Strong pathogen competition in neonatal gut colonization

Tommi Mäklin[1,*], ..., Harry, Anna, Rebecca, Yan, Alan, Pål, Ørjan, Trevor, Antti Honkela [1], Jukka Corander[2,3,4,*]

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 Wellcome Sanger Institute, Hinxton, Cambridgeshire, UK

4 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 be competing against each other and other commensal species residing within this ecosystem. However, currently we are lacking a population-wide quantification of strain-level colonisation potential for the common bacterial pathogens and the relationship of this to their prevalence in disease. In addition, it is unclear how ecological factors might be modulating these competition dynamics. Using a combination of latest high-resolution metagenomics and strain-level genomic epidemiology methods, we performed such a quantification for a longitudinal cohort of neonatal gut microbiome samples. This analysis demonstrates generally a strong inter- and intra-species competition dynamics in the gut colonisation process, but also reveals a number of synergistic relationships among several species belonging to genus *Klebsiella*, including the prominent human pathogen *Klebsiella pneumoniae*. Our findings provide 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. The earlier established common ExPEC clones ST73 and ST95 are shown to be significantly more pathogenic than the more recent, globally circulating multi-drug resistant clone ST131. 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.

Keywords

Introduction

Human gut bacteria are generally considered to be commensals but some of them harbour considerable potential to cause either mild or severe infections outside the gut, one of the most prominent examples being 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, and multi-locus enzyme electrophoresis (MLEE) typing studies done on longitudinal collections of stool confirmed these conclusions in the early 1980s [3], [4].

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 and our current study aimed to address the need to assess these aspects for several of the major human pathogens residing within the gut microbiome. There are many interesting findings from experimental studies of inter-species competition in colonisation based on animal models, for example a recent demonstration that *K. michiganensis* prevents mouse gut colonisation by *E. coli* in a particular ecological setting [5]. As an example of intra-species competition, a longitudinal study was tracking 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*. 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 [6]. 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 [7].

Characterising the diversity of colonising pathogenic bacteria 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 [8].

TBC...

Results

Strain-level analysis of neonatal gut microbiome data

We analysed 1679 sets of short reads from gut metagenome samples that had been sequenced as part of a previously published study on opportunistic pathogen colonization in newborn babies [9]. Our study extends the previous analysis by providing strain-level characterization of these samples for several important pathogenic species (Table Abbreviations) as well as a more detailed exploration of the diversity within the *Klebsiella* genus. In both the strain-level characterization and the *Klebsiella* species analysis we applied the recent mSWEEP and mGEMS methods [10], [11] to a bespoke set of reference sequences (Methods section). The analysis pipeline is described in more detail in Supplementary Figure NiceFlowChart and in the Methods section. Results from the analyses will be presented in the following sections.

Abbreviation	Full name
E. col	*Escherichia coli*
E. fcm	*Enterococcus faecium*
E. fcs	*Enterococcus faecalis*
K. aer	*Klebsiella aerogenes*
K. gri	*Klebsiella grimontii*

K. hua	*Klebsiella huaxiensis*
K. mic	*Klebsiella michiganensis*
K. orn	*Klebsiella ornithinolytica*
K. oxy	*Klebsiella oxytoca*
K. pas	*Klebsiella pasteurii*
K. pla	*Klebsiella planticola*
K. pne	*Klebsiella pneumoniae*
K. qpq	*??*
K. qps	*??*
K. spa	*Klebsiella spallanzanii ?*
K. var	*Klebsiella variicola*
S. aur	*Staphylococcus aureus*
Table Abbreviations Abbreviations for names of the pathogens examined.	

Caesarean-section birth disturbs the gut biome composition

Comparing the overall differences in species distribution between the caesarean section and the vaginal delivery cohorts estimated using mSWEEP confirms the results presented in the original study (cite). *E. coli* is more commonly found in the latter cohort (Figure Cohort Differences, Supplementary Figure E Col Strain Abundances), while the *Klebsiella* species, *E. faecalis* and *E. faecium* are more common in the former (Figure Cohort Differences, Supplementary Figures E. fcs/E.fcm/K. x).

Supplementary figures: strain abundances plots (lots of them in results/strain_abundances).

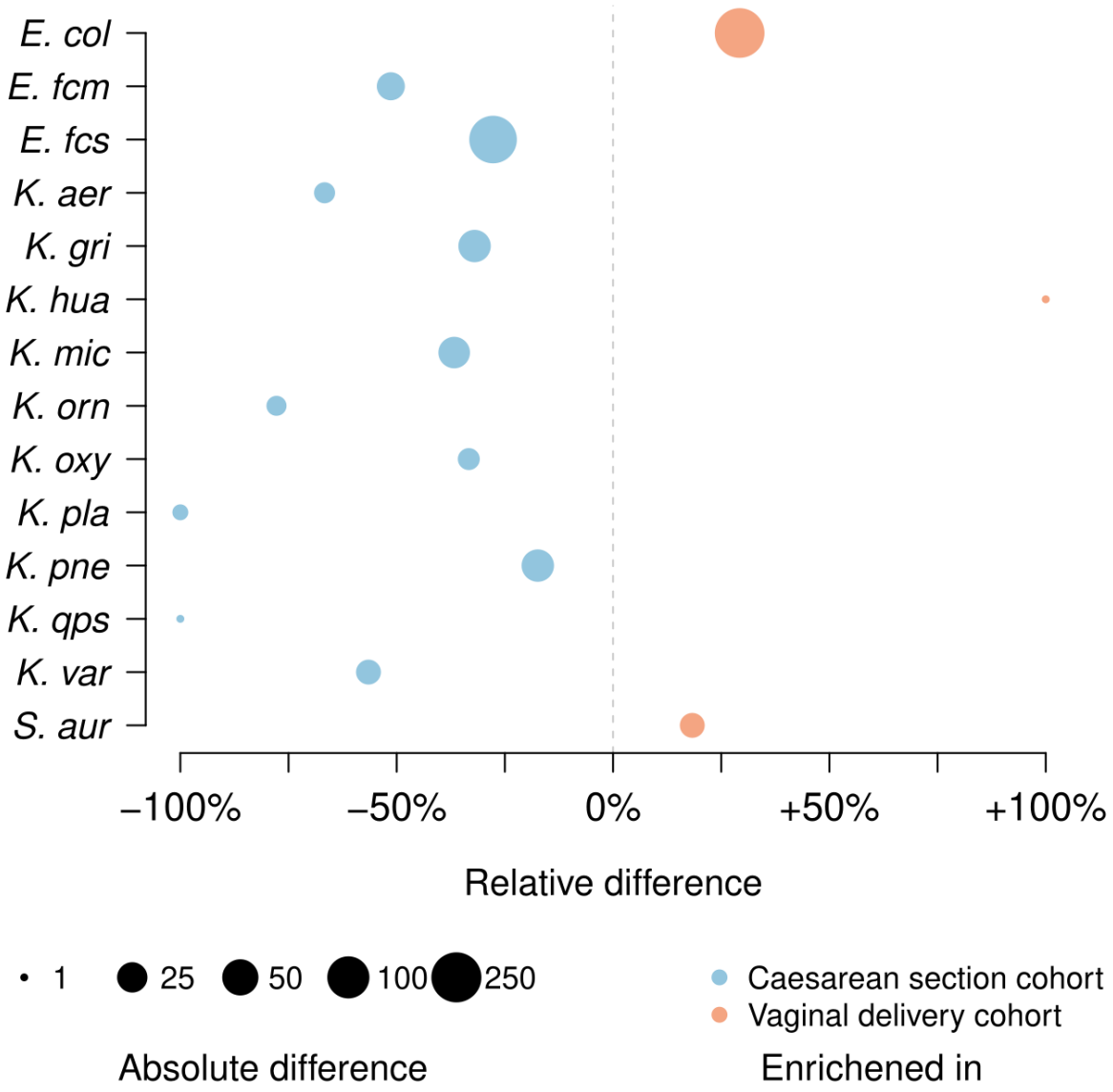


Figure Cohort differences Differences in pathogen loads between cohorts. The figure shows differences in the number of reliably identified pathogens in each cohort. Area of the circles displays the absolute difference between the cohorts, while the horizontal placement of the circles displays the relative 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 light blue.

Competition and synergistic relationships drive Enterobacteriaceae colonization

We first looked at the inter-species competition dynamics between various Enterobacteriaceae species and other major pathogen species (*Enterococcus*

faecalis, Enterococcus faecium, and *Staphylococccus aureus*). Chiefly, we identified statistically significant ($p < 0.05$, permutation test) antagonistic relationships between *E. coli* and *K. grimontii, K. michiganensis,* and *K. pneumoniae* (Figure Correlations a). The existence of this relationship is also suggested by the absence of colonization by *Klebsiella* species in the vaginally delivered cohort (Figure XX) and has been previously verified for the *E. coli* - *K. michiganensis* pair in a mouse gut model [5].

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

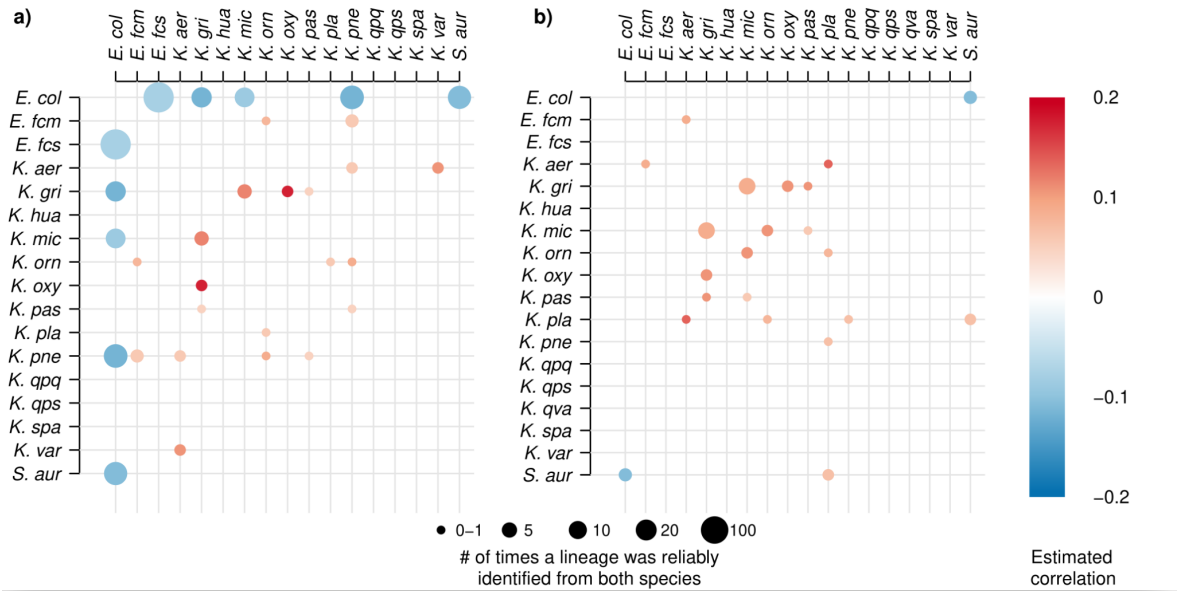


Figure Correlations Correlations between the identified priority pathogens. The figure shows statistically significant ($p < 0.05$, permutation test) positive and negative correlations for our species of interest. Panel **a)** shows the correlations within the vaginally delivered cohort, and panel **b)**

within the caesarean section delivered cohort. Darker shades of red and blue represent higher positive/negative correlation, while larger areas of the circles stand for a higher number of samples where the correlated pair was reliably identified.

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 analyzing 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 UpSet plot). 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 taxon 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 Correlation Plot).

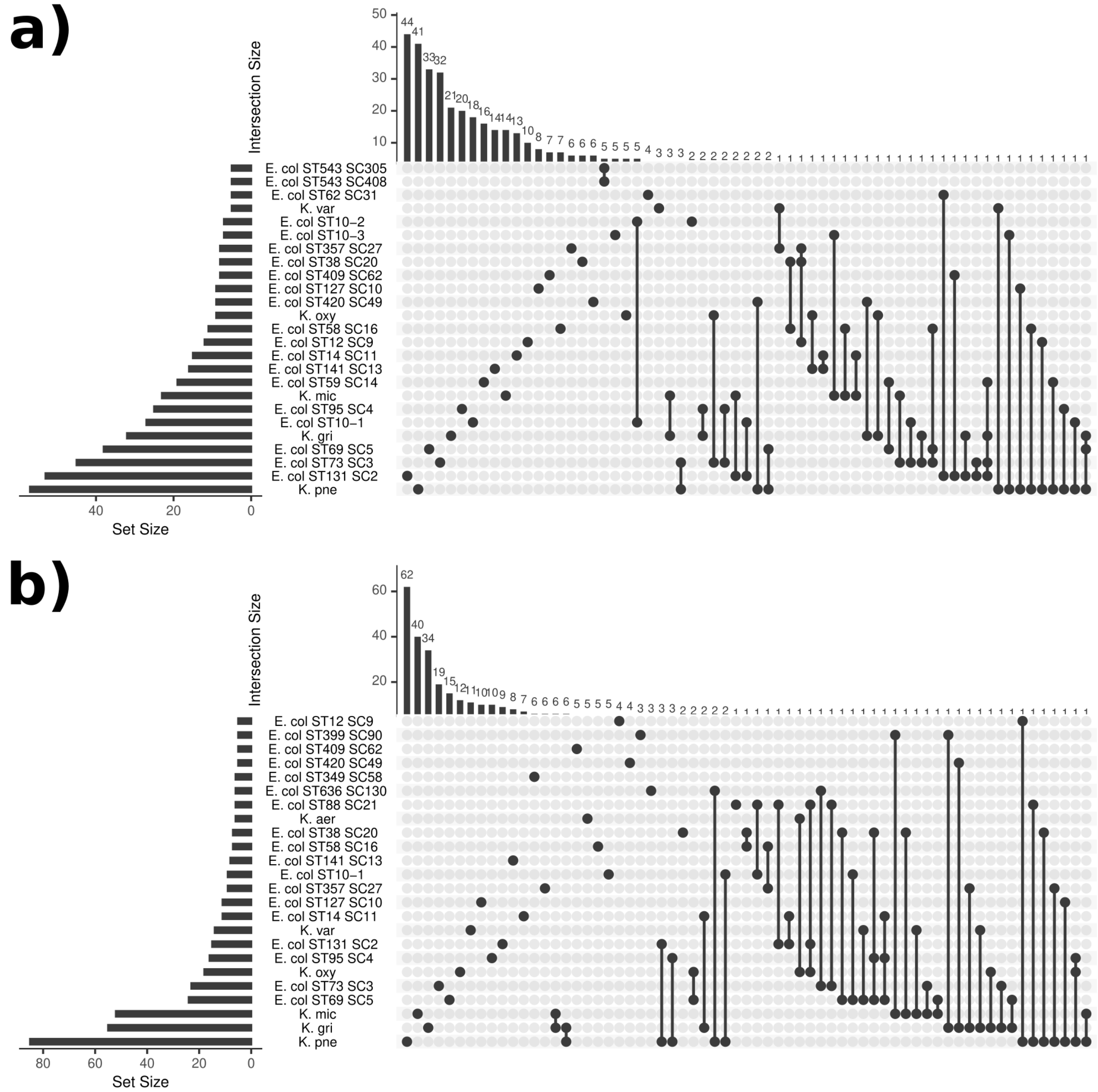


Figure UpSet plot UpSet plot showing coexistence of *E. coli* lineages with *Klebsiella* species. The plot displays the number of times the *E. coli* lineages and *Klebsiella* taxonomic units were found either alone (single dots) or together in a sample (connected dots) at least five times in the plotted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

Neonatal gut colonization adheres to the first-come, first-served principle

We then looked in more detail into at what time point the *E. coli* lineages appear to colonize the neonatal gut, to what degree they are inherited from the mothers, and whether the lineages that are successful within the first 21 days persist in the infancy period sampling 3-6 months later. Examining the time courses for each individual, we found the colonization to already haven taken place at the very first sampling time at 4 days in the vaginally delivered cohort and subsequently nearly all of the infants colonized at the next sampling time at 7 days (Figure E_col_colonization panel a). Conversely, in the caesarean section delivered cohort there were markedly fewer *E. coli* found overall in the early time points, with some signs of the initial colonization happening slightly later at 7 days (Figure E_col_colonization panel b).

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 XX cases where the same lineage was shared between the mother and the newborn, indicating potential transmission. However, most of the initial colonizations appear to have been obtained from the environment rather than transmitted from the mother. Additionally, the lineages identified in the infancy period showed …

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 colonized the gut at the 4 or 7 days time point either persisted into the final 21 days sampling point or vanished completely. Transitions to another lineage within the first 21 days of life were relatively uncommon and primarily occurred between the final time point and the infancy period sampling, where a longer period of time had passed.

Finally, looking at the number of samples which were observed to contain several *E. coli* lineages shows that co-occurrence seldom occurs within the first 21 days. In the infancy period the percentage of co-colonized individuals increases with the vaginally delivered cohort exhibiting slightly more variety. Similar rates of co-colonization are seen in the established microbiomes of the mothers. Taken together with the findings from the individual time courses, these results indicate a substantial competitive advantage for the

first strain to colonize the gut which dissipates only after several months have passed.

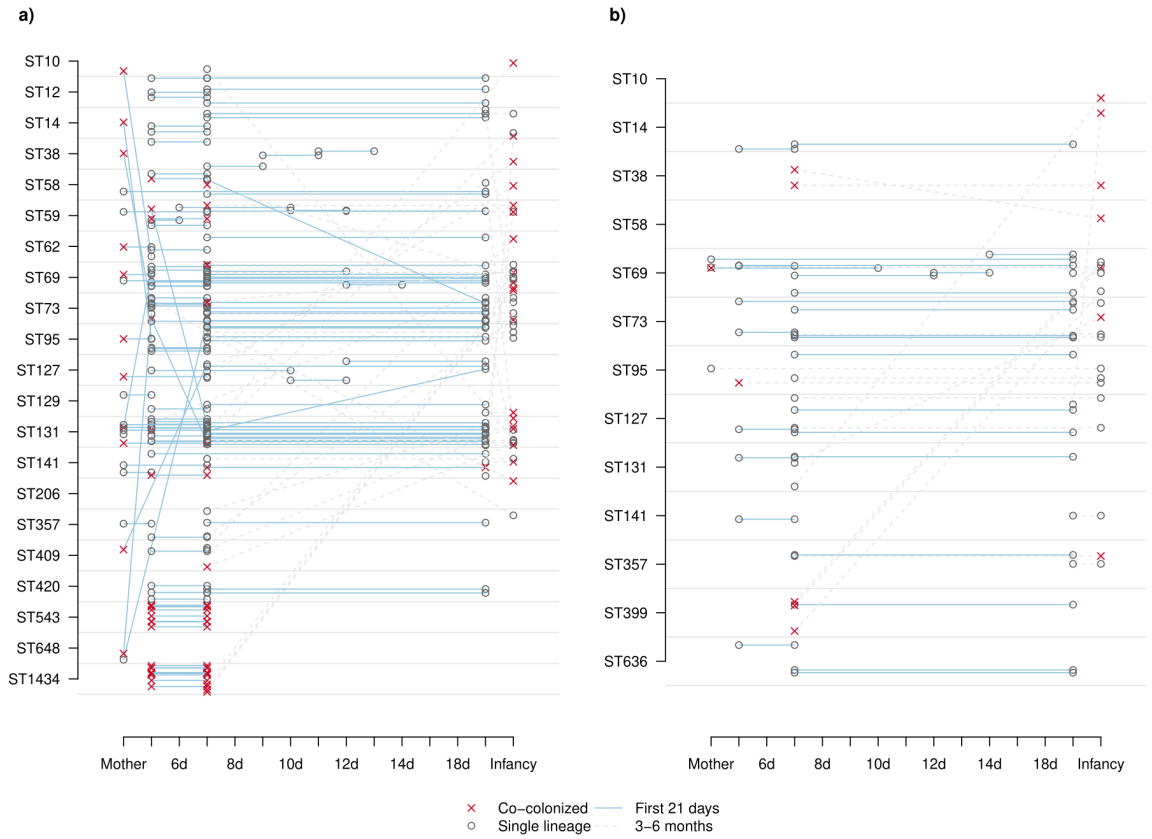


Figure *E_col* colonization Longitudinal chart showing *E. coli* colonization over time. The plot shows positive identifications of *E. coli* sequence types (rows) in a sample taken at a certain time point (columns). anel **a)** displays the data for the vaginally delivered cohort, and panel **b)** for the caesarean section delivered cohort. Empty circles are reliable identifications of a single sequence type in the sample and red crosses identifications of coexisting sequence types. Connected solid or dashed lines represent the samples taken from a single individual (time points labelled with the number or days or ‘Infancy’) or their mother. A solid light blue line denotes the samples taken within the first 21 days of life and a dashed grey line a follow-up sampling 3-6 months after. Horizontal lines signify identification of the lineage at several time points and angled lines signify a switch from one lineage to another. Only lineages which were identified at least five times are shown.

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

We further examined the dynamics of transitioning from carriage of one *E. coli* lineage to carriage of another by constructing a transition matrix for the lineages that were observed at least twice across the sets of samples. Looking at the samples from the first 21 days only (Figure E_col_transitions panels a and b) shows a strong preference for persistence of the first strain to colonize the gut, with 80% of the transitions happening on the diagonal (persistence of the same strain between two subsequent time points). Including the infancy period (Figure E_col_transitions panels c and d) contains slightly more variety especially in the vaginally delivered cohort (Figure E_col_transitions panel c) but nevertheless 70% of the transitions remain on the diagonal. The full transition matrix including all observed transitions is available in Supplementary Figure VeryLargeMatrix.

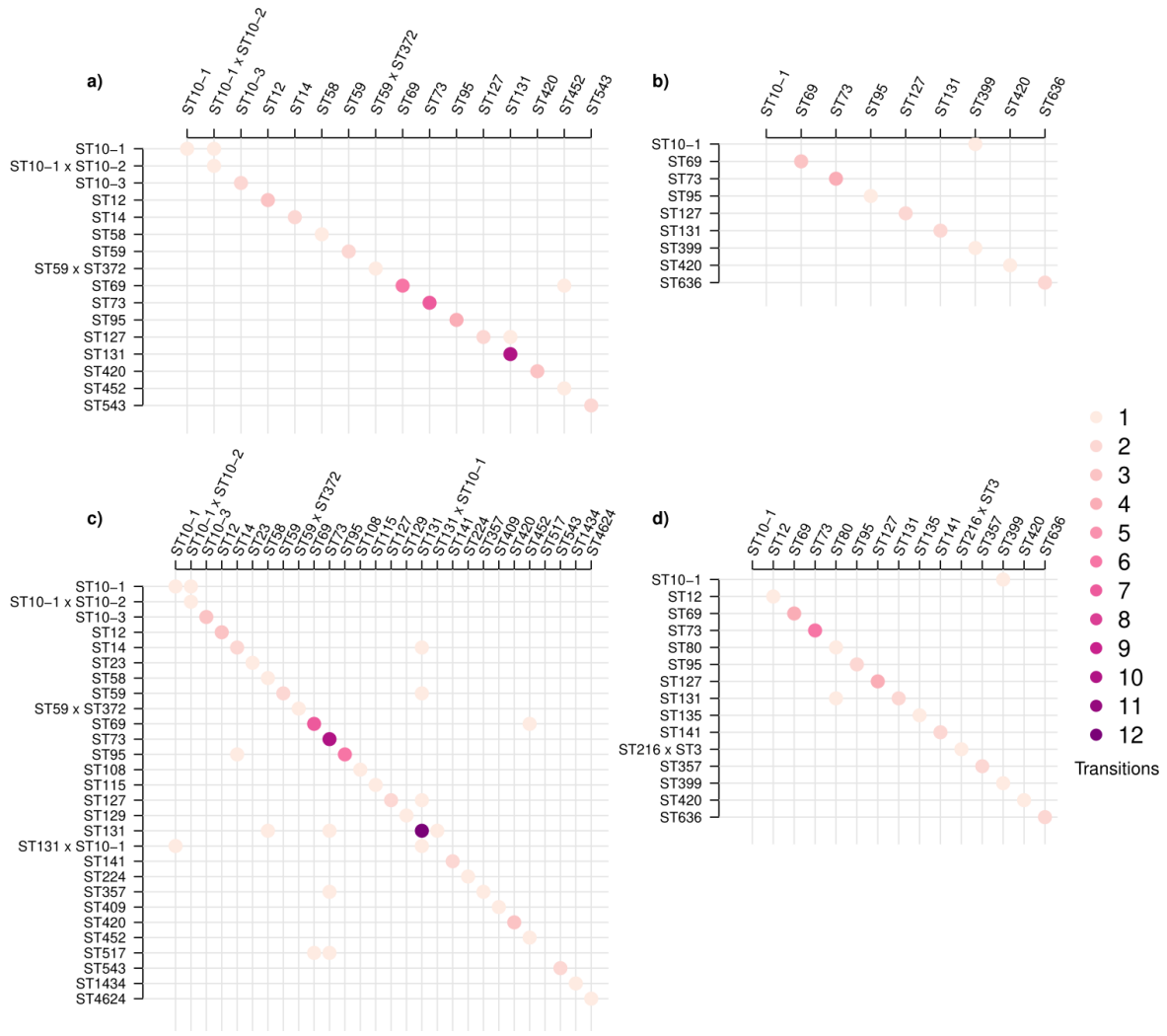
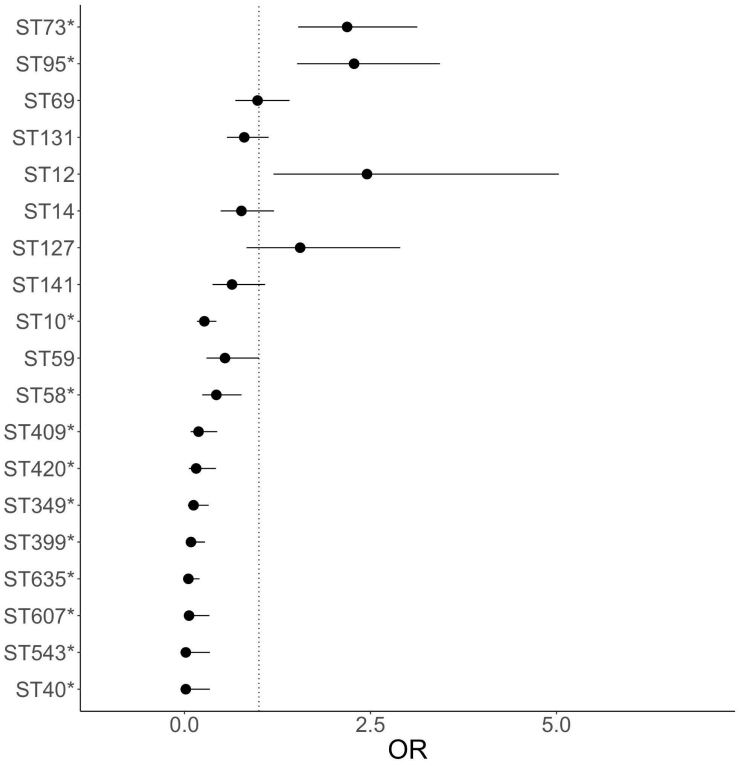


Figure E_col transitions Transition matrix showing switches from one *E. coli* lineage to another. The figure shows transitions from one *E. coli lineage* (rows) to another *E. coli* lineage (columns) or persistence of the same lineage (diagonal). Panel **a)** shows transitions for the vaginally delivered cohort with samples from the infancy period excluded, panel **b)** shows the caesarean section delivered cohort with infancy period excluded, panel **c)** the vaginally delivered cohort with the infancy period included, and panel **d)** the caesarean section delivered cohort with the infancy period included. Darker shades of purple denote more common transitions. Lineages shown were visited at least twice with the rest hidden.

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

We determined the relative invasiveness of *E. coli* lineages using odds ratios for the frequency of each lineage in carriage compared to two disease collections (Supplementary table). The interpretation of the ORs (>1 invasive, <1 commensal) were not influenced by the disease collection with the exception of ST131. For the BSAC data, that sampled years during which the prevalence of ST131 changed markedly, ST131's invasiveness is overestimated (Supplementary figure). We observed that the prevalence in carriage and the NORM disease collection was similar for ST69 and ST131 unlike ST73 and ST95 which were overrepresented in disease isolates (Figure XA). Numerous other lineages were observed to have significant ORs predominantly 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 XB). C1 and C2 ORs were more intermediate with wider confidence intervals. C2 estimates were particularly affected by which disease collection was used at a comparitor due to known differences in prevalence over time and between the UK and Norway.


ST73*
ST95*
ST69
ST131
ST12
ST14
ST127
ST141
ST10*
ST59
ST58*
ST409*
ST420*
ST349*
ST399*
ST635*
ST607*
ST543*
ST40*
0.0
2.5
5.0
OR

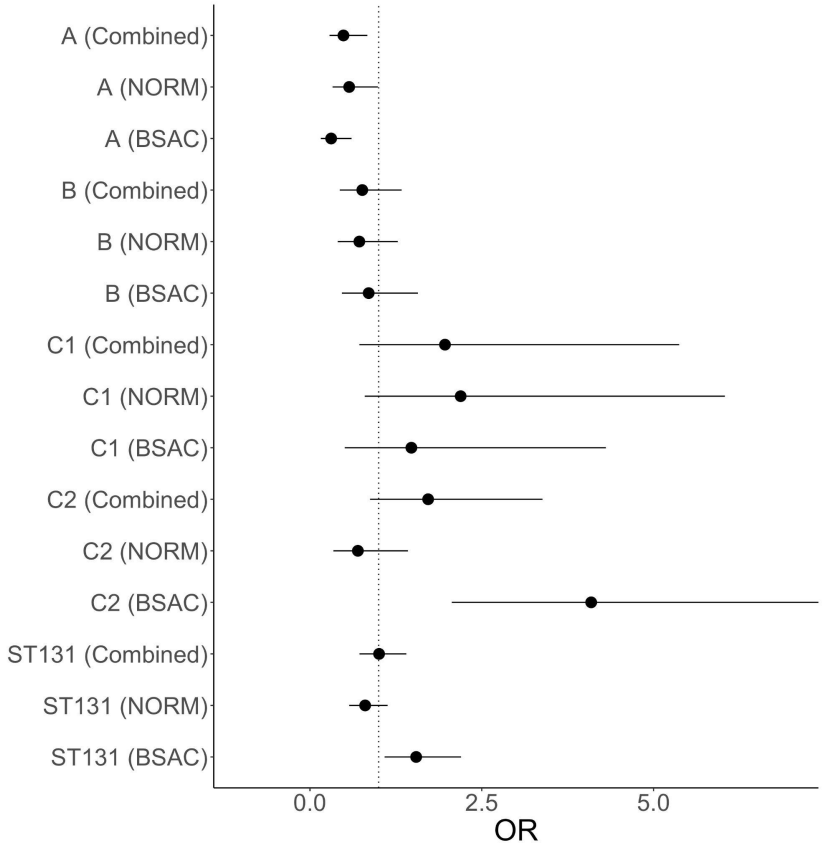

A (Combined)
A (NORM)
A (BSAC)
B (Combined)
B (NORM)
B (BSAC)
C1 (Combined)
C1 (NORM)
C1 (BSAC)
C2 (Combined)
C2 (NORM)
C2 (BSAC)
ST131 (Combined)
ST131 (NORM)
ST131 (BSAC)
0.0
2.5
5.0
OR

Figure *E_col* Odd Ratios Showing relative *E. coli* invasiveness. The figure shows the odds ratios for invasiveness and the 95% confidence interval, where a OR of > than 1 is invasive and <1 commensal, for **a)** the top 10 lineages (ST73 through to ST59) in Norwegian bloodstream infections (Gladstone et al Lancet microbe 2021) and all additional lineages with an OR significantly different to 1. Lineages for which a significant OR was observed after correcting for multiple testing are denoted with * after the lineage name in the y-axis. **b)** the main clades of ST131 using the BSAC and/or NORM collections as the comparator.

ST131 phylogeny?

E. faecalis phylogenetic tree / E. faecium results - are hospital clades more prevalent?

***Klebsiella* species (other than *K. pneumoniae*) compared to established microbiomes**

AMR and virulence profiles of Klebsiella in vaginal vs c-section babies, are there any differences?

Discussion

Invasiveness

We used a Norwegian dataset known to have a similar BSI population structure collected to the UK but with common samping years with the colonisation data presented here and critically after the emergence and stabilisation of ST131, to limit bias in estimating ORs for invasiveness. We subsequently found that a finding reported on biorxiv of ST131 being particularly invasive was not observed [12], only ST95 and ST73 were determined to be invasive. The observed relative invasiveness of ST131 clades were consistent with estimates inferred from the phylogenetic signal of BSI in Gladstone et al 2021, with A and B less invasive than C1 and C2 suggesting invasiveness can be robustly estimated from genomic collections of disease [14]. The variation in OR estimates for C2 demonstrates the importance of

matching location and time when calculating ORs for invasiveness when prevalence is known to vary temporally or geographically.

Methods

Sequencing data

We used sequencing data from a previous study [9] that has been published in the European Nucleotide Archive under accession numbers ERP115334 (whole-genome shotgun metagenomics sequencing data) and ERP024601 (isolate sequencing data).

Species detection

We used MetaPhlAn 3.0 (v3.0.13, [13], default options).

Reference sequences

We combined 805 isolate assemblies from the source study for the sequencing data [9] with a previously constructed reference database containing sequences for priority pathogens (Table Abbreviations) and common commensal and contaminant species from several studies. This reference database was further extended by including extra *E. coli* sequences from a recent surveillance study [14]. After running MetaPhlAn on the WGS reads, we checked the results for species that were identified as being present but had no representation in the reference database. 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. All reference sequences were processed with a script (available from XX0) that concatenated sequences consisting of several contigs by adding a 300bp gap between the contigs and collected the concatenated sequences in a single multifasta file.

We indexed the multifasta file with Themisto (v2.1.0, [11], no-colors option enabled) and used Themisto (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 implies that a pseudoalignment to possibly several reference sequences of a species is

reported simply as a single pseudoalignment to somewhere within the species. For the species in the priority pathogens group (Table Abbreviations), 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

We used a sequence database containing reference sequences for several nosocomial pathogens from a previous study that had been assigned to groups roughly corresponding to clonal complexes with PopPUNK [15] in the same study. This database was augmented with the sequences from the species identified with MetaPhlAn, which were assigned to groups corresponding to their species names.

Strain identification

We used a hierarchical approach consisting of a species detection step and a strain 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, [10], [16], write-probs option enabled). The output from mSWEEP and the input reads were processed with mGEMS (v1.2.0, [11], [17], default options), creating separate bins for each species detected in the sample with an abundance of at least 0.000001.

In the strain analysis step, reads contained in bins 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, [18]) 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, [19], default settings) and the corrected reads assembled with shovill (v1.1.0, [20],

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, [21], [22] 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 XXX-YYY were created using R v4.0.5 [23]. The scripts used to create the visualizations are available from YYYY. The UpSet plot [24] was created using the UpSetR package (v1.4.0, [25]).

Odds ratios for invasiveness

As the colonisation collection sampled mother-infant pairs multiple times that do not represent independent sampling the pooled presence of a lineage in a pair for all time points was used. PopPUNK clusters assigned to a Norwegian Blood Stream Infection (BSI) collection (2002-2017, [14]), a UK BSAC BSI collection (2001-2012, [26]), and the colonisation collection presented in this paper's source study (2014-2017, [9]) allowing the relative frequencies of lineages to be compared between carriage and disease, for lineages with a count of greater than 1. For tables with any zero value 0.5 was added to all cells in the table before calculating the Odds Ratio [27]. For tables with any cell value <5 Fisher's exact test was applied, otherwise the chi-squared test was used. An adjustment for multiple testing was made using the Benjamini-Hochberg method [28]. SKA was used to generate an alignment and subsequent tree of the ST131 carriage isolates and NORM ST131 [14] from which clade membership could be inferred.

Data availability

Data used are available from XXXX.

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. R.G. and A.P. were funded by the AMR grant from Trond Mohn Foundation.

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] 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.

[6] 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.

[7] 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.

[8] 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.
[9] 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.
[10] 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.
[11] T. Mäklin *et al.*, "Bacterial genomic epidemiology with mixed samples," *Microb. Genomics*, vol. 7, no. 11, Nov. 2021, doi: 10.1099/mgen.0.000691.
[12] J. Marin *et al.*, "The population genomics of increased virulence and antibiotic resistance in human commensal *Escherichia coli* over 30 years in France," Genomics, preprint, Jun. 2021. doi: 10.1101/2021.06.24.449745.
[13] 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.
[14] 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.
[15] 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.
[16] T. Mäklin and A. Honkela, *PROBIC/mSWEEP: mSWEEP-v1.6.0 (15 November 2021)*. Zenodo, 2022. doi: 10.5281/ZENODO.6523380.
[17] T. Mäklin, *PROBIC/mGEMS: mGEMS-v1.2.0 (20 November 2021)*. Zenodo, 2021. doi: 10.5281/ZENODO.5715888.
[18] H. Thorpe, *harry-thorpe/demix_check: demix_check 18470d3 (25 October 2021)*. GitHub, 2021. [Online]. Available: https://github.com/harry-thorpe/demix_check
[19] 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.
[20] T. Seemann, *tseemann/Shovill: Shovill-v1.1.0 (13 March 2020)*. GitHub, 2020. Accessed: Nov. 18, 2021. [Online]. Available: https://github.com/tseemann/shovill
[21] 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.

[22] 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.

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

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.