs significantly smaller than 1, corresponding predominantly to commensal lineages, for example ST10, with a limited capacity to cause disease.

We also estimated the invasiveness of the major clades of ST131 (A, B, C1 and C2) using the combined disease collections (NORM and BSAC) and found clades A and B to be the most commensal in nature (Figure 6b). For clades C1 and C2, the ORs were more intermediate with wider confidence intervals. However, when estimated separately using either the NORM (Norway) or the BSAC (UK) collection, the confidence intervals for the C2 clade were particularly affected by which disease collection was used in the comparison. This was likely a result of known differences in prevalence over time and between the UK and Norway [18]. In total, the BSAC (UK) and NORM (Norway) collections contained 83 genomes from clade A, 101 from B, 82 from C1, and 140 from C2.

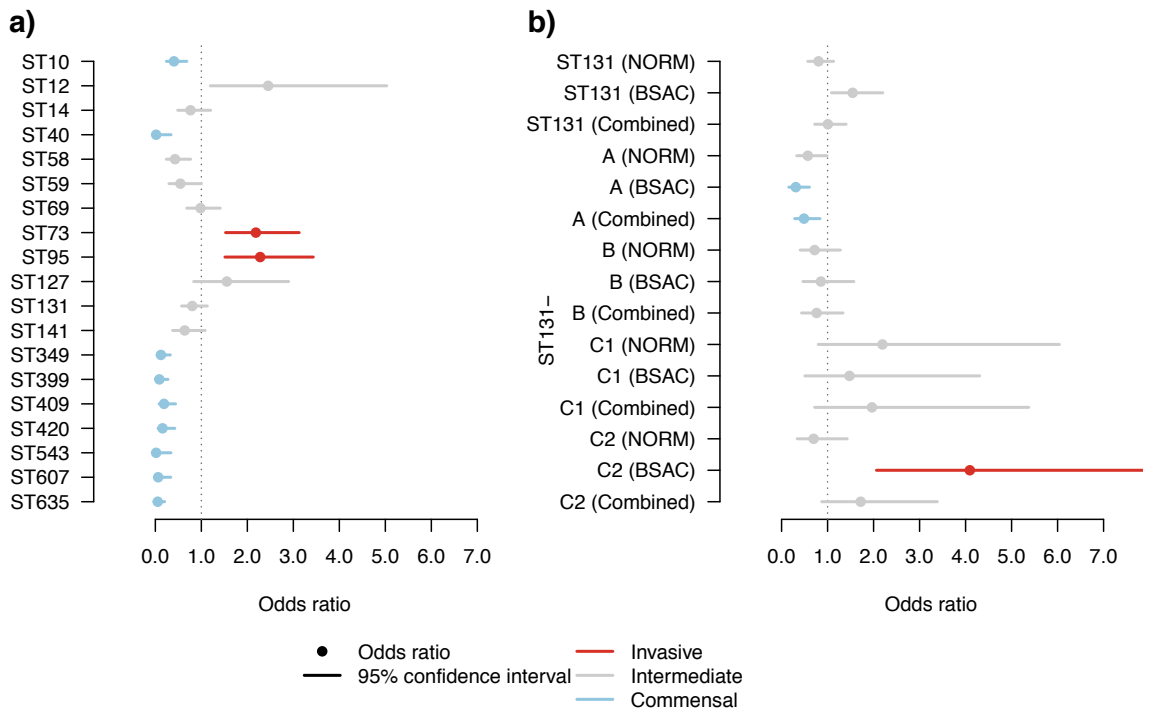

Figure 6 Odds ratios for relative *E. coli* invasiveness. The odds ratios for invasiveness are displayed with the 95% confidence interval. The confidence intervals with a statistically significant ($p < 0.05$, Pearson's chi-squared test with Benjamini-Hochberg adjustment for multiple testing) lower bounds greater than 1 correspond to more invasive lineages (coloured red), and confidence intervals with a statistically significant upper bound less than 1 to more commensal lineages (coloured light blue). Lineages with a confidence interval that contains the value 1, or are not statistically significantly different from 1, are labelled as intermediate (coloured grey). The confidence intervals are shown in panel **a)** for both the top 10 most frequent lineages in Norwegian (NORM) bloodstream infections (STs 10, 12, 14, 59, 69, 73, 95, 127, 131, and 141) and all additional lineages, where the OR was estimated to differ from 1 (significantly or not). Panel **b)** shows the confidence intervals estimated for both ST131 as a whole and additionally for its established sublineages (ST131 A, B, C1, and C2). For both ST131 and the subclades, the estimates are reported using either the BSAC (UK) collection, the NORM (Norway) collection, or combined collections for comparison.

Neonatal *E. faecalis* colonisation is not characterised by hospital-associated clusters

We further investigated the specific features of *E. faecalis* colonisation in the delivery cohorts by comparing them to a previously characterised *E. faecalis* species collection [20] using clustering and phylogenetic analyses. While the lineages identified in the neonatal cohort were widely dispersed in the species tree and many of the major STs were well represented among them, the neonates harboured none of the long-term persistent hospital-adapted lineages ST6 and ST9 and contained only a few cases of colonisation by ST28 strains (Figure 7), which have also been previously defined as hospital-adapted [20]. Furthermore, we found no significant differences ($p > 0.05$, chi-square test) between the delivery cohorts with respect to identified strains harbouring antibiotic resistance-conferring genes to major antibiotic classes (Supplementary Table 1).

Similar to the *E. coli* analysis, we examined the dynamics of transition from carriage of one *E. faecalis* lineage to carriage of another by constructing an event matrix for the lineages that were observed at least twice across the sets of samples (Figure 8). Again, this analysis revealed a strong preference for the persistence of the first strain to colonise the gut, with a vast majority of the babies not experiencing a switch between sequence types during the first weeks of life and rare appearance of co-colonisation with different sequence types. (Figure 8, Supplementary Figures 6-7).

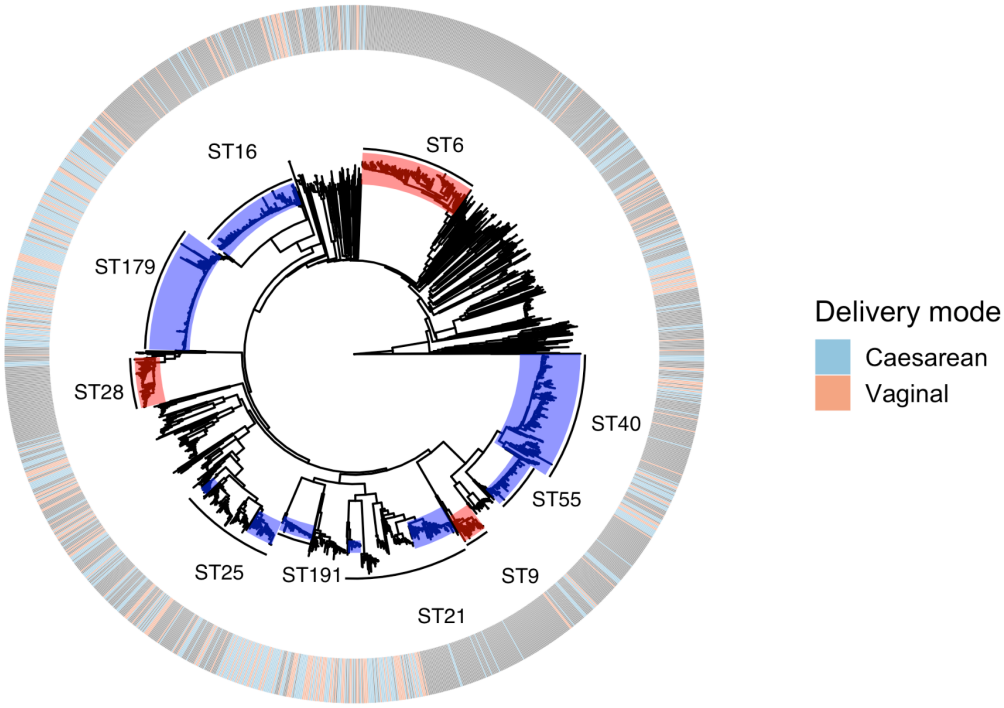

Figure 7 Embedding of the neonatal cohorts within a general *E. faecalis* species-wide collection. Outer metadata blocks depict the delivery mode of the cohorts (caesarean, light blue; vaginal, orange), aligned against the neighbour-joining (NJ) phylogeny from the core distances defined estimated with PopPUNK [21]. Ten largest sequence types in the combined collections are highlighted within the branches, and the sequence types previously defined as hospital-adapted [20] are coloured in red.

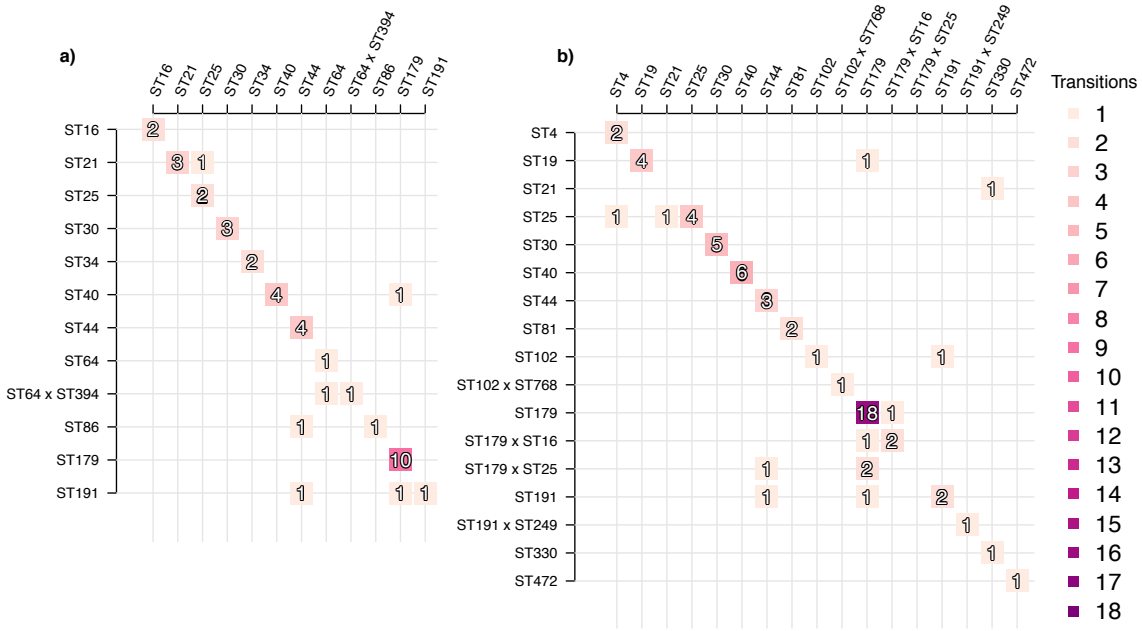

Figure 8 Event matrix displaying colonisation identities with respect to *E. faecalis* lineages between subsequent time points. The figure shows events corresponding to either transition from one *E. faecalis* lineage (rows) to another *E. faecalis* lineage (columns) or persistence of the same lineage (diagonal). Panel **a)** shows events for the vaginal delivery cohort with samples from the infancy period excluded, and panel **b)** shows the caesarean section delivery cohort with infancy period excluded. Darker shades of purple denote more common events. Lineages shown were visited at least twice across the whole set of samples.

Antimicrobial resistance and virulence genes in *Klebsiella* strains

We used Kleborate [22] to detect antimicrobial resistance (AMR) and virulence genes in the 449 *Klebsiella* assemblies obtained from the metagenomic data. We detected no *mcr* or carbapenemase genes and only 5/186 (3%) of the *K. pneumoniae* assemblies harboured a CTX-M-15 extended-spectrum beta-lactamase (ESBL) gene (Figure 9). These assemblies correspond to four individuals (two caesarean and two vaginally born). The assemblies harbouring this gene were assigned to two different sequence clusters (two to sequence cluster (SC) 1, one ST336 and one ST17; and two SC24, both ST323). Both SC1 and SC24 were found in each delivery cohort, and in one individual the same ST336 sequence was observed on both day 7 and 21, indicating sustained carriage. Of the STs detected, at least ST323 has

been previously associated with ESBL carriage and nosocomial transmission [23].

Compared to the limited number of AMR genes, we detected a more diverse set of virulence genes that were spread widely across the *Klebsiella* species (Figure 9). There were three assemblies of the hypervirulent *K. pneumoniae* ST23 clone from two individuals (both vaginally-born), which harboured colibactin (cbl), aerobactin (iuc), salmochelin (iro), and RmpADC; a suite of virulence factors that are often carried together. Clb2 was carried on ICE*Kp10*, and *iuc1*, *iro1*, *rmp1* were carried together on the virulence plasmid KpVP-1 (confirmed by Kleborate and subsequent mapping). We also observed 32 *K. pneumoniae* assemblies from 12 different SCs that harboured a diverse set of loci of the siderophore yersiniabactin (ybt). We detected other virulence factors that were very common in the more environmental species (ybt in *K. grimontii, K. michiganensis, K. ornithinolytica, K. oxytoca* and iro in *K. aeruginosa*). Although this may seem concerning, we note that these are classified as "unknown" alleles, meaning that although the genes from this locus are present (> 90% nucleotide identity and > 80% coverage), they are not the same alleles as those found in clinical strains of *K. pneumoniae*. Little is known about the prevalence of ybt and iro loci in these other *Klebsiella* species, but their presence should not be unexpected; a small study previously found that 5/11 isolates in the study (representing 5 different *Klebsiella* species) carried ybt [24], and another larger study based on whole-genome sequencing data showed that *K. ornithinolytica* commonly carries ybt [25]. More research is needed to explore the evolutionary and clinical significance of these loci.

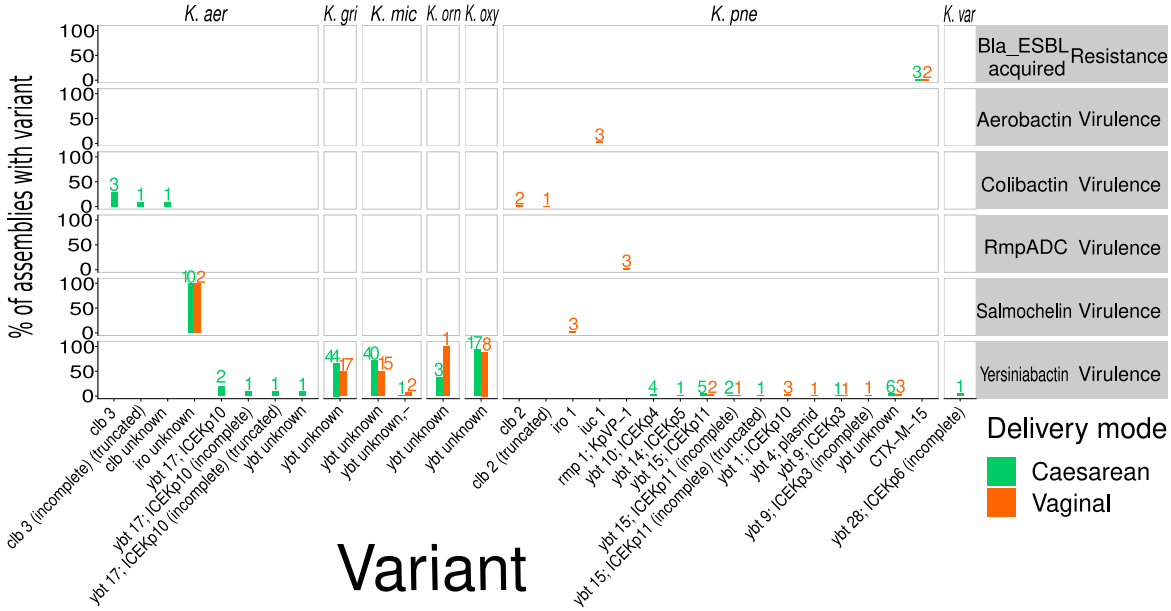

Figure 9 Summary of AMR and virulence genes in the *Klebsiella* assemblies. The plot shows frequencies of the single AMR factor and the four virulence factors identified in the *Klebsiella* species using Kleborate [22]. Sequence assemblies from the caesarean section delivered cohort are highlighted in green, and assemblies from the vaginally delivered cohort in orange.

Discussion

Colonisation of the gut by opportunistic bacterial pathogens has been a topic of intensive research for decades. Data from observational *in vivo* studies, human bacterial challenge experiments, and animal models have pointed to both antagonistic and synergistic relationships between various species. However, clear insight into the competition dynamics at both the intra- and inter-species level in the human gut has been missing in particular for the early phase of life when the gut microbiome opens up for colonisation. Here, we were able to advance this understanding thanks to the deep sequencing of neonatal stool samples in a previous landmark study [14]. Combined with recent methodology [15], [16] and high-precision genomic reference libraries, these results allowed us to identify and assemble single genomes from metagenomic sequencing data at the level of resolution required for standard bacterial genomic epidemiology, including phylogenetic analysis. The degree of resolution constitutes the main difference between our analysis and other available standard reference genome -based metagenomics approaches,

which either do not allow extracting the assembled genomes [26] or do not perform well when strain-level variation is present in the samples [27], [28].

As the purpose of our study was to unravel the competition dynamics of gut colonisation for selected key human pathogens, any possible implications for the actual host health, either during infancy or later in life, are left out of the scope of this study. For the same reason we do not reflect on our findings in light of the extensive literature on human microbiome variation in a more general context. Nevertheless, it is important to note that none of our focal pathogens are expected to cause disease in young children, unless they were immunocompromised or underwent surgery which would significantly increase the risk of the bacteria entering the bloodstream.

We found no or very few examples of nosocomially adapted lineages of *K. pneumoniae* and *E. faecalis* in the neonatal samples. Moreover, the few cases we did detect were found evenly distributed between the two delivery cohorts, suggesting that neither group is preferentially colonised with such organisms despite differences in the length of hospital stay and opportunities to become colonised during birth. In particular we saw no difference between the two cohorts in terms of the frequency of virulence or antibiotic resistance elements called from the assembled genomes of the identified *K. pneumoniae* and *E. faecalis* strains. This is despite the massive microbiome shift observed in the original study [14] that resulted in a stunted microbiome of the caesarean born babies. Taken together, this suggests that in the absence of a strong selective pressure stemming from the use of antibiotics, the hospital-adapted lineages are generally at a disadvantage when attempting to colonise healthy individuals in the hospital environment. Alternatively, the hospital-adapted lineages may have adapted to succeed in the presence of a mature microbiome and a corresponding metabolome, implying that they are strong competitors but poor pioneers. Overall, the colonisation process of a sterile infant gut appears stochastic without strong selection for the pioneering lineages, and the lineage pool associated with birth in the hospital environment differs from the known nosocomial pool that is typically sampled under antibiotic pressure.

For the most abundant organisms we were able to identify a large total number of different lineages present across the birth cohorts. Despite all this

diversity, only rarely were two or more lineages of the same species detected in any single individual at the same time point, and even transition from carrying one lineage to another lineage in the next time point was uncommon. This implies that the lineages generally compete strongly for the colonisation opportunity and the process in neonates adheres to the 'first come, first served' principle. This stands in stark contrast with the ESBL *E. coli* colonisation dynamics study performed among healthy volunteers at a hospital in Laos, where different clones were frequently replacing each other [29].

Surprisingly to us, we detected abundant diversity of species from the genus *Klebsiella*, which was particularly marked in the caesarean section delivery cohort. These differences in diversity between the cohorts were also reflected in the antagonistic relationship between *E. coli* and several of the *Klebsiella* species, the former of which was much more dominant in the vaginal delivery cohort. In both cohorts, *K. pneumoniae* was the most abundant colonising *Klebsiella* species, but it is typically considered a low-abundance member of the healthy gut microbiome and with relatively small population prevalence [30]–[32] dependent on geographical location [33]. A recent large cross-sectional study also found 16% carriage rate of *K. pneumoniae* in a general adult cohort based on selective culturing approach [34]. In our study, the high frequency and diversity of *Klebsiella* detected in the neonates implies that these species are likely to be commonly present in the adult gut to allow for such an ample transmission to the babies, but may remain undetectable with the gold standard genus-specific culturing based approach. Furthermore, similar additional variation might be hidden for *E. coli*, such that additional lineages could be present in the gut but their growth is hindered by the dominating strain(s) and thus they remain under the detection limit with the currently used sequencing depth. We do acknowledge the possibility of transmission of *Klebsiella* and *E. coli* from other body sites, however, for both species the gut is the primary site of colonisation and thus colonisation at other sites likely plays a minor role in transmission.

Our present study opened the possibility to systematically compare the colonisation abilities of *E. coli* lineages with their ability to cause bloodstream infections. Despite that such infections mostly happen in the elderly, the neonatal colonisation and competition processes are assumed to be reflective

of the population colonisation frequencies of the *E. coli* lineages in the adult gut, since diverse environmental exposure (rather than vertical transmission from mothers) is expected to be a major source of strain transmission during early life and has been observed in several pathobiont species colonising the infant gut [35], [36]. Previous attempts in the literature to estimate the level of relative invasiveness (as measured by the odds ratio of a bloodstream infection vs colonisation event) of particular strains in this respect appear to be scarce. Here we used for comparison both a Norwegian [18] and a UK [19] genomic cohort of *E. coli* bloodstream infections that were collected in a systematic manner and representative for the frequencies of lineages in the disease population.

Comparisons of the relative frequencies in colonisation vs disease revealed that the relative invasiveness of the multidrug-resistant *E. coli* ST131 was not particularly high, but its frequency in disease was overall more reflective of that in colonisation. This demonstrates the importance of matching location and time when calculating ORs for invasiveness when prevalence is known to vary temporally or geographically. Interestingly, when comparing the odds ratios at the level of sub-lineages of ST131, we found that their observed relative invasiveness was consistent with indirect estimates inferred from the phylogenetic expansion modelling in a previous study [18], with sublineages A and B determined as less invasive than sublineages C1 and C2. Out of the other successful *E. coli* sequence types frequently found in bloodstream infections, only ST73 and ST95 were found to have a significantly elevated level of invasiveness when contrasting with their estimated frequencies from our neonatal colonisation data. Additionally, ST12 showed an elevated OR but this was not statistically significant when corrected for multiple testing.

Another recent study investigated the polymicrobial nature of recurrent urinary tract infections (rUTIs) and found that *E. coli* populations in the gut and the bladder were comparable between women with and without a history of rUTI in both relative abundance and phylogroup [37]. However, a deeper lineage level analysis, such as the one performed here, combined with screening of genetic determinants of persistent bladder colonisation from the metagenome assemblies could provide a more refined characterisation of the possible population differences between gut and

the bladder. In addition, since the bladder is a more resource-poor niche compared with the gut, it is possible that the strain-level competition in the colonisation process as observed in our study would be even stronger in the bladder environment, where both efficient iron acquisition systems [38] and oxygen stress tolerance will be of heightened importance [39]. Such an analysis would nevertheless require substantially increased sequencing depths to enable high-resolution genomic epidemiology.

Added benefits of deep sequencing in analysis of heterogeneous colonisation have also been demonstrated in study of within-host diversity in pneumococcal carriage [40]. Using a similar approach to ours on sequenced DNA from plate sweeps, deep sequencing enabled investigation of implications in disease, analysis of the evolutionary responses to antibiotic treatment, identification of lineages present in samples, and performing resistance and other variant calling [40]. Given the expected continuing decrease in sequencing cost and the increased availability of relevant genomic reference libraries, we anticipate that genomic epidemiology from mixed samples, either based on direct metagenomics or on enriched or targeted DNA will offer wide future opportunities to improve our understanding of pathogen persistence, transmission and evolution.

Methods

Sequencing data

We used sequencing data from a previous study [14] that has been published in the European Nucleotide Archive under accession numbers ERP115334 (whole-genome shotgun metagenomics sequencing data) and ERP024601 (isolate sequencing data). The isolate sequencing data was only used as reference data for the reference-based methods. The WGS metagenomics data was the basis for all reported results.

Reference sequences

We combined 805 isolate assemblies from the source study for the sequencing data [14] with a bespoke reference database containing sequences for priority

pathogens (Table 2) and common commensal and contaminant species. For the species that were analysed in more detail (*E. coli, E. faecalis, Klebsiella* species), we constructed additional species-specific assembly collections, which were used to explore the within-species diversity. The *E. coli* species database consisted of data from two studies: a curated collection of ~10 000 *E. coli* sequence assemblies [41] augmented by additional ~3300 assemblies from a Norwegian bloodstream infection collection [18]. For *E. faecalis*, we used ~2000 assemblies from several European countries [20], and for the *Klebsiella* species ~3000 assemblies collected in Italy [25]. The isolate assemblies from the source study [14] belonging to *E. coli*, *E. faecalis*, or *Klebsiella* species were also included in the respective collections.

We also ran the bacterial species identification method MetaPhlAn (v3.0, [42], default options) on the metagenomics reads to check for species that had no representation in our already collected reference sequences. For these species, we downloaded the reference and representative genomes available in the NCBI database as of 30 October 2021 and added them to our reference. After collecting the full reference database as described, all of the sequences were processed with a script (available from https://github.com/tmaklin/baby-microbiome-paper-plots) that concatenated sequences consisting of several contigs by adding a 300bp gap (twice the read length in the reads) between the contigs, and collected the concatenated sequences in a single multifasta file.

Pseudoalignment index construction

We first indexed the complete multifasta file with the Themisto short-read pseudoaligner (v2.1.0, [16], no-colors option enabled) and used Themisto again (load-dbg and color-file options enabled) to colour the resulting index according to the species designation of the reference sequences. Colouring the index in this manner means that a pseudoalignment to possibly several reference sequences of the same species is reported simply as a single pseudoalignment to somewhere within the species.

For the species in the priority pathogens list (Table 2), we extracted the reference sequences belonging to each of the priority pathogen species from the multifasta file, and used Themisto (default options) to build additional

species-specific pseudoalignment indexes for these target species. In this index, we did not incorporate any colours, instead representing each sequence individually, which is the format required for the further within-species analyses.

Reference sequence grouping

For each species-specific set of reference sequences, we next used the PopPUNK [21] clustering method for bacterial genomes to identify lineages within the species that (roughly) correspond to clonal complexes. This was done by running PopPUNK (v2.4.0) multiple times for each species-specific set of reference sequences with several different options (bgmm model with the number of components K ranging from 2 to 32, the dbscan model, and model refinement for all produced bgmm models and the dbscan model), and selecting the model that had both high quality metrics as reported by PopPUNK and was also deemed consistent with the MLST designations for the species in a manual inspection. The options that produced the chosen clustering for each species are included in Table 3.

Species	**PopPUNK options chosen**
E. coli	DBSCAN + refinement
E. faecalis	DBSCAN + refinement
K. aerogenes	6-component BGMM + refinement
K. grimontii	8-component BGMM
K. huaxiensis	DBSCAN + refinement
K. michiganensis	8-component BGMM
K. ornithinolytica	DBSCAN + refinement
K. oxytoca	6-component BGMM
K. pasteurii	DBSCAN + refinement
K. planticola	3-component BGMM
K. pneumoniae	3-component BGMM

K. quasipneumoniae subsp. quasipneumoniae	-
K. quasipneumoniae subsp. similipneumoniae	3-component BGMM
K. spallanzanii	-
K. variicola	4-component BGMM + refinement

Table 3 PopPUNK options chosen for each species that was analysed at the lineage-level. DBSCAN refers to running PopPUNK with the "--fit-model dbscan" option, and the *k*-component BGMM (Bayesian Gaussian Mixture Model) to running with the "--fit-model bgmm -K *k*" option, where *k* is the number of components for the mixture model. For the species marked with a -, the clustering provided with the assemblies [25] was sufficient since no new isolate data was available, negating the need to recluster the sequences.

Lineage identification

Next, we began processing the WGS metagenomics dataset with a hierarchical approach consisting of an initial species detection step and a subsequent lineage analysis step performed for the species in the priority pathogens list. In the initial species detection step, the metagenomics reads were first pseudoaligned [43] with Themisto (v2.1.0, reverse complement handling and output sorting options enabled) against the previously constructed colored pseudoalignment index, which contained reference sequences for all species considered in our analysis (described in the preceding Reference sequences section). These pseudoalignments were used as input to mSWEEP, which is a method for estimating the relative abundances of some predefined groups in short-read sequencing data [15]. The output from mSWEEP (we used v1.6.0, [44], with the write-probs option enabled) was further processed with mGEMS (v1.2.0, [45], default options), which is a method for partitioning a set of metagenomics reads into several sets of reads (called bins), each corresponding to one of the groups that were used as the basis for relative abundance estimation in mSWEEP [16]. In this initial species detection step, the groups used were the species and we used mGEMS to extract the bins for all species in our target pathogen list.

In the subsequent lineage analysis step, we analysed the species-specific read bins created in the species detection step. If a read bin had been created for one of the species belonging to the priority pathogens list (Table 2), we pseudoaligned the reads in the bin with Themisto (v2.1.0, reverse complement handling and output sorting options enabled) against the corresponding species-specific pseudoalignment index (constructed as described in the Pseudoalignment index construction subsection). These pseudoalignments were then processed with mSWEEP, this time using the lineage definitions from PopPUNK as the target groups, to estimate the relative abundances of the lineages in the reads assigned to the species bin. We then ran mGEMS on the results from mSWEEP, creating another set of read bins that now corresponded to the lineages identified at a relative abundance greater than 0.01. Finally, we ran the quality control tool demix_check (commit 18470d3, [46]) on the resulting lineage-level bins. This tool uses the mash [47] distances between the reads assigned to each bin to provide confidence scores for whether the reads in the bin originate from the lineage they're assigned to, or if the bin was spuriously created due to the actual lineage missing from the reference data. The results from demix_check were used to filter out lineage-level bins that received a low or lowest confidence value (3 or 4 in the demix_check output). The bins that received a high or very high score (1 or 2) are called "reliable identifications" in this manuscript, and the list of these is provided in Supplementary Table 2.

Assembly

We assembled both the isolate sequencing data for their inclusion in the reference sequence collection, and the lineage-level bins obtained from the WGS metagenomics reads using mSWEEP and mGEMS as described above. The same assembly approach was used for both the isolate data and the lineage-level bins. Before assembly, the reads in question were quality controlled and corrected with the fastp preprocessing tool for fastq files (v0.23.1, [48], default settings). The corrected reads were then assembled with the shovill assembly pipeline (v1.1.0, [49], with read correction disabled).

Correlation estimation

We used the FastSpar correlation estimation tool for compositional data (v1.0.0, [50], [51] 1000 iterations with 200 exclusion iterations) to infer the correlations between species-level abundances for the WGS metagenomics data obtained with mSWEEP as described earlier. Since FastSpar requires operational taxonomic units (counts) as the input data, we converted the output from mSWEEP to count data by multiplying the relative abundance estimates for each WGS metagenomics sample with the total number of reads reported as pseudoaligned by Themisto. Statistical significance of the resulting correlation values was calculated using the permutation test functionality from FastSpar (10 000 permutations with 5 FastSpar iterations per permutation) and significance was determined at the $p < 0.05$ level.

Odds ratios for invasiveness

As the colonisation collection [14] sampled mother-infant pairs multiple times, the collection does not represent independent sampling. Because of this, we pooled the presence of a lineage in a mother-infant pair for all time points for the invasiveness analyses. PopPUNK clusters were assigned (as described earlier) to the combined NORM collection (Norwegian bloodstream infections from 2002-2017, [18]), the BSAC collection (UK bloodstream infections from 2001-2012, [19]), and the colonisation collections presented in this paper's source study (healthy newborns sampled during 2014-2017, [14]), allowing the relative frequencies of lineages to be compared between carriage and disease. We only compared lineages which were identified more than one time.

In estimating the odds ratios and their confidence intervals, for tables with any zero values, 0.5 was added to all cells in the table before calculating the odds ratios according to established methodology [52]. Statistical significance of the confidence intervals was determined with Fisher's exact test for tables with cell value <5, and otherwise the chi-squared test was used. An adjustment for multiple testing was made using the Benjamini-Hochberg method [53]. In the ST131 sublineage analysis, SKA [54] was used to generate an alignment and a subsequent tree embedding the ST131 carriage assemblies from this study and the ST131 assemblies from both the NORM (Norway) collection [18] and the

BSAC (UK) collection [19], from which clade membership of the assemblies in this study could be inferred.

Comparative analysis of *E. faecalis* population structures

We used the query option from PopPUNK (v.2.2.0, [21]) to embed the *E. faecalis* assemblies from the neonatal cohort into an alignment-free clustering of *E. faecalis* sequences from another collection [20]. This allowed us to compare the differences between these two collections by unifying the clustering across them. Then, we constructed a comparative neighbour-joining phylogeny of both the neonatal cohorts and the other *E. faecalis* collection [20] from core distances estimated with PopPUNK [21]. Multi-locus sequence types were retrieved from the assemblies using fastMLST v.0.0.15 [55], and antibiotic resistance profiles screened using AMRFinderPlus v.3.10.18 [56] against the NCBI database (version 2021-12-21.1) with the '--plus' option enabled and with minimum identity of 75% and minimum coverage of 80%. The differences in presence of genes conferring resistance to major antibiotic classes (aminoglycosides, macrolides, lincosamides, tetracyclines, and phenicols) between the different delivery modes (caesarean vs vaginal) were compared by using Pearson's chi-squared test with Benjamini-Hochberg adjustment for multiple testing in R v.4.2.0 [57]. Vancomycin (glycopeptide) was omitted as none of the assemblies from the isolate sequencing data in the neonate source study [14] harboured *van* genes.

Detection of AMR and virulence genes in *Klebsiella* strains

We used the Kleborate tool (v2.1.0 [22]) to detect and screen for AMR and virulence genes in the assemblies created from the mGEMS lineage-level bins belonging to the *Klebsiella* genus. Compared to other approaches such as AMRFinderPlus, Kleborate is a specialised tool for genomic profiling of sequences specifically from the *Klebsiella* genus that leverages a curated database for this particular genus.

Data availability

Sequencing data used are available from the European Nucleotide Archive under accession numbers ERP115334 (whole-genome shotgun metagenomics sequencing data) and ERP024601 (isolate sequencing data).

The species-level pseudoalignment index for Themisto v2.1.0, the *E. coli*, *E. faecalis*, and *Klebsiella* species-specific indexes are all available from Zenodo (species-level index doi: 10.5281/zenodo.6656881, *E. coli* index doi: 10.5281/zenodo.6656897, *E. faecalis* index: 10.5281/zenodo.6656903, *Klebsiella* index: 10.5281/zenodo.6656911). Results from the mGEMS pipeline which were assigned high or very high confidence scores using demix_check, forming the core of the analyses presented, are listed in Supplementary Table 2.

Code availability

Figures 1-3 and Figures 5-8 were created using R v4.0.5 [57]. Scripts used to create the visualisations and batch job scripts containing the commands and options used to run the analyses are available from GitHub at https://github.com/tmaklin/baby-microbiome-paper-plots. The UpSet plot [58] in Figure 3 was created using the UpSetR package (v1.4.0, [59]). Figure 4 was created using SankeyMATIC (https://github.com/nowthis/sankeymatic).

Acknowledgements

The authors wish to thank the Finnish Grid and Cloud Infrastructure (FGCI) for supporting this project with computational and data storage resources. J.C. and H.T. were funded by ERC grant no. 742158 and J.C. additionally by NFR grant no. 299941 and Academy of Finland EuroHPC grant. J.C. and A.H. were supported by the Academy of Finland Flagship Finnish Center for Artificial Intelligence FCAI. R.A.G. and A.K.P. were funded by the AMR grant from Trond Mohn Foundation. Y.S. and T.D.L. are supported by the Wellcome Trust (206194 and 108413/A/15/D).

References

[1] O. Tenaillon, D. Skurnik, B. Picard, and E. Denamur, “The population genetics of commensal *Escherichia coli*,” *Nat. Rev. Microbiol.*, vol. 8, no. 3, pp. 207–217, Mar. 2010, doi: 10.1038/nrmicro2298.

[2] F. Ørskov, I. Ørskov, D. J. Evans, R. B. Sack, D. A. Sack, and T. Wadström, “Special *Escherichia coli* serotypes among enterotoxigenic strains from diarrhoea in adults and children,” *Med. Microbiol. Immunol. (Berl.)*, vol. 162, no. 2, pp. 73–80, Jun. 1976, doi: 10.1007/BF02121318.

[3] D. A. Caugant, B. R. Levin, and R. K. Selander, “Genetic diversity and temporal variation in the *E. coli* population of a human host,” *Genetics*, vol. 98, no. 3, pp. 467–490, Jul. 1981, doi: 10.1093/genetics/98.3.467.

[4] D. A. Caugant, B. R. Levin, and R. K. Selander, “Distribution of multilocus genotypes of *Escherichia coli* within and between host families,” *J. Hyg. (Lond.)*, vol. 92, no. 3, pp. 377–384, Jun. 1984, doi: 10.1017/S0022172400064597.

[5] A. Ortiz, N. M. Vega, C. Ratzke, and J. Gore, “Interspecies bacterial competition regulates community assembly in the *C. elegans* intestine,” *ISME J.*, vol. 15, no. 7, pp. 2131–2145, Jul. 2021, doi: 10.1038/s41396-021-00910-4.

[6] M. L. Patnode *et al.*, “Interspecies Competition Impacts Targeted Manipulation of Human Gut Bacteria by Fiber-Derived Glycans,” *Cell*, vol. 179, no. 1, pp. 59-73.e13, Sep. 2019, doi: 10.1016/j.cell.2019.08.011.

[7] E. Deriu *et al.*, “Probiotic Bacteria Reduce *Salmonella Typhimurium* Intestinal Colonization by Competing for Iron,” *Cell Host Microbe*, vol. 14, no. 1, pp. 26–37, Jul. 2013, doi: 10.1016/j.chom.2013.06.007.

[8] R. A. Oliveira *et al.*, “*Klebsiella michiganensis* transmission enhances resistance to Enterobacteriaceae gut invasion by nutrition competition,” *Nat. Microbiol.*, vol. 5, no. 4, pp. 630–641, Apr. 2020, doi: 10.1038/s41564-019-0658-4.

[9] A. Onderdonk, B. Marshall, R. Cisneros, and S. B. Levy, “Competition between congenic *Escherichia coli* K-12 strains in vivo,” *Infect. Immun.*, vol. 32, no. 1, pp. 74–79, Apr. 1981, doi: 10.1128/iai.32.1.74-79.1981.

[10] L. H. Lam and D. M. Monack, “Intraspecies Competition for Niches in the Distal Gut Dictate Transmission during Persistent Salmonella Infection,” *PLoS Pathog.*, vol. 10, no. 12, p. e1004527, Dec. 2014, doi: 10.1371/journal.ppat.1004527.

[11] X. Fang *et al.*, “Metagenomics-Based, Strain-Level Analysis of *Escherichia coli* From a Time-Series of Microbiome Samples From a Crohn’s Disease Patient,” *Front. Microbiol.*, vol. 9, p. 2559, Oct. 2018, doi: 10.3389/fmicb.2018.02559.

[12] T. K. S. Richter, J. M. Michalski, L. Zanetti, S. M. Tennant, W. H. Chen, and D. A. Rasko, “Responses of the Human Gut *Escherichia coli* Population to Pathogen and Antibiotic Disturbances,” *mSystems*, vol. 3, no. 4, pp. e00047-18, Aug. 2018, doi: 10.1128/mSystems.00047-18.

[13] M. G. J. de Vos, M. Zagorski, A. McNally, and T. Bollenbach, "Interaction networks, ecological stability, and collective antibiotic tolerance in polymicrobial infections," *Proc. Natl. Acad. Sci.*, vol. 114, no. 40, pp. 10666–10671, Oct. 2017, doi: 10.1073/pnas.1713372114.

[14] Y. Shao *et al.*, "Stunted microbiota and opportunistic pathogen colonization in caesarean-section birth," *Nature*, vol. 574, no. 7776, pp. 117–121, Oct. 2019, doi: 10.1038/s41586-019-1560-1.

[15] T. Mäklin *et al.*, "High-resolution sweep metagenomics using fast probabilistic inference," *Wellcome Open Res.*, vol. 5, p. 14, Oct. 2021, doi: 10.12688/wellcomeopenres.15639.2.

[16] T. Mäklin *et al.*, "Bacterial genomic epidemiology with mixed samples," *Microb. Genomics*, vol. 7, no. 11, Nov. 2021, doi: 10.1099/mgen.0.000691.

[17] J. R. Brannon *et al.*, "Invasion of vaginal epithelial cells by uropathogenic *Escherichia coli*," *Nat. Commun.*, vol. 11, no. 1, p. 2803, Dec. 2020, doi: 10.1038/s41467-020-16627-5.

[18] R. A. Gladstone *et al.*, "Emergence and dissemination of antimicrobial resistance in *Escherichia coli* causing bloodstream infections in Norway in 2002–17: a nationwide, longitudinal, microbial population genomic study," *Lancet Microbe*, vol. 2, no. 7, pp. e331–e341, Jul. 2021, doi: 10.1016/S2666-5247(21)00031-8.

[19] T. Kallonen *et al.*, "Systematic longitudinal survey of invasive *Escherichia coli* in England demonstrates a stable population structure only transiently disturbed by the emergence of ST131," *Genome Res.*, vol. 27, no. 8, pp. 1437–1449, Aug. 2017, doi: 10.1101/gr.216606.116.

[20] A. K. Pöntinen *et al.*, "Apparent nosocomial adaptation of *Enterococcus faecalis* predates the modern hospital era," *Nat. Commun.*, vol. 12, no. 1, p. 1523, Dec. 2021, doi: 10.1038/s41467-021-21749-5.

[21] J. A. Lees *et al.*, "Fast and flexible bacterial genomic epidemiology with PopPUNK," *Genome Res.*, vol. 29, no. 2, pp. 304–316, Feb. 2019, doi: 10.1101/gr.241455.118.

[22] M. M. C. Lam, R. R. Wick, S. C. Watts, L. T. Cerdeira, K. L. Wyres, and K. E. Holt, "A genomic surveillance framework and genotyping tool for *Klebsiella pneumoniae* and its related species complex," *Nat. Commun.*, vol. 12, no. 1, p. 4188, Dec. 2021, doi: 10.1038/s41467-021-24448-3.

[23] C. L. Gorrie *et al.*, "Genomic dissection of *Klebsiella pneumoniae* infections in hospital patients reveals insights into an opportunistic

pathogen," *Nat. Commun.*, vol. 13, no. 1, p. 3017, Dec. 2022, doi: 10.1038/s41467-022-30717-6.

[24] S. Bach, A. Almeida, and E. Carniel, "The *Yersinia* high-pathogenicity island is present in different members of the family *Enterobacteriaceae*," *FEMS Microbiol. Lett.*, vol. 183, no. 2, pp. 289–294, Feb. 2000, doi: 10.1111/j.1574-6968.2000.tb08973.x.

[25] H. Thorpe *et al.*, "One Health or Three? Transmission modelling of *Klebsiella* isolates reveals ecological barriers to transmission between humans, animals and the environment," *bioRxiv*, preprint, Aug. 2021. doi: 10.1101/2021.08.05.455249.

[26] L. R. van Dijk *et al.*, "StrainGE: a toolkit to track and characterize low-abundance strains in complex microbial communities," *Genome Biol.*, vol. 23, no. 1, p. 74, Dec. 2022, doi: 10.1186/s13059-022-02630-0.

[27] A. Sczyrba *et al.*, "Critical Assessment of Metagenome Interpretation—a benchmark of metagenomics software," *Nat. Methods*, vol. 14, no. 11, pp. 1063–1071, Nov. 2017, doi: 10.1038/nmeth.4458.

[28] F. Meyer *et al.*, "Critical Assessment of Metagenome Interpretation: the second round of challenges," *Nat. Methods*, vol. 19, no. 4, pp. 429–440, Apr. 2022, doi: 10.1038/s41592-022-01431-4.

[29] A. Kantele *et al.*, "Dynamics of intestinal multidrug-resistant bacteria colonisation contracted by visitors to a high-endemic setting: a prospective, daily, real-time sampling study," *Lancet Microbe*, vol. 2, no. 4, pp. e151–e158, Apr. 2021, doi: 10.1016/S2666-5247(20)30224-X.

[30] Y. Sun *et al.*, "Measurement of *Klebsiella* Intestinal Colonization Density To Assess Infection Risk," *mSphere*, vol. 6, no. 3, pp. e00500-21, Jun. 2021, doi: 10.1128/mSphere.00500-21.

[31] C. L. Gorrie *et al.*, "Gastrointestinal Carriage Is a Major Reservoir of Klebsiella pneumoniae Infection in Intensive Care Patients," *Clin. Infect. Dis.*, vol. 65, no. 2, pp. 208–215, Jul. 2017, doi: 10.1093/cid/cix270.

[32] R. M. Martin *et al.*, "Molecular Epidemiology of Colonizing and Infecting Isolates of *Klebsiella pneumoniae*," *mSphere*, vol. 1, no. 5, pp. e00261-16, Oct. 2016, doi: 10.1128/mSphere.00261-16.

[33] Y.-T. Lin *et al.*, "Seroepidemiology of Klebsiella pneumoniae colonizing the intestinal tract of healthy chinese and overseas chinese adults in Asian countries," *BMC Microbiol.*, vol. 12, no. 1, p. 13, Dec. 2012, doi: 10.1186/1471-2180-12-13.

[34] N. Raffelsberger *et al.*, “Gastrointestinal carriage of *Klebsiella pneumoniae* in a general adult population: a cross-sectional study of risk factors and bacterial genomic diversity,” *Gut Microbes*, vol. 13, no. 1, p. 1939599, Jan. 2021, doi: 10.1080/19490976.2021.1939599.

[35] K. Korpela *et al.*, “Selective maternal seeding and environment shape the human gut microbiome,” *Genome Res.*, vol. 28, no. 4, pp. 561–568, Apr. 2018, doi: 10.1101/gr.233940.117.

[36] P. Ferretti *et al.*, “*C. difficile* is overdiagnosed in adults and a commensal in infants,” *bioRxiv*, preprint, Feb. 2022. doi: 10.1101/2022.02.16.480740.

[37] C. J. Worby *et al.*, “Longitudinal multi-omics analyses link gut microbiome dysbiosis with recurrent urinary tract infections in women,” *Nat. Microbiol.*, vol. 7, no. 5, pp. 630–639, May 2022, doi: 10.1038/s41564-022-01107-x.

[38] S. D. Himpsl *et al.*, “The oxidative fumarase FumC is a key contributor for E. coli fitness under iron-limitation and during UTI,” *PLOS Pathog.*, vol. 16, no. 2, p. e1008382, Feb. 2020, doi: 10.1371/journal.ppat.1008382.

[39] S. Subashchandrabose and H. L. T. Mobley, “Virulence and Fitness Determinants of Uropathogenic *Escherichia coli*,” *Microbiol. Spectr.*, vol. 3, no. 4, p. 3.4.20, Jul. 2015, doi: 10.1128/microbiolspec.UTI-0015-2012.

[40] G. Tonkin-Hill *et al.*, “Pneumococcal within-host diversity during colonisation, transmission and treatment,” *Nat. Microbiol.*, in press 2022.

[41] G. Horesh, G. A. Blackwell, G. Tonkin-Hill, J. Corander, E. Heinz, and N. R. Thomson, “A comprehensive and high-quality collection of Escherichia coli genomes and their genes,” *Microb. Genomics*, vol. 7, no. 2, Feb. 2021, doi: 10.1099/mgen.0.000499.

[42] F. Beghini *et al.*, “Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3,” *eLife*, vol. 10, p. e65088, May 2021, doi: 10.7554/eLife.65088.

[43] N. L. Bray, H. Pimentel, P. Melsted, and L. Pachter, “Near-optimal probabilistic RNA-seq quantification,” *Nat. Biotechnol.*, vol. 34, no. 5, pp. 525–527, May 2016, doi: 10.1038/nbt.3519.

[44] T. Mäklin and A. Honkela, *PROBIC/mSWEEP: mSWEEP-v1.6.0 (15 November 2021)*. Zenodo, 2022. doi: 10.5281/ZENODO.6523380.

[45] T. Mäklin, *PROBIC/mGEMS: mGEMS-v1.2.0 (20 November 2021)*. Zenodo, 2021. doi: 10.5281/ZENODO.5715888.

[46] H. Thorpe, *harry-thorpe/demix_check: demix_check 18470d3 (25 October*

2021). GitHub, 2021. [Online]. Available: https://github.com/harry-thorpe/demix_check

[47] B. D. Ondov *et al.*, "Mash: fast genome and metagenome distance estimation using MinHash," *Genome Biol.*, vol. 17, no. 1, p. 132, Dec. 2016, doi: 10.1186/s13059-016-0997-x.

[48] S. Chen, Y. Zhou, Y. Chen, and J. Gu, "fastp: an ultra-fast all-in-one FASTQ preprocessor," *Bioinformatics*, vol. 34, no. 17, pp. i884–i890, Sep. 2018, doi: 10.1093/bioinformatics/bty560.

[49] T. Seemann, *tseemann/Shovill: Shovill-v1.1.0 (13 March 2020)*. GitHub, 2020. GitHub, 2021. [Online]. Available: https://github.com/tseemann/shovill

[50] J. Friedman and E. J. Alm, "Inferring Correlation Networks from Genomic Survey Data," *PLoS Comput. Biol.*, vol. 8, no. 9, p. e1002687, Sep. 2012, doi: 10.1371/journal.pcbi.1002687.

[51] S. C. Watts, S. C. Ritchie, M. Inouye, and K. E. Holt, "FastSpar: rapid and scalable correlation estimation for compositional data," *Bioinformatics*, vol. 35, no. 6, pp. 1064–1066, Mar. 2019, doi: 10.1093/bioinformatics/bty734.

[52] A. B. Brueggemann, D. T. Griffiths, E. Meats, T. Peto, D. W. Crook, and B. G. Spratt, "Clonal Relationships between Invasive and Carriage *Streptococcus pneumoniae* and Serotype- and Clone-Specific Differences in Invasive Disease Potential," *J. Infect. Dis.*, vol. 187, no. 9, pp. 1424–1432, May 2003, doi: 10.1086/374624.

[53] Y. Benjamini and Y. Hochberg, "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *J. R. Stat. Soc. Ser. B Methodol.*, vol. 57, no. 1, pp. 289–300, Jan. 1995, doi: 10.1111/j.2517-6161.1995.tb02031.x.

[54] S. R. Harris, "SKA: Split Kmer Analysis Toolkit for Bacterial Genomic Epidemiology," *bioRxiv*, preprint, Oct. 2018. doi: 10.1101/453142.

[55] E. Guerrero-Araya, M. Muñoz, C. Rodríguez, and D. Paredes-Sabja, "FastMLST: A Multi-core Tool for Multilocus Sequence Typing of Draft Genome Assemblies," *Bioinforma. Biol. Insights*, vol. 15, p. 117793222110592, Jan. 2021, doi: 10.1177/11779322211059238.

[56] M. Feldgarden *et al.*, "AMRFinderPlus and the Reference Gene Catalog facilitate examination of the genomic links among antimicrobial resistance, stress response, and virulence," *Sci. Rep.*, vol. 11, no. 1, p.

12728, Dec. 2021, doi: 10.1038/s41598-021-91456-0.

[57] R Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, 2021. [Online]. Available: https://www.R-project.org

[58] A. Lex, N. Gehlenborg, H. Strobelt, R. Vuillemot, and H. Pfister, "UpSet: Visualization of Intersecting Sets," *IEEE Trans. Vis. Comput. Graph.*, vol. 20, no. 12, pp. 1983–1992, Dec. 2014, doi: 10.1109/TVCG.2014.2346248.

[59] J. R. Conway, A. Lex, and N. Gehlenborg, "UpSetR: an R package for the visualization of intersecting sets and their properties," *Bioinformatics*, vol. 33, no. 18, pp. 2938–2940, Sep. 2017, doi: 10.1093/bioinformatics/btx364.