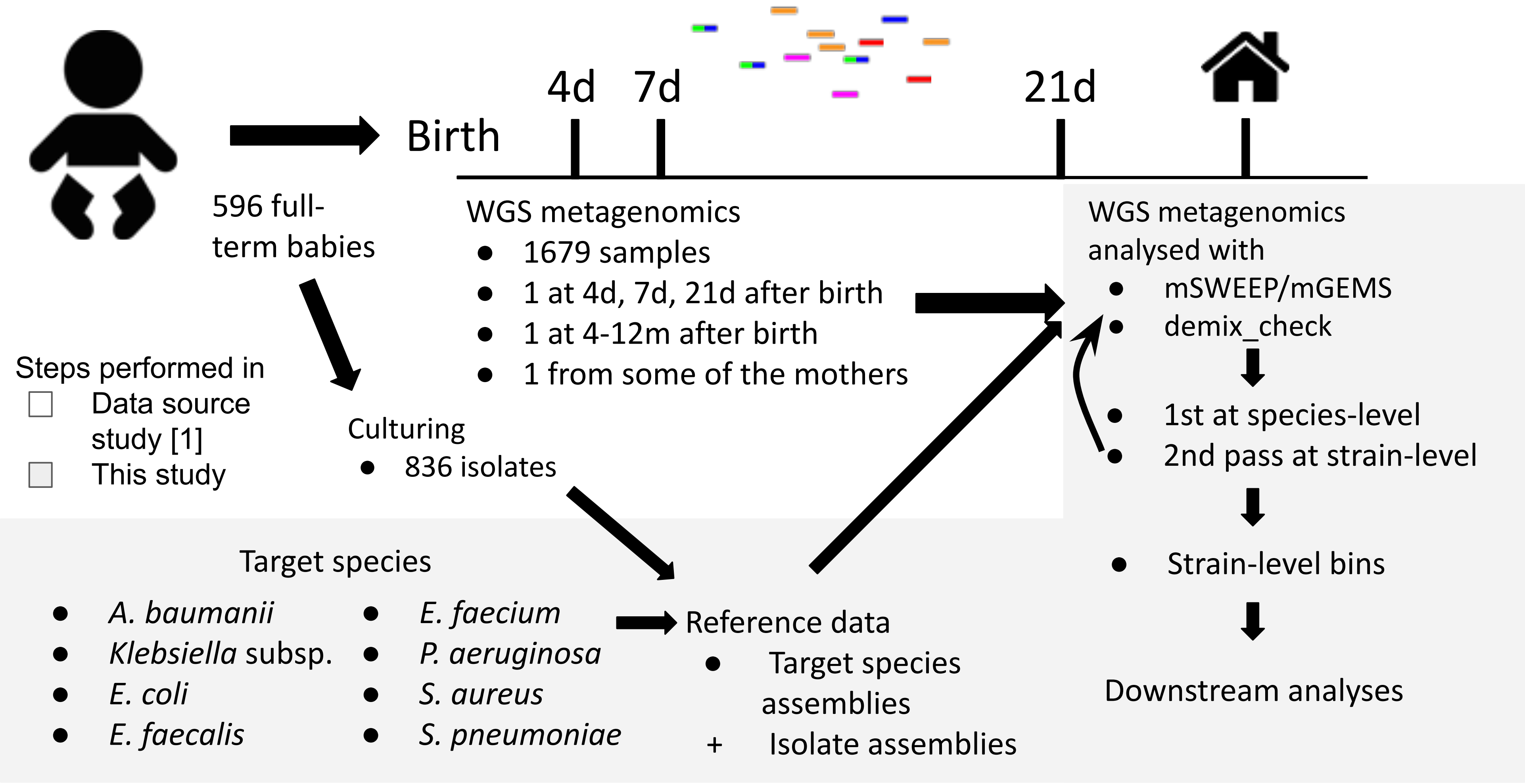
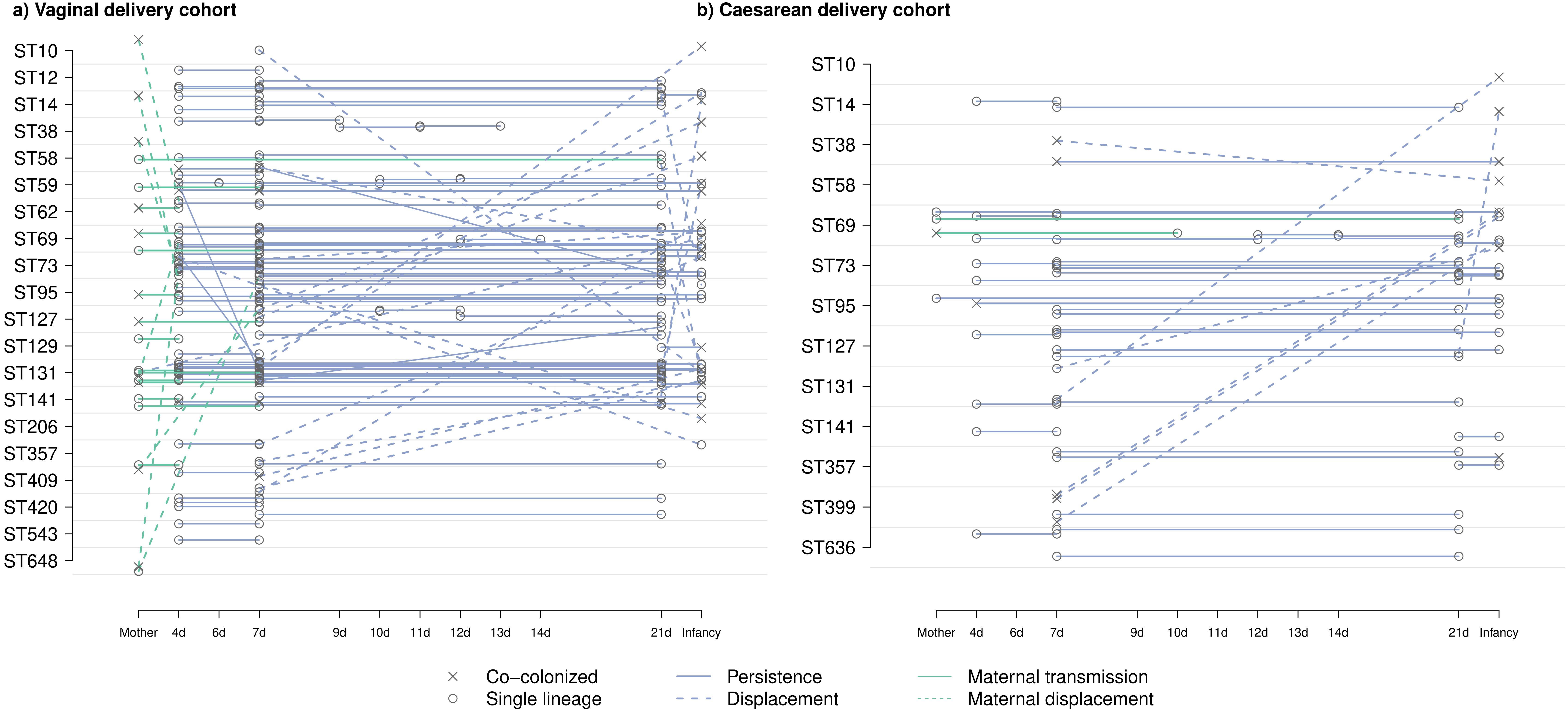
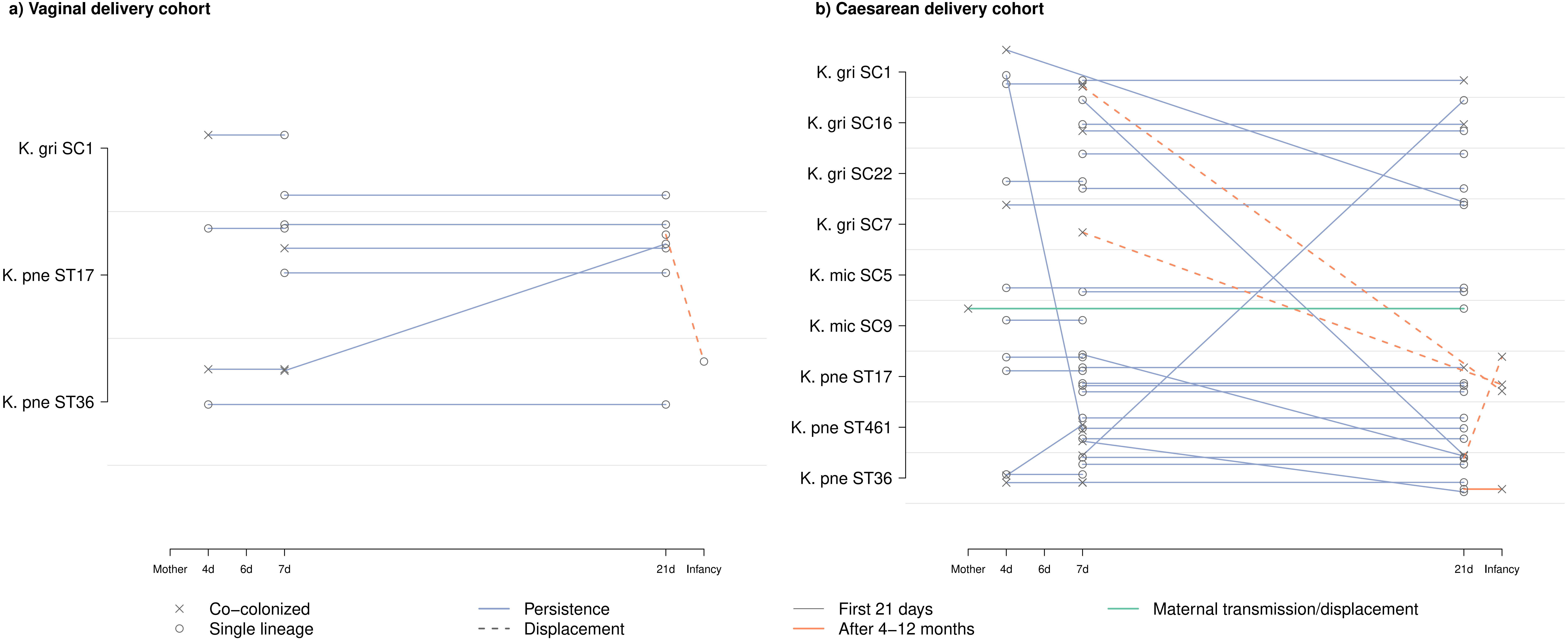


Supplementary Figure 1 Flowchart describing the data collection and analysis steps. The figure shows an overview of how the sequencing data was collected and analysed. Sampling and sequencing was performed in the source study ([1]; white background in the figure), and the analysis and reference gathering were performed in this study (grey background).

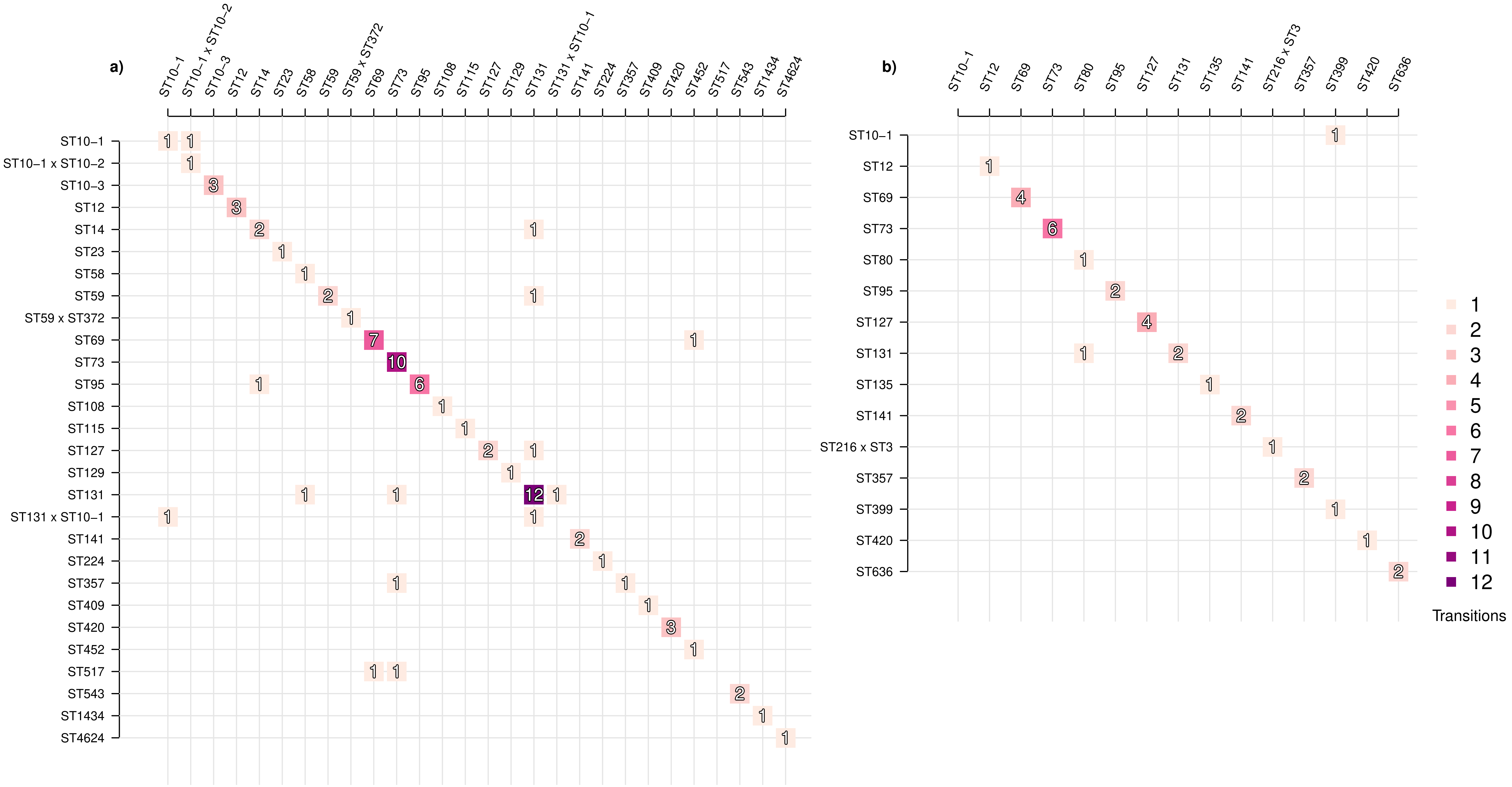
[1] Y. Shao et al., “Stunted microbiota and opportunistic pathogen colonization in caesarean-section birth,” Nature, vol. 574, no. 7776, pp. 117–121, Oct. 2019, doi: 10.1038/s41586-019-1560-1.



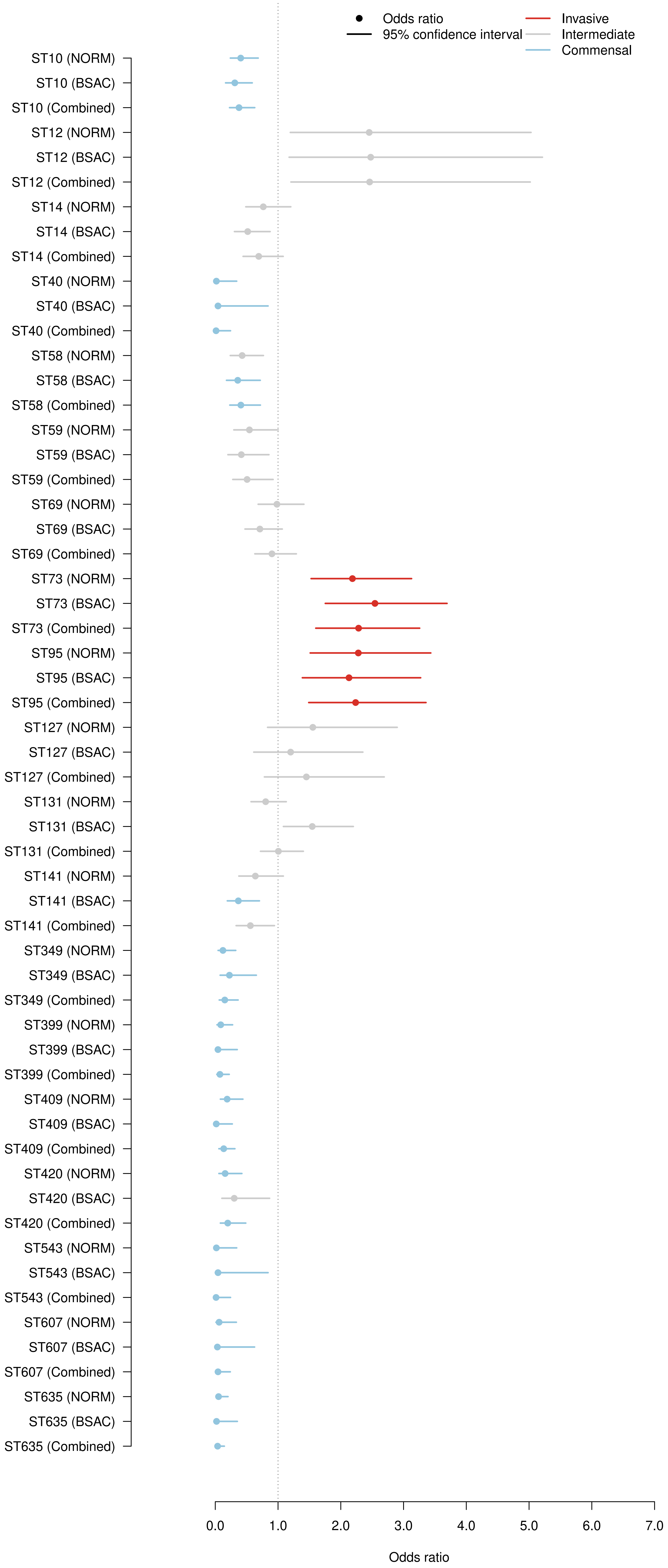
Supplementary Figure 2 Longitudinal chart showing *Escherichia coli* lineage colonisation over time. The plot shows positive identifications of *E. coli* sequence types (rows) in a sample taken at a certain time point (columns). Panel **a)** vaginal delivery cohort; panel **b)** caesarean section delivery cohort. Hollow circles represent reliable identifications of a single sequence type in the sample and black 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 line connects samples where the same lineage was identified in two consecutive time points, and a dashed line connects samples where two different lineages were identified. Lineages that were identified in both the mother and a sample from the baby are connected by a solid light green line. Dashed light green lines connect samples, where the mother carried an E. coli lineage but the sample from the baby contained a different *E. coli* lineage. Horizontal solid lines signify identification of the lineage at several time points and angled dashed lines signify a switch from one lineage to another. Only lineages which were identified at least five times are shown.



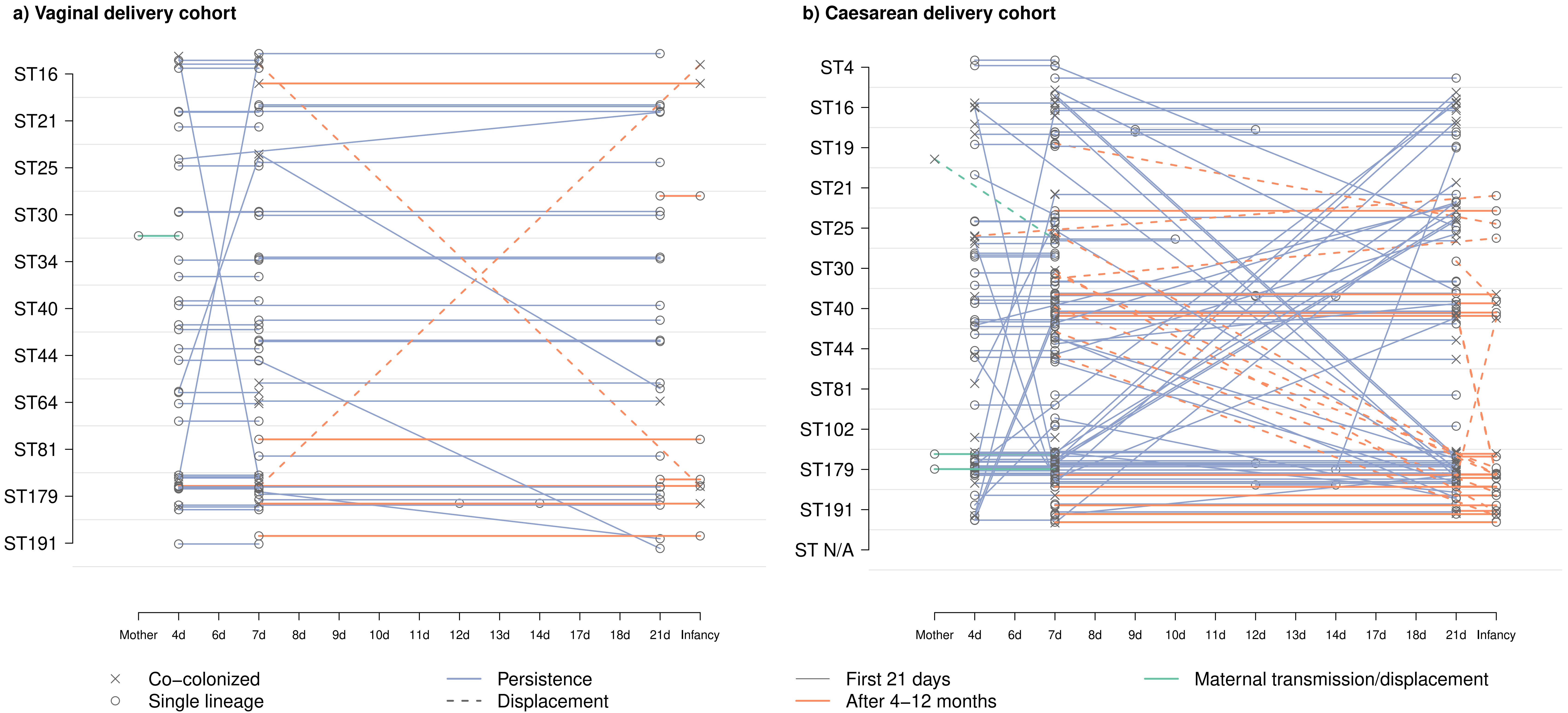
Supplementary Figure 3 Longitudinal chart showing *Klebsiella* species and lineage colonisation over time. The plot shows positive identifications of *Klebsiella* species and their sequence types (ST) or sequence clusters (SC) in the rows in a sample taken at a certain time point (columns). Panel **a)** vaginal delivery cohort; panel **b)** caesarean section delivery cohort. Hollow circles represent reliable identifications of a single sequence type in the sample and black 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, a dashed light green line transmission/displacement from the mothers, and a dashed orange line the follow-up sampling done 4-12 months later. 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.



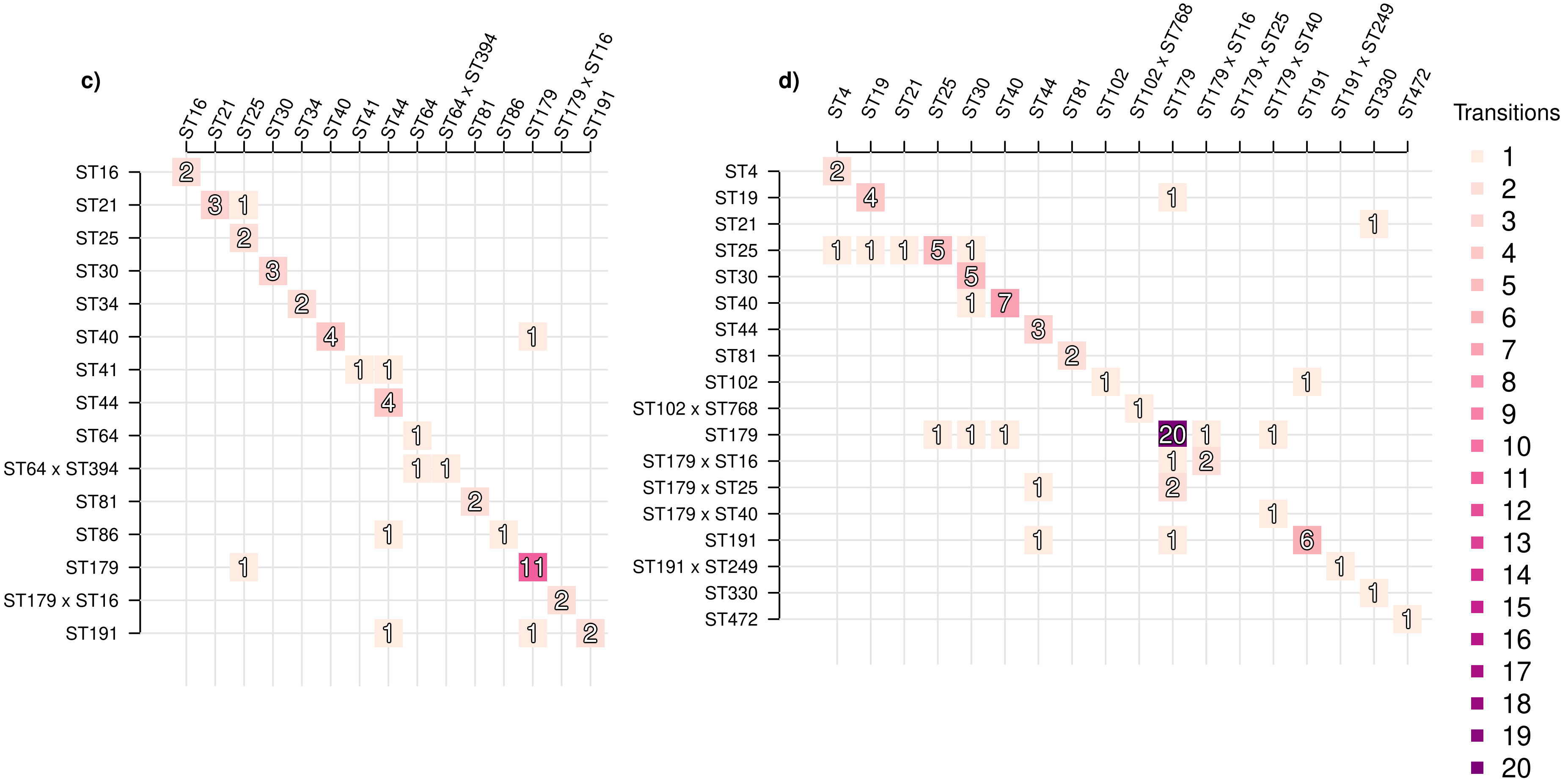
Supplementary Figure 4 Event matrix displaying colonisation identities with respect to *Escherichia coli* lineages between subsequent time points. The figure shows events corresponding to either transition from one *E. coli* lineage (rows) to another *E. coli* lineage (columns) or persistence of the same lineage (diagonal). Panel **a)** shows events for the vaginal delivery cohort with samples from the infancy period included, and panel **b)** shows the caesarean section delivery cohort with infancy period included. 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 visited at least twice across the whole set of samples.



Supplementary Figure 5 Full odds ratios for relative ***Escherichia coli*** **invasiveness.** The odds ratios (OR) for invasiveness are displayed with the 95% confidence interval, centred on the OR, where an OR of > 1 corresponds to more invasive and <1 to more commensal ST, for 19 lineages. The odds ratios are shown separately for either the NORM or the BSAC collection, or the combination of both. Lineages for which a significant OR was observed after correcting for multiple testing are coloured with a light blue or red colour. The exact p-values and sample sizes are available from the GitHub repository containing the scripts for plotting this figure (https://github.com/tmaklin/baby-microbiome-paper-plots).



Supplementary Figure 6 Longitudinal chart showing *Enterococcus faecalis* lineage colonisation over time. The plot shows positive identifications of *E. faecalis* sequence types (rows) in a sample taken at a certain time point (columns). Panel **a)** vaginal delivery cohort; panel **b)** caesarean section delivery cohort. Hollow circles represent reliable identifications of a single sequence type in the sample and black 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, a dashed light green line transmission/displacement from the mothers, and a dashed orange line the follow-up sampling done 4-12 months later. 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.



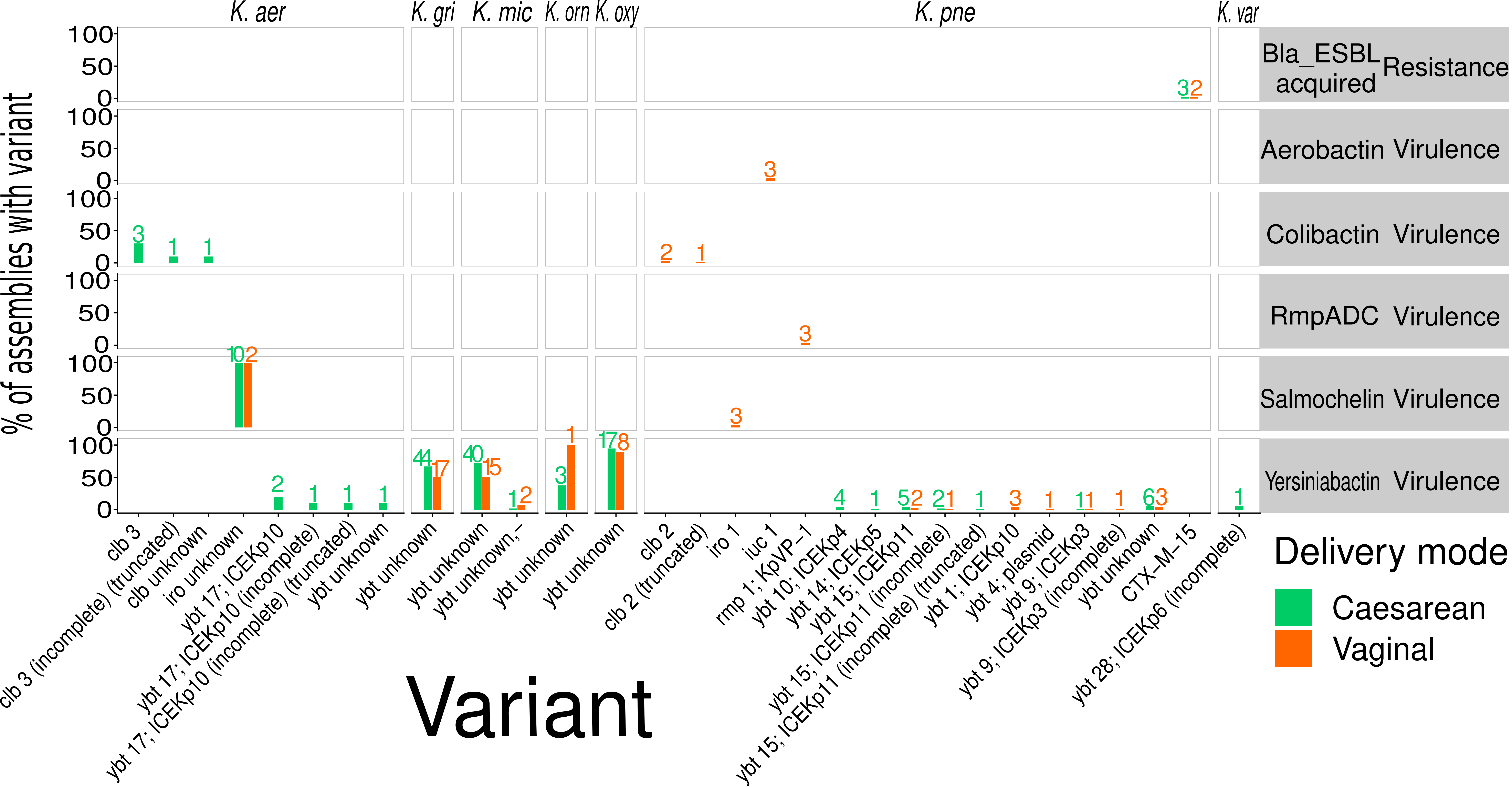
c)

	ST16	ST21	ST25	ST30	ST34	ST40	ST41	ST44	ST64	ST64 x ST394	ST81	ST86	ST179	ST179 x ST16	ST191
ST16	2														
ST21		3	1												
ST25			2												
ST30				3											
ST34					2										
ST40						4							1		
ST41							1	1							
ST44								4							
ST64									1						
ST64 x ST394									1	1					
ST81											2				
ST86								1				1			
ST179			1										11		
ST179 x ST16														2	
ST191								1					1		2

d)

	ST4	ST19	ST21	ST25	ST30	ST40	ST44	ST81	ST102	ST102 x ST768	ST179	ST179 x ST16	ST179 x ST25	ST179 x ST40	ST191	ST191 x ST249	ST330	ST472
ST4	2																	
ST19		4									1							
ST21																	1	
ST25	1	1	1	5	1													
ST30					5													
ST40					1	7												
ST44							3											
ST81								2										
ST102									1						1			
ST102 x ST768										1								
ST179				1	1	1					20	1		1				
ST179 x ST16											1	2						
ST179 x ST25							1				2							
ST179 x ST40														1				
ST191							1				1				6			
ST191 x ST249																1		
ST330																	1	
ST472																		1

Supplementary Figure 7 Event matrix displaying colonisation identities with respect to *Enterococcus 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 included, and panel **b)** shows the caesarean section delivery cohort with infancy period included. Darker shades of purple denote more common events. Lineages shown were visited at least twice across the whole set of samples.



Supplementary 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. Sequence assemblies from the caesarean section delivered cohort are highlighted in green, and assemblies from the vaginally delivered cohort in orange.

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

Supplementary Table 1 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 was used since no changes were made to the assemblies.