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 quantification of the competition 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* (ExPEC), which is the predominant facultative anaerobe in the large intestine [1]. Work on *E. coli* going back to several decades suggests strong intra-species competition in healthy colonisation based on serotypic variation [2], or the lack of thereof. Multi-locus enzyme electrophoresis (MLEE) 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.

The role of intra-species competition in colonisation has been widely studied in animal models, with certain interesting results related to the species that also colonise the human gut microbiome. For example, *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 tracked the microbiome composition in general, and competition among *E. coli* strains in particular, 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 clones were seen taking the role of the dominant *E. coli* strain in the microbiome over time, but return to a previously identified strain was never observed [11]. In another study, the diversity of *E. coli* colonising the gut prior, during and after an ecological disruption was highlighted in an entero-toxigenic *E. coli* (ETEC) 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 (UTI) 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. We performed strain identification and genome assembly for a set of species with pathogenic potential by de-mixing metagenomes from a longitudinal cohort of neonatal gut microbiome samples [14]. Stool samples were collected from the neonate cohort at 4, 7, and 21 days after birth, and later during the infancy period for a subcohort. 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 gut microbiome composition [14]. The study in question [14] performed metagenomic sequencing at an unprecedented depth, which combined with recent analytical advances in genomic epidemiology of mixed samples [15],

[16], and the vastly improved availability of large and high-quality genome libraries from studies of key human pathogens based on whole-genome sequencing of 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 short reads from gut metagenomes based on stool samples that had been sequenced as part of a previously published study on opportunistic pathogen colonisation in newborn babies representing two distinct delivery cohorts: vaginally and caesarean section born babies [14]. Our study extends the previous analysis by providing lineage-level characterization and subsequent genome assembly from these samples for several important pathogen species (Supplementary Table 1) as well as a more detailed exploration of the diversity within the *Klebsiella* genus. In both the lineage-level characterization and the *Klebsiella* species analysis we applied the recent mSWEEP and mGEMS methods [15], [16] with a bespoke set of reference sequences (Methods section). The analysis pipeline is described in more detail in Supplementary Figure 1 and in the Methods section. Comprehensive sets of results from the analyses will be presented in the following sections.

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*. 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 1a). 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) 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 *S. 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.

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 panels a and b). 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 panels a and b). 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.

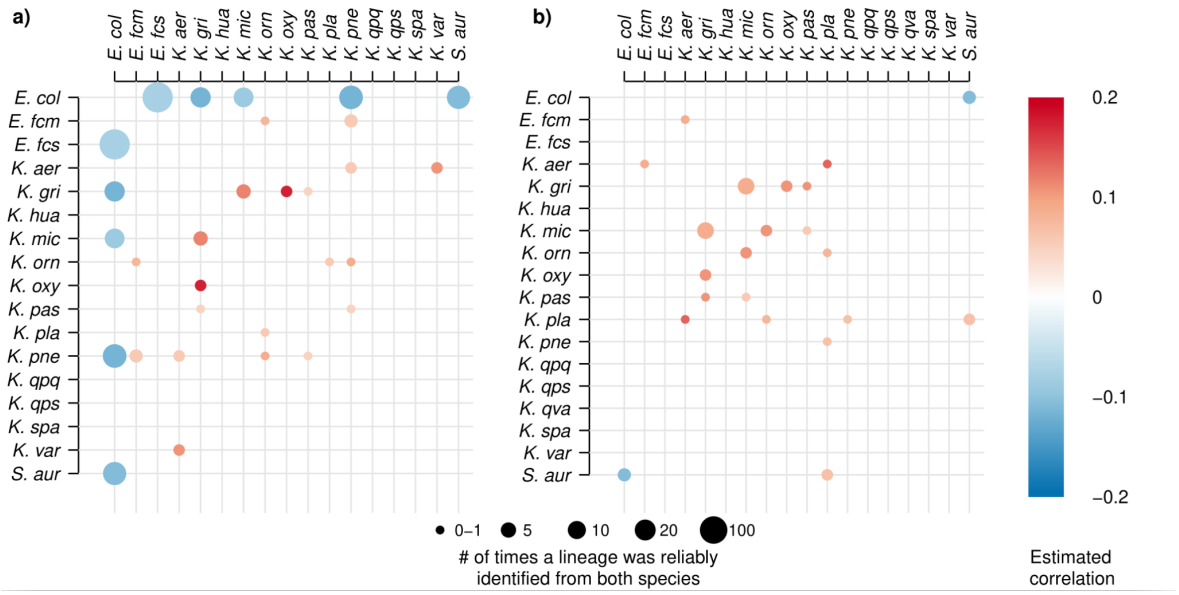

Figure 1 Correlations between the identified priority pathogens. The figure shows statistically significant ($p < 0.05$, permutation test) positive and negative correlations for the focus pathogen species. Correlations are shown within the vaginal delivery cohort (Panel **a)**) and within the caesarean section delivery cohort (Panel **b)**). Darker shades of red and blue represent stronger positive/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).

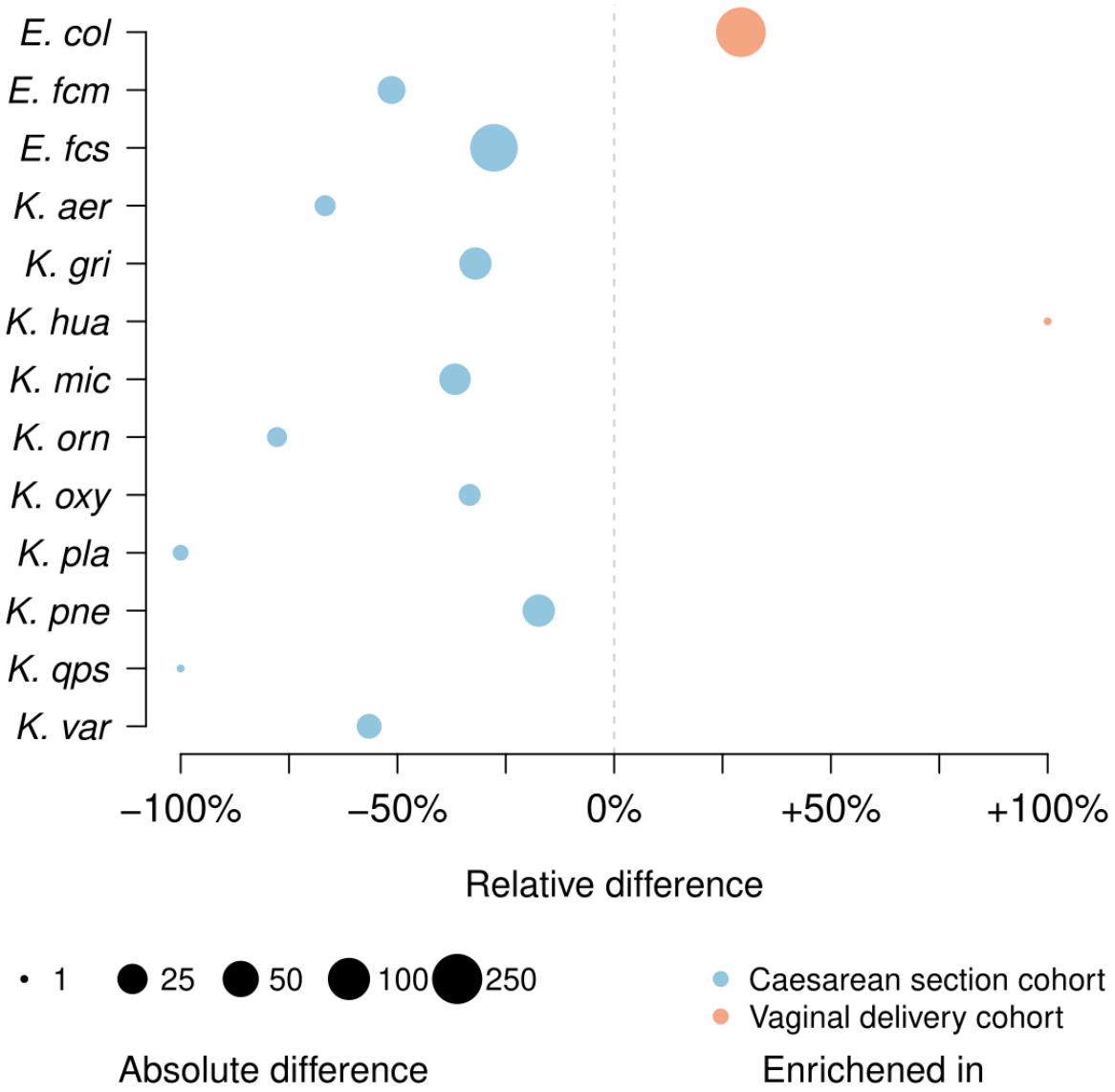

Figure 2 Differences in pathogen loads between cohorts. The figure shows differences in the number of reliably identified pathogens in each cohort. Area of the circles displays the absolute difference between the cohorts, while the horizontal placement of the circles displays the relative difference. 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 a total of six times each in the caesarean section delivered cohort and in the case of the former were also established as synergistic in the correlation analysis (Figure 1).

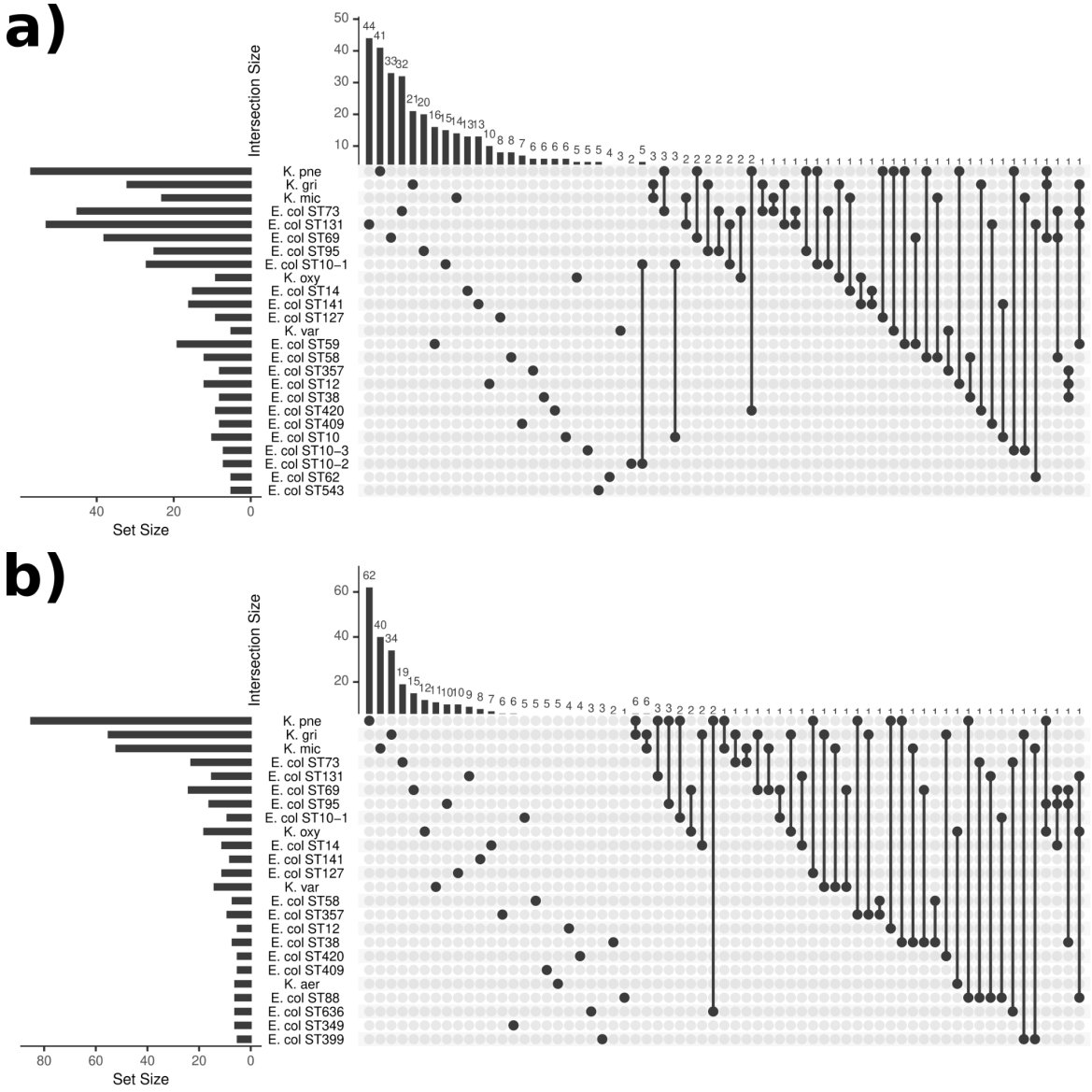

Figure 3 UpSet plot showing coexistence of *E. coli* lineages with *Klebsiella* species. The plot displays the number of times the *E. coli* lineages and various *Klebsiella* species were found either alone (single dots) or together in a sample (connected dots) at least five times in the plotted samples. Data are shown in panel **a)** for the vaginally delivery cohort, and in panel **b)** for the caesarean section delivery cohort. Set size (bottom-left panel) refers to the number of times a taxonomic unit was found in total, while intersection size (top panel) 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 to scrutinise possible variation across the time point the *E. coli* lineages appear to colonise the neonatal gut, to what degree they are inherited from the mother, and 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 (88 out of total 314 in the cohort, 184 of whom were detected to carry *E. coli* at some point in the first 21 days). In the next sampling at 7 days after birth nearly all of the infants who had detectable amounts of *E. coli* at any time point were already colonised (150 out of 184; Supplementary Figure 2a) and 66 of the infants were carrying the lineage that was also observed at 4 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; Supplementary Figure 2b). Carriage of the same lineage persisted in 11 infants out of the 21 who carried *E. coli* at 4 days.

When comparing lineages identified in the mothers to those identified in the infants either at the 4 or the 7 days time point, we found 16 infants who shared the same lineage with their mothers (out of total 43 mothers who 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; Supplementary Figure 2b). Based on these results the majority of the initial colonisations in both cohorts appear to have been obtained from the environment rather than transmitted from the mother.

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. Transitions to another lineage within the first 21 days of life were uncommon with only 4 such transitions observed. When transitions happened, they primarily occurred between the final time point from each newborn 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 showed that co-colonisation seldom occurs 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). Similarly, co-colonisation is somewhat 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 examined 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 the sets of samples. The samples from the first 21 days (Figure 4 panels a and b) 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). Including the infancy period (Figure 4 panels c and d) results in slightly more variability, especially

in the vaginal delivery cohort (Figure 4 panel c), however, the observed events still remain on the diagonal.

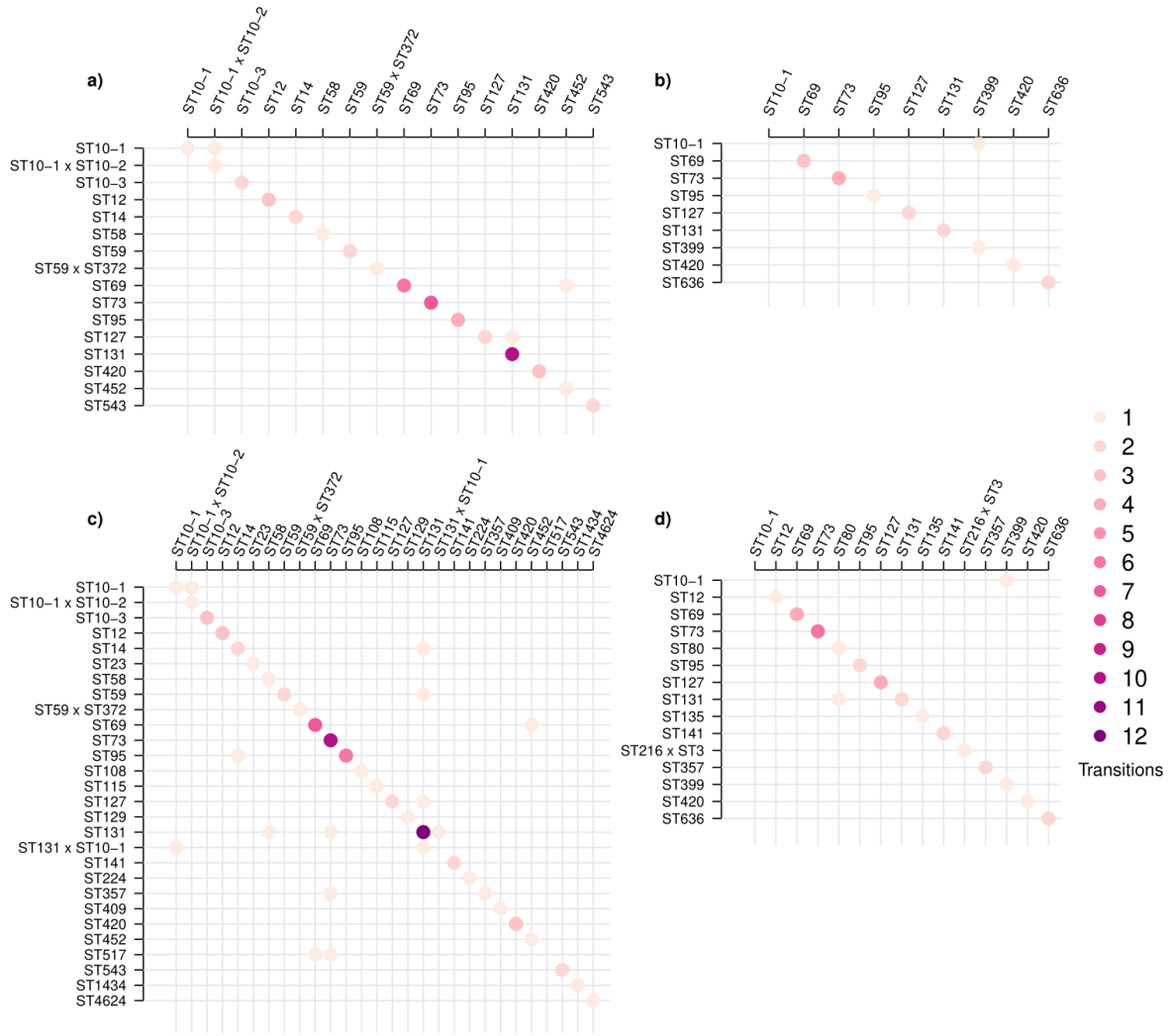

Figure 4 Event matrix displaying colonisation identities with respect to *E. coli* lineages between subsequent time points. The figure shows events corresponding to either transition from one *E. coli lineage* (rows) to another *E. coli* lineage (columns) or persistence of the same lineage (diagonal). Panel **a)** shows events for the vaginal delivery cohort with samples from the infancy period excluded, panel **b)** shows the caesarean section delivery cohort with infancy period excluded, panel **c)** the vaginal delivery cohort with the infancy period included, and panel **d)** the caesarean section delivery cohort with the infancy period included. Darker shades of purple denote more common events. Lineages shown were visited 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 using odds ratios (ORs) for the frequency of each lineage in neonatal gut colonisation compared to two systematic genomic cohorts of *E. coli* from bloodstream infections, NORM [17], and BSAC [18]. The interpretations of the ORs (>1 more invasive, <1 more commensal) were not influenced by the choice of infection cohort, with the exception of the multi-drug resistant sequence type ST131 (Supplementary Figure 4). For the BSAC data, that sampled years during which the prevalence of ST131 changed markedly, ST131's invasiveness is overestimated (Supplementary Figure 4).

We observed that the prevalence in carriage and the NORM disease collection was similar for ST69 and ST131, unlike ST73 and ST95 which were significantly overrepresented among disease isolates (Figure 5a). 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 and found A and B to be the most commensal in nature (Figure 5b). C1 and C2 ORs were more intermediate with wider confidence intervals. C2 estimates were particularly affected by which disease collection was used in the comparison due to known differences in prevalence over time and between the UK and Norway [17].

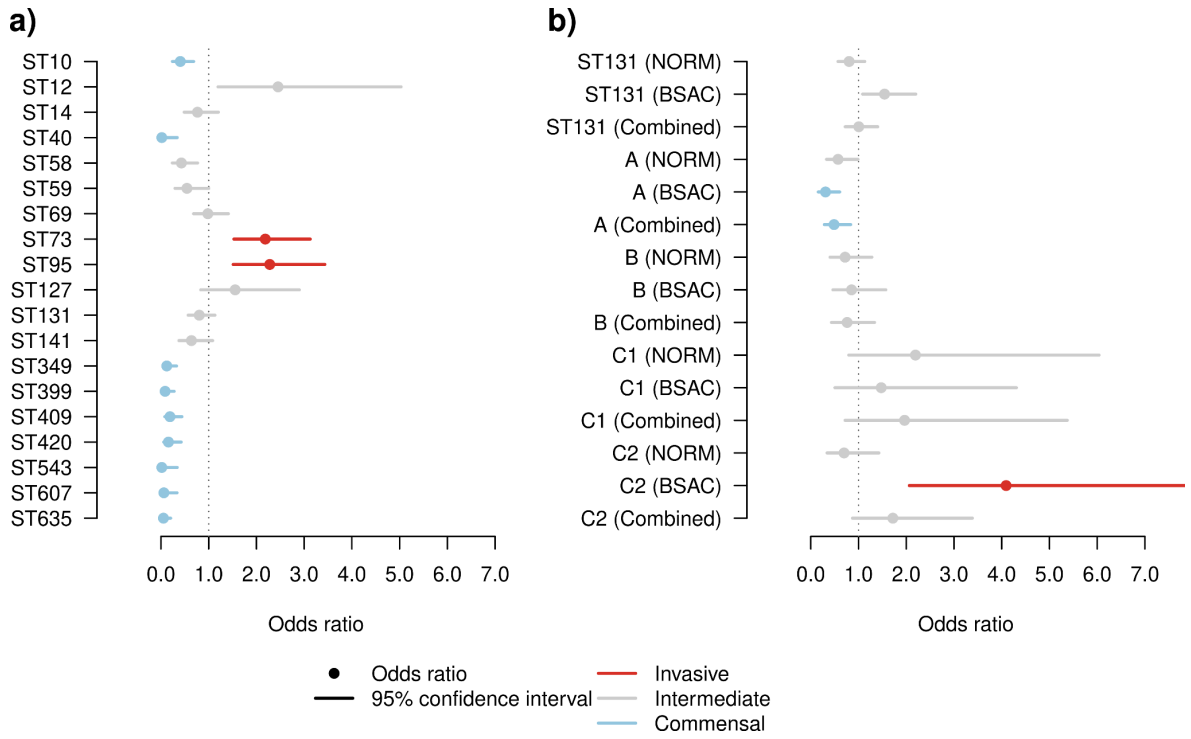

Figure 5 Odds ratios for relative *E. coli* invasiveness. The odds ratios for invasiveness are displayed with the 95% confidence interval, where an OR of > 1 corresponds to more invasive and <1 to more commensal ST, for **a)** the top 10 most frequent lineages (ST73 through to ST59) in Norwegian bloodstream infections (Gladstone et al Lancet microbe 2021) and all additional lineages with an OR significantly different from 1. Lineages for which a significant OR was observed after correcting for multiple testing are coloured with a light-blue or red colour. **b)** ORs for the main clades of ST131 using either the BSAC and/or NORM cohorts for the comparison.

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

To further investigate the specific features of the colonisation of the delivery cohorts, we compared them by clustering and phylogenetic analyses to a previously characterised *E. faecalis* species collection [19]. While the lineages identified in the delivery cohorts were otherwise widely dispersed in the species tree and while other major STs were well represented among them, they harboured none of the long-term persistent hospital-adapted lineages ST6 and ST9 and only a few ST28 strains (Figure 6) [19]. Furthermore, no significant differences ($p > 0.05$, Chi-square test) were found between the delivery cohorts with respect to identified strains harbouring antibiotic

resistance-conferring genes to major antibiotic classes (Supplementary Table 2).

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 a transition matrix for the lineages that were observed at least twice across the sets of samples (Figure 7). 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 7, Supplementary Figure 5).

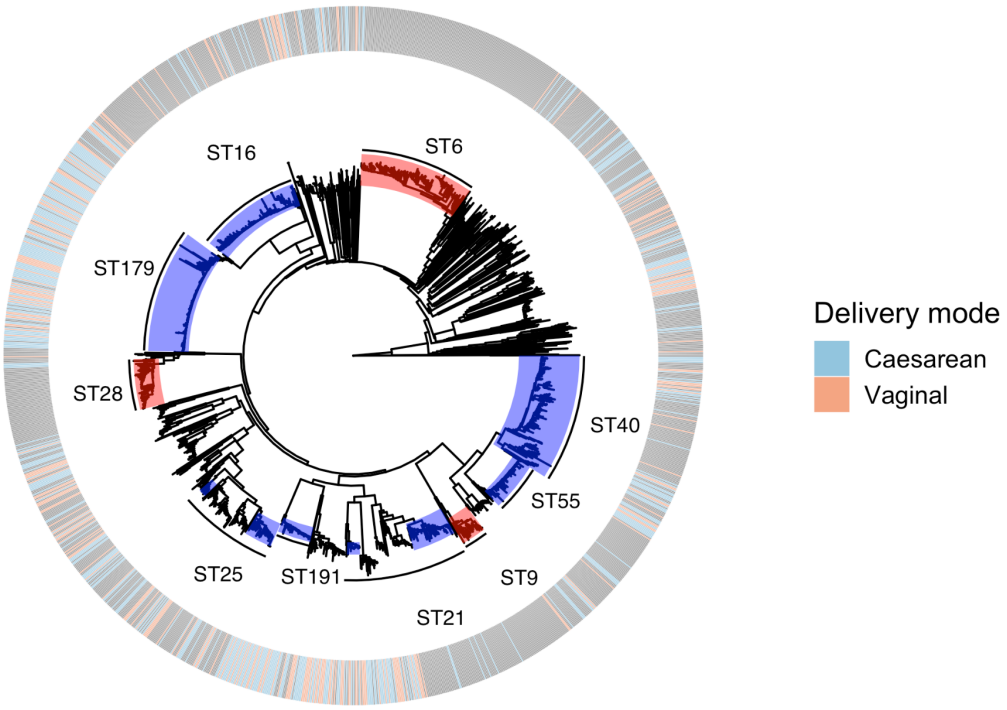

Figure 6 Embedding of the neonatal cohorts within a general *E. faecalis* species-wide collection. Outer metadata blocks depict the delivery mode of the cohorts (caesarean, light blue; vaginal, orange), aligned against the neighbour-joining (NJ) phylogeny from the core distances defined by PopPUNK [20]. Ten largest sequence types in the combined collections are highlighted within the branches, and the sequence types previously defined as hospital-adapted [19] are coloured in red.

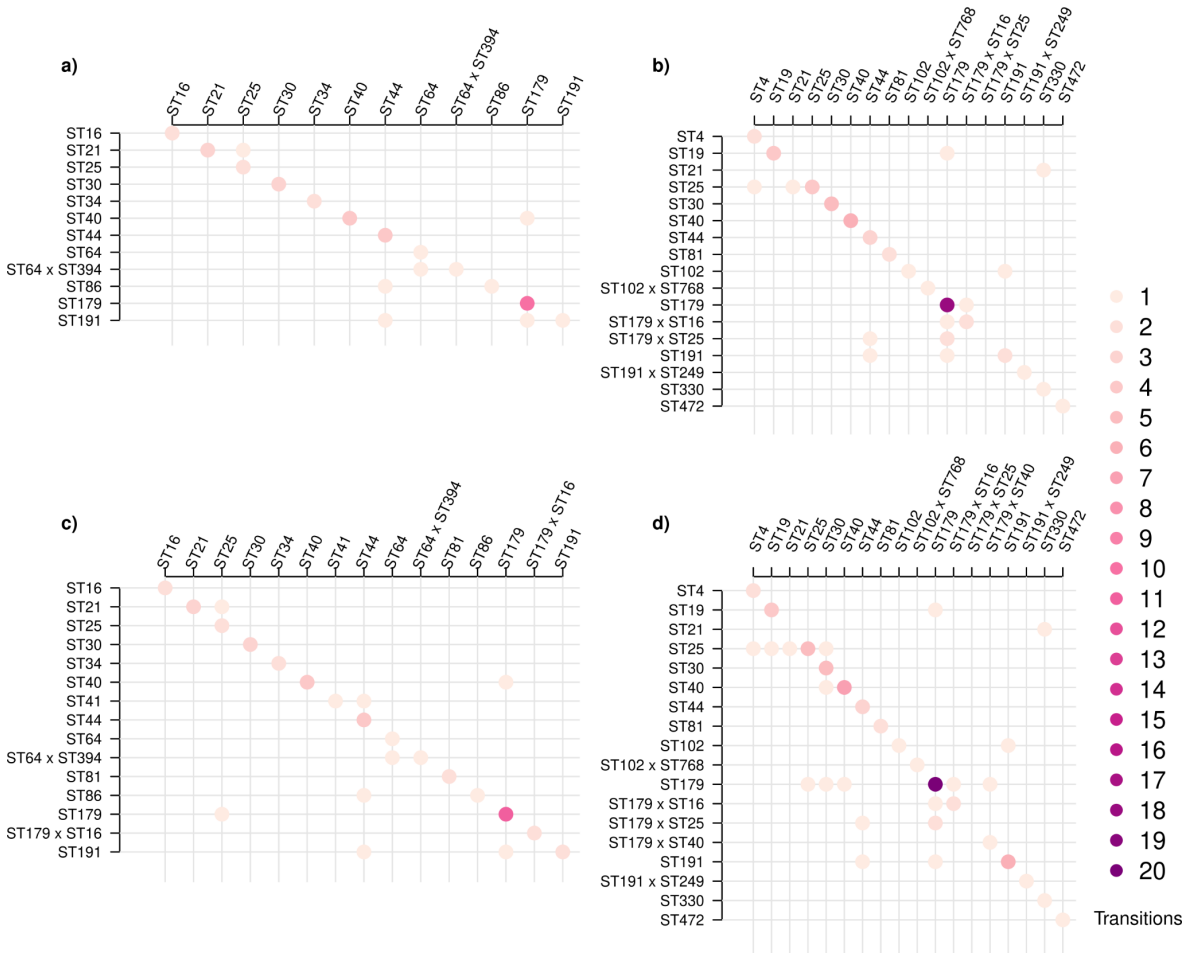

Figure 7 Event matrix displaying colonisation identities with respect to *E. faecalis* lineages between subsequent time points. The figure shows events corresponding to either transition from one *E. faecalis* lineage (rows) to another *E. faecalis* lineage (columns) or persistence of the same lineage (diagonal). Panel **a)** shows events for the vaginal delivery cohort with samples from the infancy period excluded, panel **b)** shows the caesarean section delivery cohort with infancy period excluded, panel **c)** the vaginal delivery cohort with the infancy period included, and panel **d)** the caesarean section delivery cohort with the infancy period included. Darker shades of purple denote more common events. Lineages shown were visited at least twice across the whole set of samples.

AMR and virulence genes in *Klebsiella* strains

We used Kleborate [21] to detect AMR and virulence genes in the 449 *Klebsiella* assemblies (Figure 8). We detected no *mcr* or carbapenemase genes and only 5/186 (3%) of the *K. pneumoniae* assemblies harboured a CTX-M-15 ESBL gene. These four isolates came from four individuals (two caesarean and

two vaginal delivery cases). The gene was also detected twice in one caesarean-born individual. The isolates harbouring the gene were from two different sequence clusters (SC1 containing sequences from ST336 and ST17, and SC24 containing ST323). Both sequence clusters were found in each delivery cohort. Out of these, at least ST323 has been previously associated with ESBL carriage and nosocomial transmission [22].

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 8). There were three isolates of the hypervirulent *K. pneumoniae* ST23 clone from two individuals (both vaginally-born). These isolates harboured aerobactin (iuc), colibactin (cbt), salmochelin (iro), and RmpADC; a suite of virulence factors that are often carried together on virulence plasmids. We also observed 32 *K. pneumoniae* isolates 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. Therefore we consider it likely that these are more divergent loci that are required for survival in these species, rather than true virulence factors, and we would advise caution with the interpretation of these results.

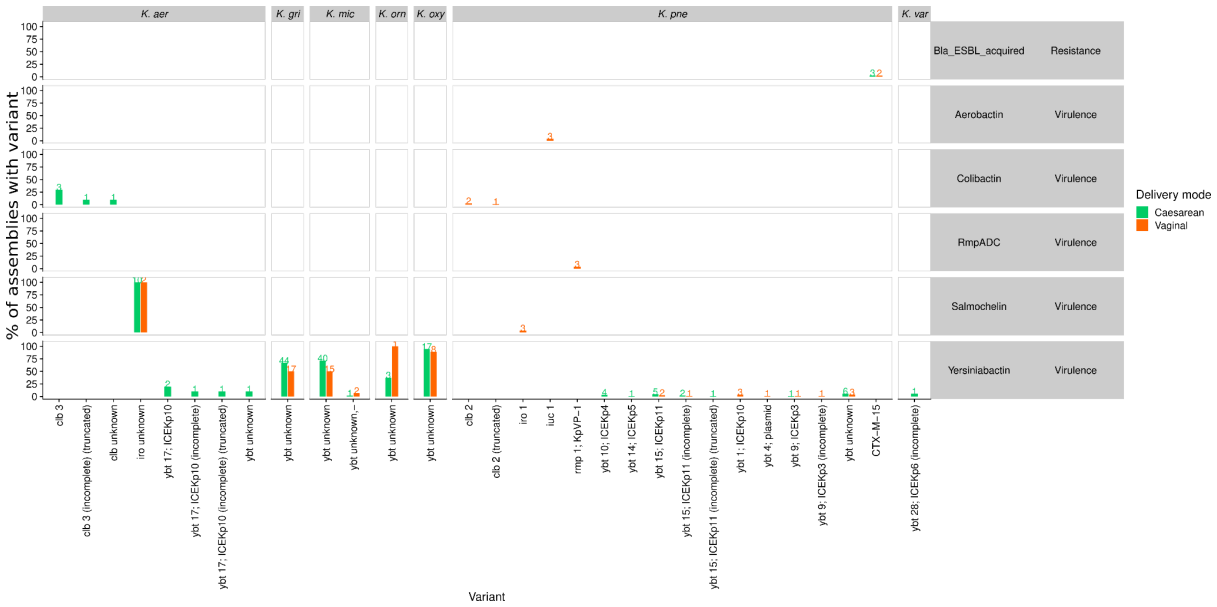

Figure 8 Summary of AMR and virulence genes in the *Klebsiella* assemblies. The plot shows frequencies of the single AMR factor and the four virulence factors identified in the *Klebsiella* species using Kleborate [21]. Sequence assemblies from the caesarean section delivered cohort are highlighted in 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 to 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 novel 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 for standard bacterial genomic epidemiology.

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 identified strains of common pathogenic gut bacteria. 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 the hospital adapted lineages are generally at a disadvantage when attempting to colonise healthy individuals in the hospital environment in the absence of strong selective pressure stemming from the use of antibiotics. 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 [23].

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 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 (Fig 1). 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 [24]–[26] dependent on geographical location [27]. A recent large cross-sectional study also found 16% carriage rate of *K. pneumoniae* in a general adult cohort based on selective culturing approach [28]. In our study, the high frequency and diversity of *Klebsiella* detected in the neonates implies that these species must 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.

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 lineages in the adult gut, since contacts with other humans are the most relevant source of transmission of the bacteria to the babies. Previous attempts in the literature to estimate the level of relative invasiveness of particular strains in this respect appear to be scarce [12] and here we used for comparison both a UK [18] and a Norwegian [17] 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 ST131 relative invasiveness was not particularly high, but its frequency in disease was overall reflective of that in colonisation (Figure 5). 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 [17], 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 colonisation data (Figure 5). Additionally, ST12 showed an elevated OR but this was not statistically significant when corrected for multiple testing.

A recent study investigated the polymicrobial nature of recurrent urinary tract infections (rUTIs) and found that *E. coli* gut and bladder populations were comparable between women with and without a history of rUTI in both relative abundance and phylogroup [29]. However, a deeper lineage level analysis, such as the one performed here, combined with screening of genetic determinants of persistent bladder colonisation from the de-mixed metagenome assemblies could provide a more refined characterisation of the possible population differences between gut and the bladder. This would nevertheless require substantially increased sequencing depths to enable high-resolution genomic epidemiology. Another recent study further demonstrated the added benefits of deep sequencing powered analysis of heterogeneous colonisation by performing deep sequencing of within-host diversity in pneumococcal carriage [30]. Using mGEMS [16] on sequenced DNA from plate sweeps allowed them to investigate implications in disease, the evolutionary responses to antibiotic treatment, identify lineages present in samples, and to perform resistance and other variant calling [30]. 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/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).

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 (Supplementary Table 1) 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 [31], and ~3300 assemblies from a Norwegian bloodstream infection collection [17]. For *E. faecalis*, we used ~2000 assemblies from several European countries [19], and for *Klebsiella* ~3000 assemblies collected in Italy [32]. The isolate assemblies from the source study [14] belonging to *E. coli*, *E. faecalis*, or *Klebsiella* were also included in the respective collections.

After collecting the initial reference database, we ran MetaPhlAn (v3.0, [33], default options) on the WGS metagenomics reads to identify species in the samples that had no representation in the already collected 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 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 indexed the produced multifasta file with Themisto (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 group (Supplementary Table 1), we also built individual species-level level indexes with Themisto (default options) incorporating only the reference sequences of that particular species and no colours.

Reference sequence grouping

Within each species-specific reference database, we used PopPUNK [20] to assign the sequences to groups that roughly correspond to clonal complexes. This was done by running the method 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). Out of these, we chose the result that had both high quality metrics as reported by PopPUNK and was relatively consistent with the species' MLST designations in a manual inspection.

Lineage identification

We used a hierarchical approach consisting of a species detection step and a lineage analysis step. In the species detection step, reads were first aligned with Themisto (v2.1.0, reverse complement handling and output sorting options enabled) against the colored index and species abundances estimated from the alignments with mSWEEP (v1.6.0, [15], [34], write-probs option enabled). The output from mSWEEP was combined with the input reads and processed with mGEMS (v1.2.0, [16], [35], default options), creating separate bins for each species detected in the sample with an abundance of at least 0.000001.

In the lineage analysis step, reads from each species-level bin belonging to the priority pathogens group were aligned against their corresponding

species-level indexes and processed with mSWEEP and mGEMS in the same way as in the species detection step. We then ran demix_check (commit 18470d3, [36]) on the resulting bins to filter out cases where the bins contained reads that did not match any reference lineage in our reference sequences.

Assembly

Read files were quality controlled and corrected with fastp (v0.23.1, [37], default settings) and the corrected reads assembled with shovill (v1.1.0, [38], with read correction disabled). Both the isolate and the binned reads were assembled with the same approach.

Correlation estimation

We used FastSpar (v1.0.0, [39], [40] 1000 iterations with 200 exclusion iterations) to infer the correlations between operational taxonomic units constructed by multiplying the relative abundance of a reference taxonomic unit from mSWEEP with the total number of pseudoaligned reads in the sample. Statistical significance of the correlation values was calculated using the permutation test functionality from FastSpar (10 000 permutations with 5 FastSpar iterations per permutation).

Visualisations

Figures 1-7 were created using R v4.0.5 [41]. The scripts used to create the visualisations are available from GitHub at https://github.com/tmaklin/baby-microbiome-paper-plots. The UpSet plot [42] was created using the UpSetR package (v1.4.0, [43]).

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 Norwegian bloodstream infection (BSI, 2002-2017, [17]), the UK BSAC BSI collection (2001-2012,

[18]), and the colonisation collections presented in this paper's source study (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. For tables with any zero values, 0.5 was added to all cells in the table before calculating the odds ratios [44]. Statistical significance 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 [45]. In the ST131 sublineage analysis, SKA [46] was used to generate an alignment and a subsequent tree embedding the ST131 carriage isolates from this study and the NORM collection ST131 [17] into the BSAC collection [18], from which clade membership could be inferred.

Comparative analysis of *E. faecalis* population structures

We used the query option from PopPUNK v.2.2.0 [20] to embed another *E. faecalis* collection [19] and the de-mixed metagenomics assemblies from the neonatal cohort into the alignment-free clustering of the *E. faecalis* sequences. This allowed us to compare the differences between these three 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 [19] from the core distances defined by PopPUNK [20]. Multi-locus sequence types were retrieved from assemblies using fastMLST v.0.0.15 [47], and antibiotic resistance profiles screened from assemblies using AMRFinderPlus v.3.10.18 [48] 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 [41]. Vancomycin (glycopeptide) was omitted as none of the isolates in the neonate source study [14] harboured *van* genes.

Detection of AMR and virulence genes in *Klebsiella* strains

We used Kleborate v2.1.0 [21] to detect AMR and virulence genes in the *Klebsiella* assemblies.

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 3.

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] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] G. Tonkin-Hill *et al.*, "Pneumococcal within-host diversity during colonisation, transmission and treatment," Genomics, preprint, Feb. 2022. doi: 10.1101/2022.02.20.480002.

[31] 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.

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

[33] 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.

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

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

[36] H. Thorpe, *harry-thorpe/demix_check: demix_check 18470d3 (25 October 2021)*. GitHub, 2021. [Online]. Available: https://github.com/harry-thorpe/demix_check

[37] 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.

[38] T. Seemann, *tseemann/Shovill: Shovill-v1.1.0 (13 March 2020)*. GitHub, 2020. Accessed: Nov. 18, 2021. [Online]. Available: https://github.com/tseemann/shovill

[39] 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.

[40] 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.

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

[42] 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.

[43] 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.

[44] 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.

[45] 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.

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

[47] 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.

[48] 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.