Group 2 and 3 ABC-transporter dependant capsular K-loci contribute significantly to variation in the invasive potential of *Escherichia coli*

Rebecca A. Gladstone[1], Maiju Pesonen[2], Anna K. Pöntinen[1], Tommi Mäklin[3], Neil MacAlasdair[1,4], Harry Thorpe[1], Yan Shao[4], Sudaraka Mallawaarachchi[1,5,6], Sergio Arredondo-Alonso[1], Benjamin J. Parcell[7], Jake David Turnbull[8], Gerry Tonkin-hill[1,5,6,9,10], Pål J. Johnsen[11], Ørjan Samuelsen[12], Nicholas R. Thomson[4,13], Trevor Lawley[4], Jukka Corander[1,3,4]

1 Department of Biostatistics, University of Oslo, Oslo, Norway
2 Oslo Centre for Biostatistics and Epidemiology, Oslo University Hospital, Oslo, Norway
3 Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland
4 Parasites and Microbes, Wellcome Sanger Institute, Hinxton, UK
5 Peter MacCallum Cancer Centre, Melbourne, Australia
6 Sir Peter MacCallum Department of Oncology, University of Melbourne, Melbourne, Australia
7 Medical Microbiology, Ninewells Hospital and Medical School, Dundee, UK
8 The National Collection of Type Cultures, Culture Collections, UK Health Security Agency
9 Department of Microbiology and Immunology, The University of Melbourne, at the Peter
10 Doherty Institute for Infection and Immunity, Melbourne, Australia
11 Department of Pharmacy, Faculty of Health Sciences, UiT The Arctic University of Norway, Tromsø, Norway
12 Norwegian National Advisory Unit on Detection of Antimicrobial Resistance, Department of Microbiology and Infection Control, University Hospital of North Norway, Tromsø, Norway
13 London School of Hygiene and Tropical Medicine, London, WC1E 7HT, United Kingdom

Corresponding authors: Rebecca A. Gladstone (r.a.gladstone@medisin.uio.no), Jukka Corander (jukka.corander@medisin.uio.no)

Funding: The project was funded by the Trond Mohn Foundation (grant identifier TMS2019TMT04 to A.K.P., R.A.G., Ø.S., P.J.J., and J.C.). The presented work has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Actions (grant No. 801,133 to S.A.-A. and A.K.P.), from Wellcome Trust (grant no. 220540/Z/20/A to YS, TL) and has also been supported by the European Research Council (grant No. 742158 to J.C.).

Abstract

The major opportunistic pathogen *Escherichia coli* is the largest cause of antimicrobial resistance (AMR) associated infections and deaths globally. Considerable antigenic diversity has been documented in Extra-intestinal pathogenic *E. coli* (ExPEC). Still, the need for systematic genomic surveys of asymptomatic colonisation and invasive disease has precluded the quantification of K-type invasive potential across different ExPEC lineages. We assembled and curated an *in-silico* capsular typing database for group 2 and group 3 K-loci from >20,000 genomes and applied it to paired carriage and disease cohorts to investigate K-type epidemiology. The most virulent circulating capsules have estimated odds ratios of >10 for being found in bloodstream infections versus carriage. The invasive potential differed markedly between lineages, and subclades of the global multi-drug resistant ST131, which displayed limited O and H antigens but substantial K-type diversity. We also discovered that insertion sequence elements contribute to the evolutionary dynamics of group 2 and group 3 K-loci by importing new capsular genes. Furthermore, the level of capsule diversity was positively correlated with more recombinogenic lineages that could adapt their antigenic repertoire faster. Our investigation highlights several K-types and lineages that contribute disproportionately to invasive ExPEC disease, which are associated with high levels of AMR. These results have significant translational potential, including improved ExPEC diagnostics, personalised therapy options, and the ability to build predictive regional risk maps by combining genomic surveys with demographic and patient frailty data.

Introduction

Capsules are major virulence determinants in bacterial pathogens. They have many critical roles, including shielding bacteria from the immune system, influencing the ability to cause invasive infections and acting as a barrier to antimicrobials. They can also influence DNA exchange, phage predation, and plasmid uptake, all of which have evolutionary underpinnings. These capsular polysaccharide antigens are often well-studied and used as a central component of bacterial disease surveillance and as effective vaccine targets, including for *Streptococcus pneumoniae*, *Haemophilus influenzae* and *Neisseria meningitidis*[1]. Whilst *Escherichia coli* is the leading cause of bacterial bloodstream infections (BSIs) globally, little is known about the current diversity of capsules in extra-intestinal (ExPEC) infections. *E. coli* has additionally been shown to represent the species causing the highest burden of antimicrobial-resistant BSI-associated deaths, making vaccine development and improved diagnostics an even greater priority[2]. Capsules are generally considered good targets for translative research to develop new antimicrobials, phage therapy and vaccines. Although some K-antigens mimic vertebrate glycoconjugates, hindering vaccine development,[3,4] there have been early *in vivo* successes demonstrating phage therapy's potential in *E. coli* invasive infections[5].

Over 80 *E. coli* K-antigen types can be identified with traditional capsular typing using a set of antisera. However, this method is labour-intensive and no longer in use. Unlike other important *E. coli* antigens, such as the O and H-antigens[6], no genotypic method exists for typing K-antigens, and a lack of paired K-phenotype and K-genotype data has precluded its development. *E. coli* capsule loci have been classified into four groups based on whether they are *wzy*-dependant (groups 1 and 4) or *kps*-ABC-transporter dependant (groups 2 and 3) and further subdivided by the genetic organisation of the capsule locus[7,8]. Small collections of isolates have highlighted associations between these capsular groups, phylogroups and disease types. The ExPEC-associated B2, D, and F phylogroups almost always carry group 2 (G2) capsules[9]. Conversely, the phylogroups A, B1, C and E are mainly associated with intestinal carriage, and these commonly carry group 1 or 4 capsules, which are rare in invasive disease[9]. The ExPEC-associated G2 and group 3 (G3) capsule loci have sets of conserved capsular polysaccharide (*kps)* genes found in regions 1 involved in assembly and export, and region 3 involved in transport. These genes flank the K-antigen determining genes in region 2 encoding glycosyl transferases and sugar synthetases. G2 and G3 K-loci are found near *serA* in the chromosome; however, recently, a subgroup of G3 K-loci have also been observed on

plasmids[10]. The conserved gene content and organisation of G2 and G3 loci make them an ideal candidate for in-silico typing. *In-silico* methods for the broader assessment of capsules across bacterial families and species have already increased our understanding of these important antigens[11,12]. A G2 and G3 *E. coli* typing scheme will complement the considerable work of others on phenotypes and structures to stimulate further study of *E. coli* K-antigens[13], as has been demonstrated for typing schemes in other species[14,15].

Here, we catalogue the incredible diversity of *E. coli* G2 and G3 capsular loci, present their prevalence in European BSIs and carriage, and quantify their relative invasiveness and associations with DNA exchange in homologous recombination, mobile genetic elements, and antimicrobial resistance (AMR). Our results highlight key virulent capsule types and genotypes that require renewed scrutiny and attention in translational efforts to reduce the future public health burden of *E. coli* invasive disease.

Results

By investigating a collection of >20,000 high-quality *E. coli* genomes, we observed 101 unique capsular ABC-transporter-dependent G2 (n=79) and G3 (n=22) K-loci, based on gene presence-absence patterns. Twenty-one of 26 known G2 phenotypes and all seven known G3 phenotypes are represented in this set. These 101 K-loci were used to create a G2 and G3 *in silico* K-typing database available at https://github.com/rgladstone/EC-K-typing, compatible with Kaptive[16]. The K-typing database enabled the assignment of K-loci to two large longitudinal genomic surveys of bloodstream infections (BSIs) from Norway[17] (2002-2017, n=3,254) and the UK[18,19] (2001-2017, n=2,219). In addition, K-typing of a UK carriage dataset from a metagenomic survey[20,21] (2014-2017, n=1,089) provided a comparator to UK BSIs. This gave unparalleled insight into the underlying colonisation dynamics that influence disease epidemiology and was essential to quantifying the risk that different K-types pose to human health.

K-locus epidemiology

Most BSIs were caused by strains with a G2 or G3 capsule (85.1%). G2 dominated; only a minority had G3 loci (3.4%). The vast majority of BSI isolates in phylogroups B2 (94.4%), D (85.4%) and F (95.4%) had G2 K-loci, and G3 K-loci were most common in phylogroup in D, B2 and A (Supplementary Data T1). The top five K-types causing BSIs were K1, K5, the K52-like K-locus (KL)189, K2 and the K14-like KL137 (Supplementary Data T2). These common K-types accounted for over fifty per cent of BSIs. Whilst G2 and G3 K-loci in UK carriage were less common (55.4%), K1 and K5 were still the most common colonising K-types.

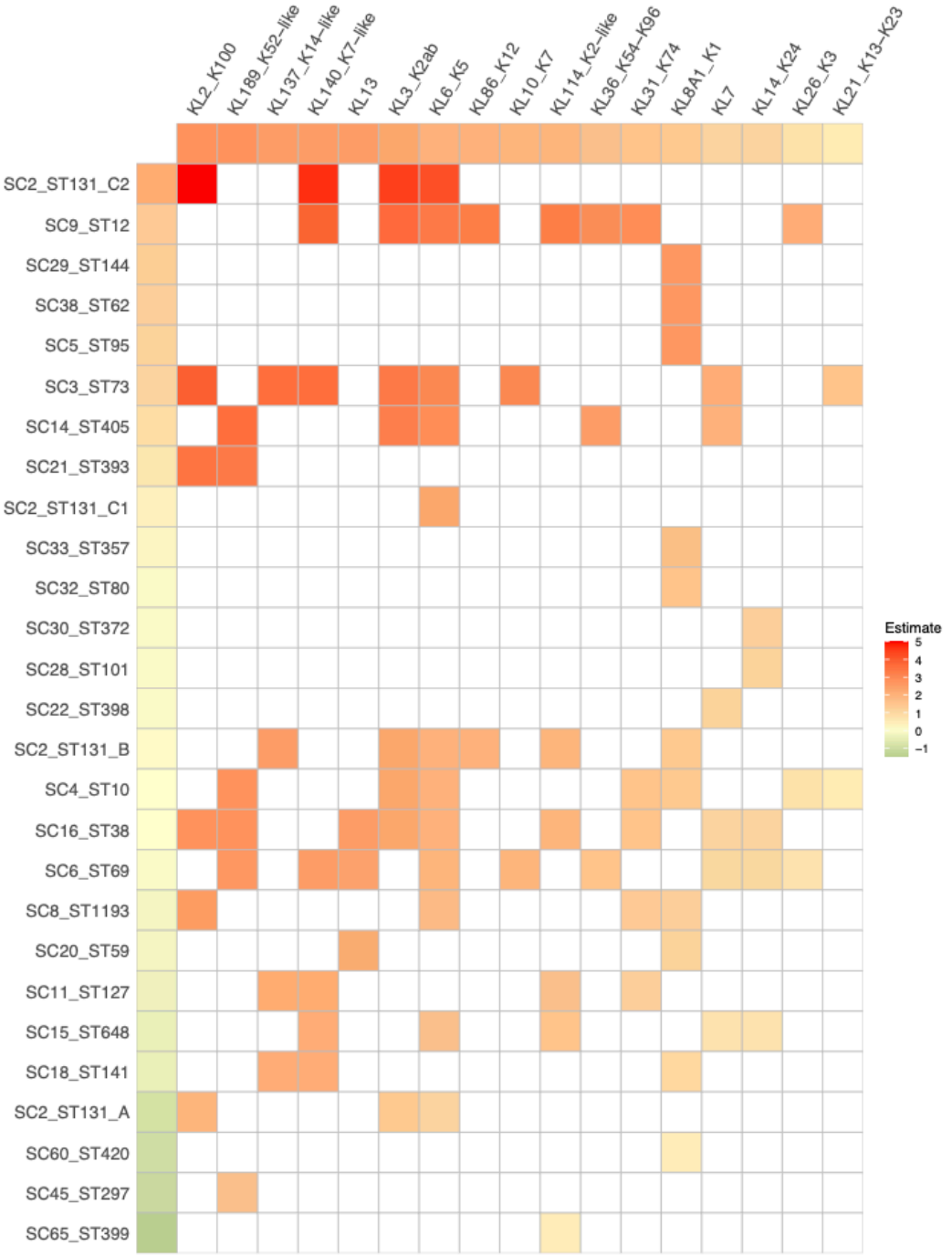
The invasive potential of K-types

Using a mixed-effect regression model for UK isolates from carriage and BSIs, we estimated the contributions of each K-locus and lineage to invasiveness (Figure 1, Supplementary Data T3-T5). We determined that K100 had the highest invasive potential (OR=16.9, 95% CI: (5.4, 53.1), p<0.0001) of the 17 most common K-loci in comparison to the average untypeable isolate (*i.e.* G2/G3 negative or rare loci). The lineage-specific odds of K100 being found in BSI was estimated to vary between 6.8 (ST131 clade A) and 151.0 (ST131 clade C2). There was also variation within ST131 clades B and C despite sharing the same O-H type. We previously described the over-representation of ST131 clade A in Norway and C2 in the UK[17]. With

knowledge of K-type, we could determine that specifically K100-clade C2 was overrepresented in the UK ($p<0.0001$), which we estimate was acquired by C2 at the end of the 20th century (1992 [1987-1996]), whilst it is K100-clade A, (the ancestral K-type in clade A, 1982 [1972-1988]) that is overrepresented in Norway ($p<0.0001$). The K52-like KL189 had the second highest invasive potential (OR=15.5, 95% CI: (7.1, 33.7), $p<0.0001$). The odd ratios for the K-types K1, K2, and K5 were ranked 13th, 6th, and 7th, respectively, and all three were significantly greater than one. These findings highlight an extensive variability in the propensity of K-type and lineage to cause infection.

We also considered the burden of *E. coli* BSI by age and sex due to immunosenescence in the elderly and a higher burden of UTIs in younger women. Patient age group and sex data were available for the UK BSI collection, and we observed that K1 was overrepresented in the <1 year ($p=0.0069$) and 40-49 years ($p=0.0048$) age groups compared to the collection as a whole, and K2 (KL3 + KL18) was over-represented in the 20-39 age group ($p=0.0234$), this burden primarily rested in females (n=11/14). We observed further sex differences in the proportion of BSIs within an age group. The K52-like KL189, predominantly found in the uropathogrenic ST69 and its double locus variant ST393, was observed in females in every decade of life but only occurred in older men aged ≥50, making it significantly overrepresented in females ($p<0.0001$), presumably due to disease progression from urinary tract infections. This may drive the general observation that the proportion of BSIs in the 1-59 age group was significantly higher for females (60%) than males (40%, $p<0.0001$) but approximately even across the life course (females 52%, males 48%). This highlights how systematic population-based K-type screening can advance understanding of the age and sex-specific epidemiology of ExPEC infections.

Figure 1. The estimated marginal and combined invasive potential of different K-loci (columns) and lineages (rows). The colours represent regression coefficients on a logit scale (log-odds) estimated from a generalised mixed model with clinical manifest (infection/carriage) as a binary outcome, K-loci (KL) as a fixed effect and lineages (sequence cluster, SC) as random effects. Red tones are associated with higher and green tones with lower estimated invasive potential. The reference category for K-loci was untypeable (*i.e.* G2/G3 negative or rare loci). The rows and columns are sorted by the invasiveness estimates from highest to lowest. The ST131 lineage is split into its major clades (A, B, C1 and C2). The K-type where known is given for each K-locus.



Group 2 K-loci association with *E.coli* and the ExPEC pathotype

E. coli G2 region 1 and 2 *kps* genes are not often found in other species. In a published collection of 661,000 assemblies[22], 11,737 were positive for the highly conserved and essential *kpsF* gene at 90% identity. These were nearly exclusively *E. coli* (99.0%, 11623/11737). Outside *E. coli*, *kpsF* was most commonly observed in *Salmonella enterica subspecies enterica*, *Klebsiella pneumoniae* (n=17) and *Staphylococcus aureus* (n=11). Using a published pangenome of ~7,500 *E. coli*[23], we assessed the pathotype association for G2. ExPECs accounted for 85.8% (1590/1853) of the *kpsF*-positive isolates with pathotype information, and non-ExPEC pathotypes accounted for 97.5% (3647/3741) of the *kpsF*-negative isolates. Shiga toxin-producing *E. coli* (STEC) accounted for the majority of non-ExPEC pathotypes positive for *kpsF* (90.5%, 238/263) and these STEC were found in multiple lineages, including ST442, ST10, ST25, ST504 and ST675.

K-loci in major MDR lineages

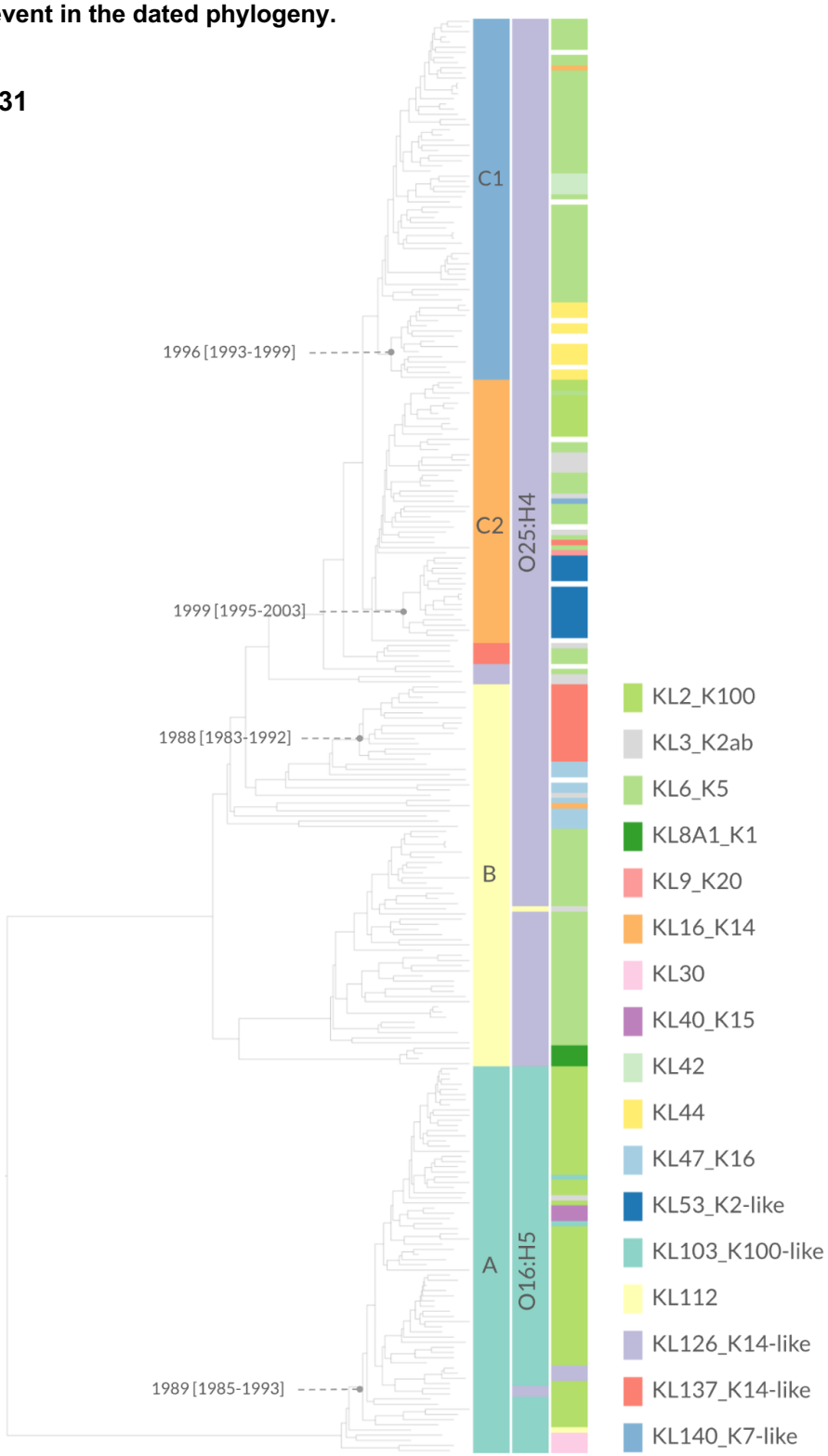
The globally disseminated multidrug-resistant (MDR) lineage ST131 (Figure 2A) is know to have two dominant O-H types. However, we observed at least 16 different K-loci that have been introduced into ST131, highlighting the rapid capsule diversification in this lineage. The ancestral K-type for the B/C clades is K5 (n=231/513) and K100 in clade A (n=93/117). Other K-loci well established in ST131 BSIs include; KL53 (K2-like, n=41/631), KL3 (K2, 34/631), KL137 (K14-like, n=29/631) and KL44 (n=25/631) all were acquired between 1980-2000. In total, three different K2/K2-like K-loci, differing by a single region 2 gene, were acquired nine independent times across the lineage. Interestingly, KL44 in ST131 is an example of a K-locus with an atypical architecture, where region 1 is followed by region 3, and region 2 is at the end of the locus. Its acquisition was followed by subsequent fragmentation and complete G2 loss events.

In another globally distributed MDR ExPEC lineage, ST69, we observed 11 O-types, 8 H-types and 18 K-loci. O-antigen diversity in this lineage is higher than for ST131 (Simpson's Diversity Index (SDI) ST69-O: 0.65, ST131-O: 0.39) but is still less than ST69 K-antigen diversity (SDI ST69-K: 0.76). K-loci belonging to G2, atypical-G2 and G3 are all found in this lineage (Figure 2B). The most common K-locus was the highly invasive K52-like KL189. On the internal node of the tree immediately following the acquisition of KL189, we detected a highly probable expansion event based on the phylodynamic properties of the dated phylogeny[24]. Subsequent switches from KL189 to the unrelated atypical KL13 occurred three separate times between 1992 and 2003. A second detectable expansion event overlaps with a subclade that has

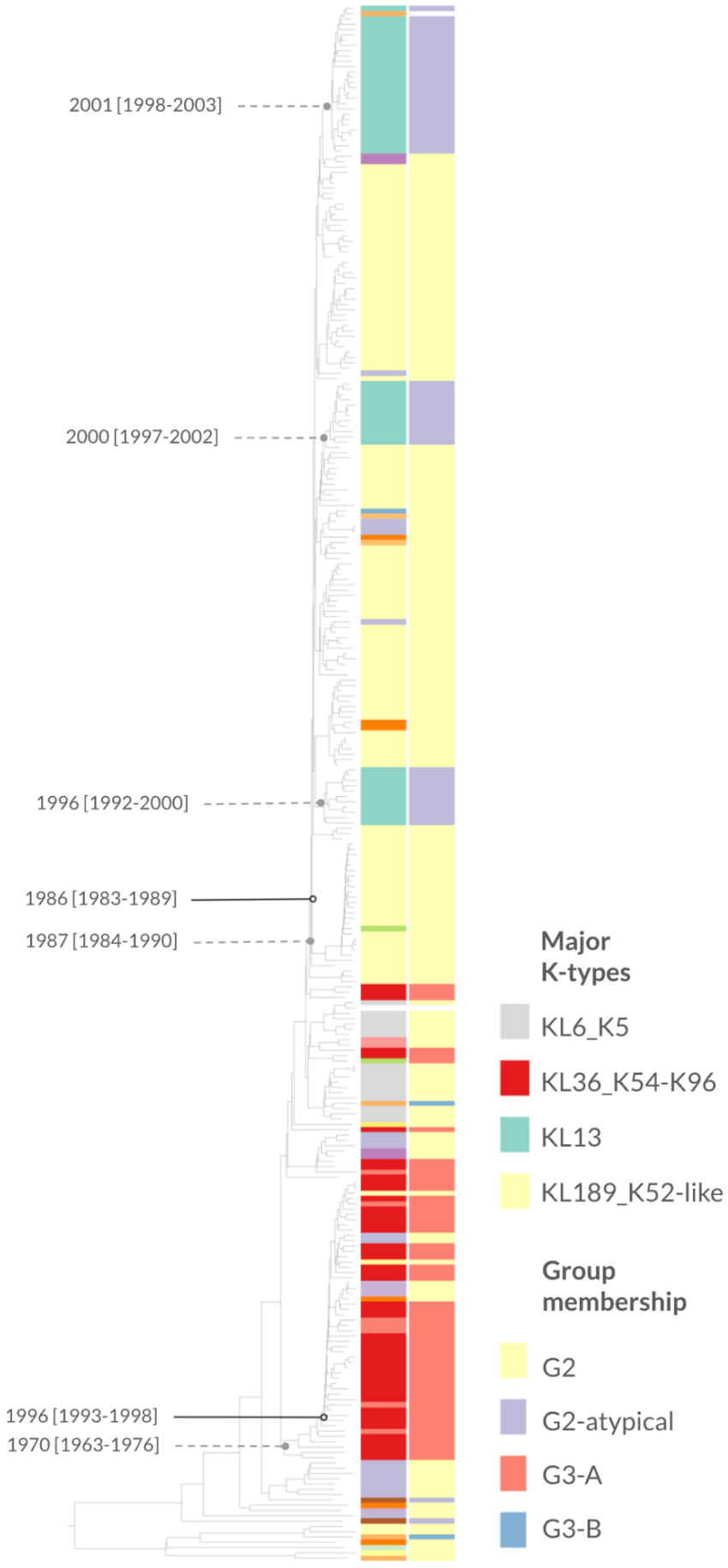
194 acquired the G3 K-locus KL36 of K54/K96 (Figure 2B). The ability to diversify the capsule locus
195 in these two MDR lineages may have facilitated their rapid expansion in BSIs in the 21st
196 century.[17–19]

Figure 2. Diversity of K-loci across the clades of ST131 and ST69 from Norwegian BSIs. Major capsule switch events are indicated with a light grey circle and dashed line from the relevant nodes with the time to the most recent common ancestor and confidence intervals. A) The four major ST131 clades are labelled in column one. The major O-H types are displayed in column 2. The K-loci are colour-coded in column 3, with the adjacent key displaying the known or closely related phenotype where phenotypic data was available. B). The major ST69 K-loci are colour-coded in column 1, and the K-group in column 2. An open circle and black line denote a detected expansion event in the dated phylogeny.

A) ST131



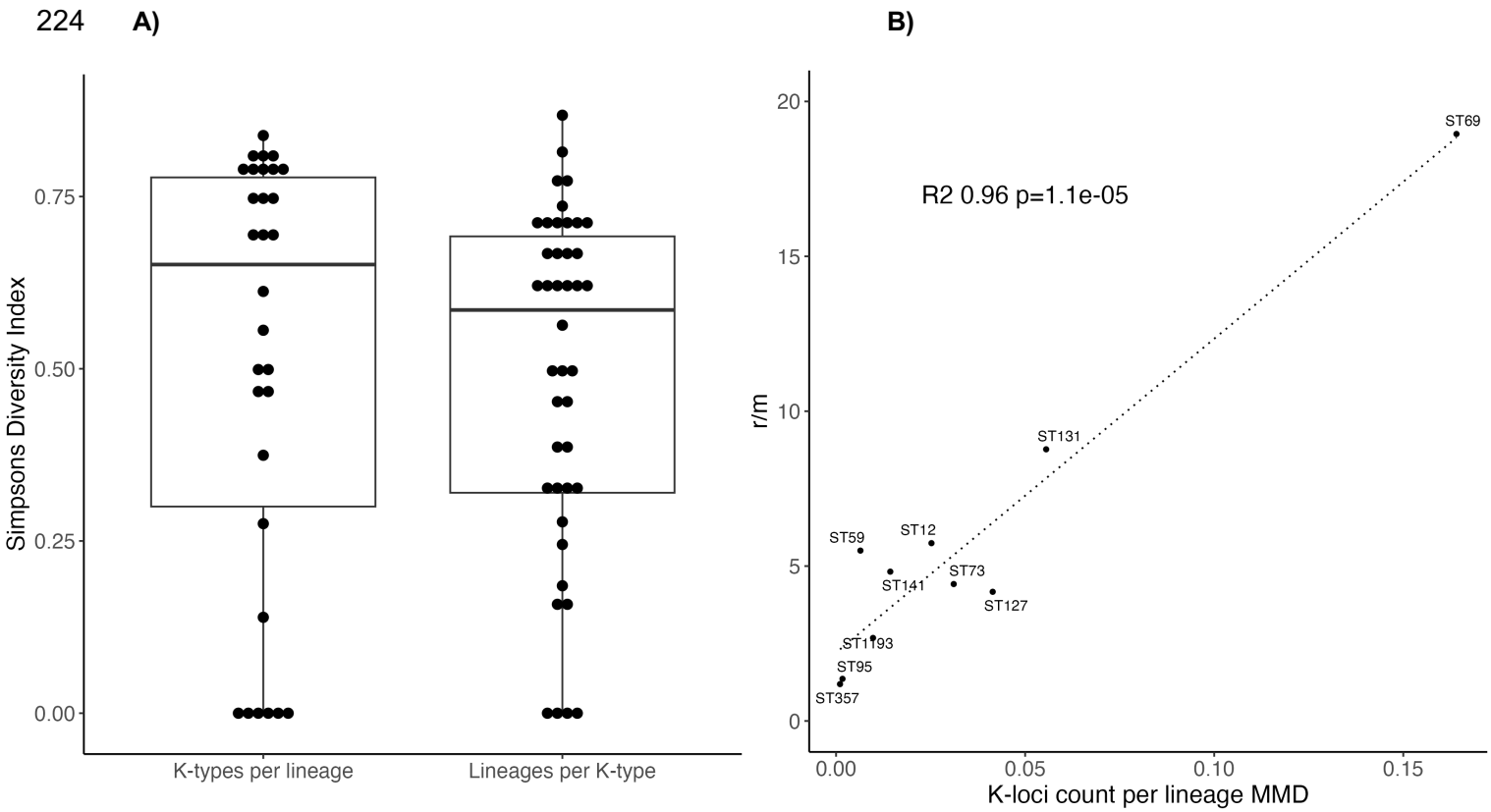
B) **ST69**



K-locus evolution

Whilst K-diversity was higher than O or H diversity in over half of the lineages observed in BSIs, lineages varied greatly in their K-type diversity, suggesting differing evolutionary processes (Figure 3A). Indeed, the number of K-loci in a lineage was strongly correlated with the r/m (recombination to mutation ratio) of that lineage, even when controlling for the lineage diversity using the recombination-free median mutational pairwise distance (MMD, R2=0.96, p<0.0001, Figure 3B). There was no correlation between r/m and MMD (R2=-0.35, p=0.3). Lineages proficient in recombination could contribute to the changes in the K-locus genetic architecture, including the observed atypical K-loci.

Figure 3. A) Simpson's diversity index for the richness and evenness of K-types within lineages and of lineages within a K-type. B) Pearson correlation between the recombination to mutation ratio (r/m) and the number of K-loci per lineage adjusted by the lineage diversity (median recombination-free mutational pairwise distance, MMD)

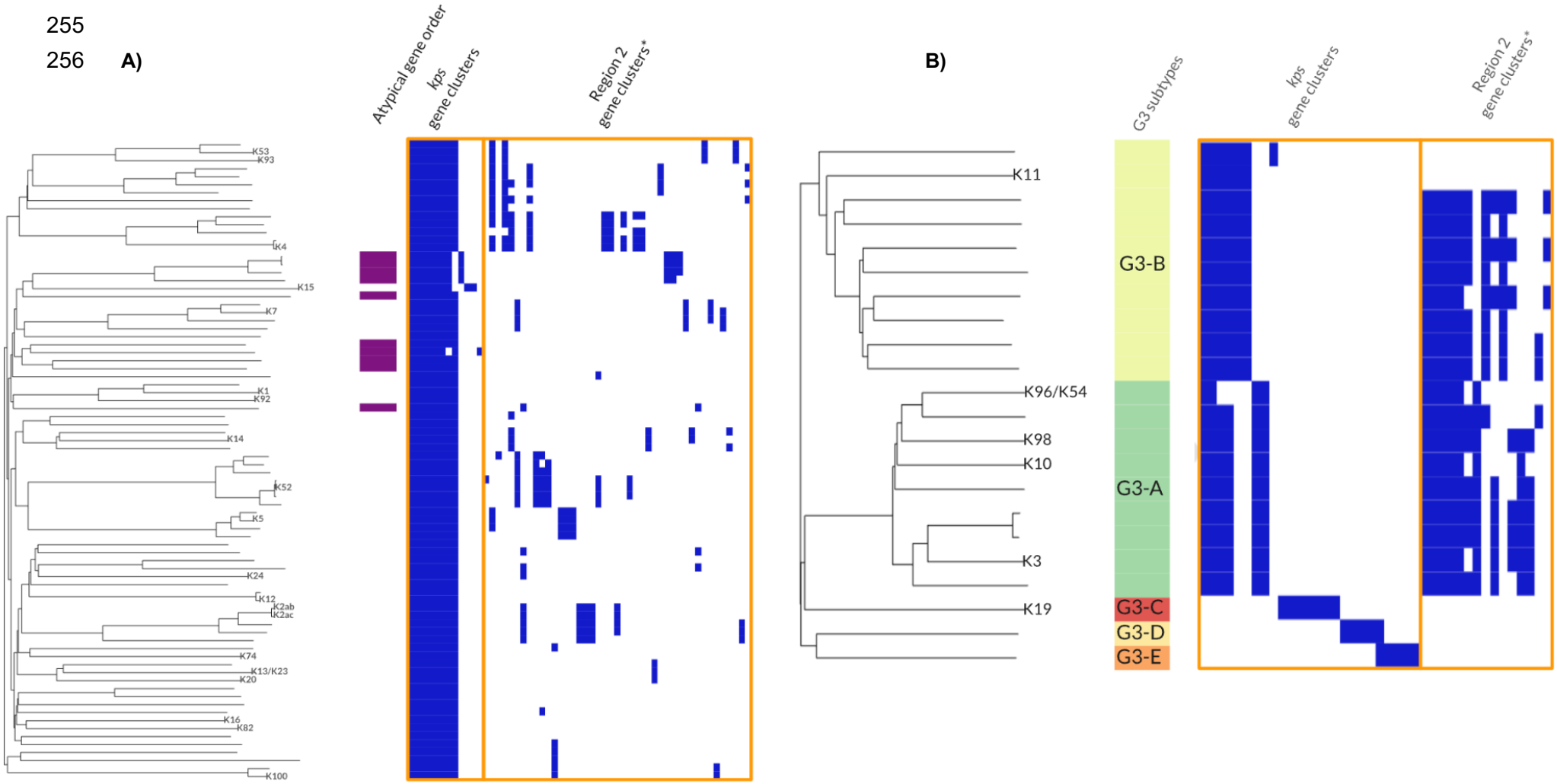


K-loci structural variation

Known K-phenotypes only account for a fraction of the numerous K-loci identified in this study, and many K-loci corresponded to deep ancestral branches that likely represent distinct K-types (Figure 4). In total, there were 229 region 2 capsular gene clusters (>70% sequence identity within clusters), of which, 48% were annotated as “hypothetical protein”, demonstrating considerable unexplored capsular determining gene variation. The eight expected *kps* genes in G2 K-loci formed one major gene cluster each. Seven of the G2 K-loci contained a *kps* gene in a separate gene cluster for at least one *kpsM, kpsT, kpsS* and/or *kpsC* gene, which are important in forming the biosynthesis-export complex[25]. Ten K-loci feature an atypical locus structure that has not been previously observed, where region 2 is outside of regions 1 and 3 (Figure 4A, Supplementary Figure 2). Although the atypical K-locus organisations are relatively rare in BSIs (3.5%), they were spread across the species phylogeny: present in phylogroups A, B2, D and F. There were also sizable clusters of atypical loci in major *E. coli* BSI lineages: ST69 (KL13, n=84/449), ST131 (KL44, n=25/631), ST73 (KL44 n=20/981), ST59 (KL13 n=14/93).

Unlike in G2, the diverse *kps* genes in G3 form two major gene clusters delineating the G3-A and G3-B subtypes (Figure 4B). No G3 *kps* genes clustered with G2 *kps* genes. G3-A and most of G3-B share the five region 2 gene clusters *wsaE* and *rmlBDAC*. Meanwhile, K11, KL66, K19, KL65 and KL75 loci also have divergent and unique region 2 gene sets (data not shown). K19, KL65 and KL75 are extremely divergent K-loci that do not share any gene clusters with G2, other G3 or each other except for the 7th *kpsS* gene cluster (*kpsS*_7) between KL65 and KL75. K19 was observed in phylogroup A in BSIs, KL66 has previously been classified as G3, and G3 have been suggested to be divergent G2 K-loci elsewhere[10,26]. Therefore, we denote them here as G3-C to G3-E.

Figure 4. Neighbour-joining trees of G2 (A) and G3 (B) K-loci, based on a distance matrix of the proportion of non-shared mash hashes, annotated with known K-phenotypes at the tips. The first metadata column denotes the K-loci with atypical gene organisation (A) and subtypes of G3 (B). The gene cluster presence/absence is presented in blue/white. The region 1 and 3 *kps* gene clusters are displayed in block one (yellow border), and the most common region 2 gene clusters are highlighted in block two (yellow border). *only gene-clusters present in >2 K-loci are presented.



K-loci diversification and mobility

This genomic survey of G2 and G3 K-loci allowed us to interrogate the mechanisms that may have generated this considerable diversity. The proportion of unique K-loci sequences (at least 1bp difference) with one or more insertion sequences (IS) was 64% (3182/4996). While K1 was only observed with an IS once (n=1/692), KL6-K5 always had at least 1 IS (n=706). IS1 and IS3 were the most common of the ten IS families observed in the K-loci collection. Of the database's 101 K-loci (unique capsular gene presence-absence patterns), 55 were observed in at least one isolate with an IS element, and an IS-element-free K-locus was not observed for 32/101 K-loci. Importantly, the IS overlapped with capsular coding sequences (CDS) in 17 K-loci and in nine of these, the IS carried a putative capsule gene in region 2 as cargo.

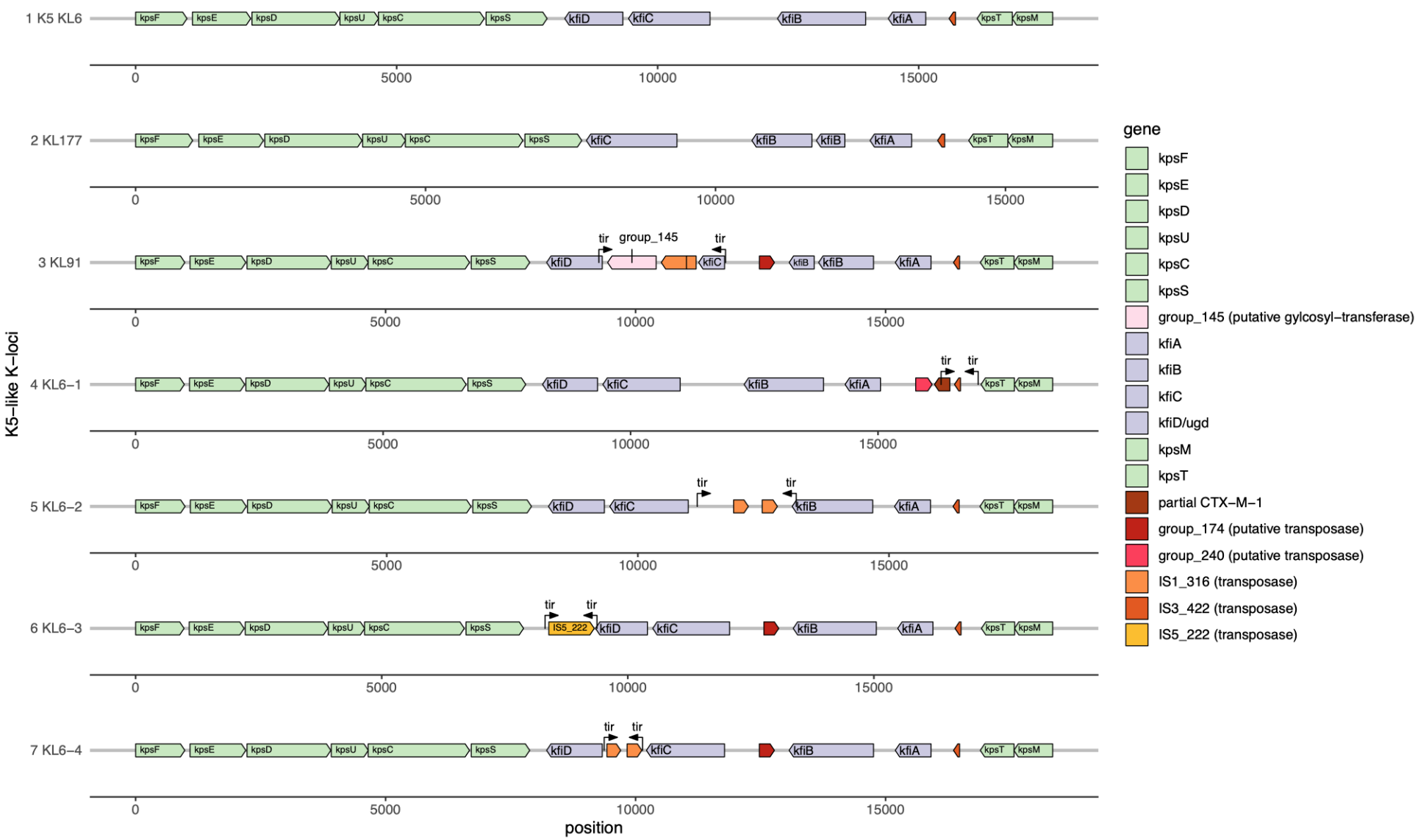
The K5 reference (KL6) was one of three closely related K-loci gene presence-absence patterns after IS genes were excluded. The KL177 locus differs from KL6_K5 as it has lost *kfiD,* though it was rare and only observed in carriage (n=3). In KL91 an IS1 316 element had a putative glycosyl-transferase family two protein (group 145) and a truncated *kfiC* within it's terminal inverted repeats (TIRs, Figure 5A) suggesting the IS moves as a larger unit. This locus was observed 4 times in BSIs and 5 times in carriage. A variant of the KL6 locus (KL6-1) seen in a single ST131 isolate had an additional putative transposase (group 240) and a fragment of the extended-spectrum β-lactamase gene $bla_{\text{CTX-M}}$ near the IS3 remnant seen in all the K5-like loci. The $bla_{\text{CTX-M}}$ and transposase gene were observed together in multiple plasmid types within ST131 (100% identity), suggesting that both genes may have then moved into this K-locus from a plasmid via homologues recombination with the existing IS in these K-loci. There were three more examples of IS in a KL6 background without any alteration to the K-locus-related gene content, each observed once in the collection used to populate the database, suggesting IS movement in and out of KL6 is happening continuously.

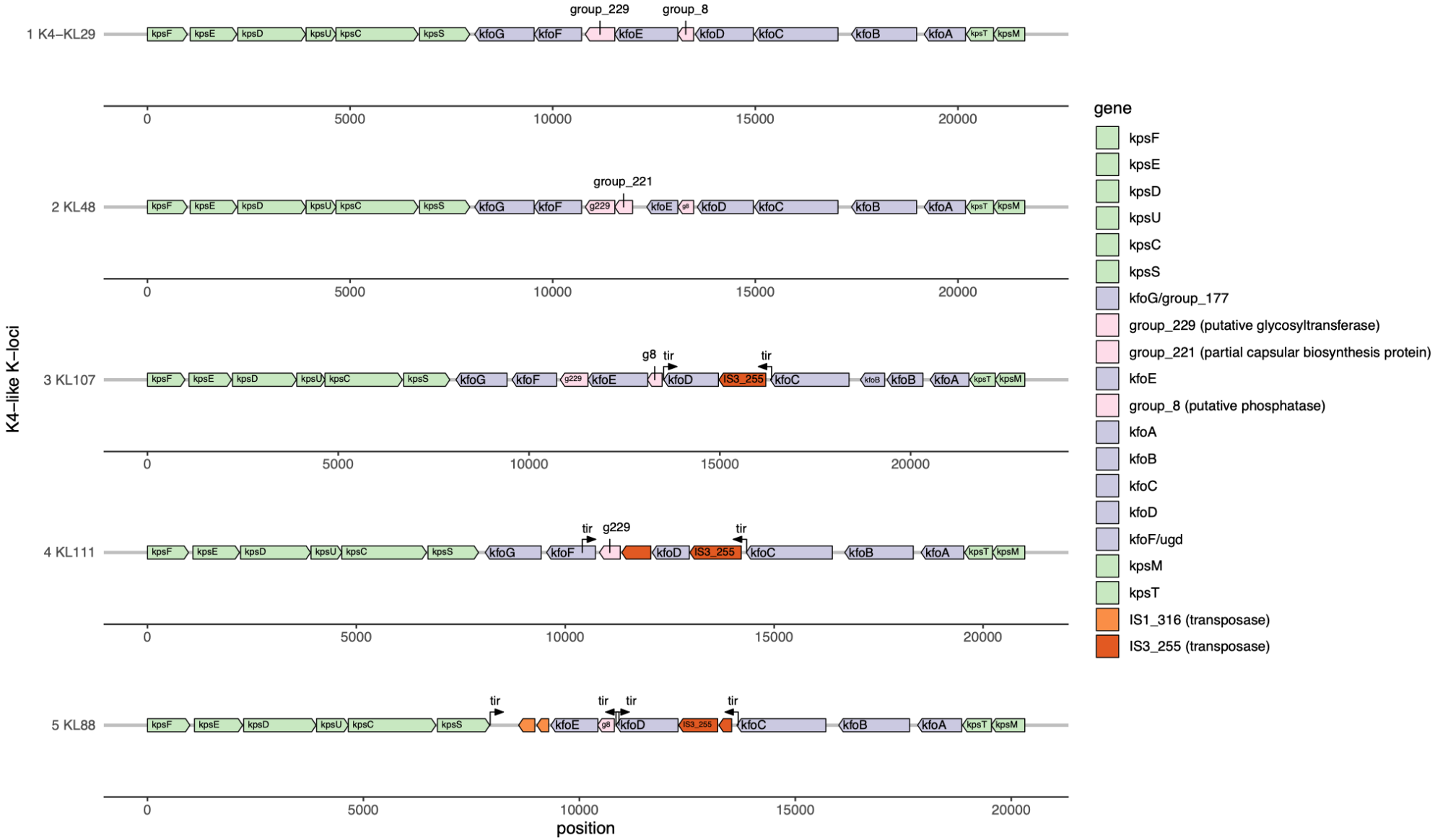
There were five different K4-related K-loci with differing gene presence-absence patterns when IS genes were excluded (Figure 5B). The *kfoA, kfoB, kfoC,* and *kfoD* genes are present in all five, and kfoF (gene-cluster name *udg*) and *kfoE* were only absent in KL88 and KL111, respectively. KL107 is an IS variant of K4-KL29 where *kfoB* is split into two fragments and kfoD is contained within an IS element. IS elements were observed between *kpsS* and *kfoC* and likely contributed to the different patterns of capsular-specific genes within region 2 for this K-locus cluster. K4-KL29 and the K4-like KL107 were the most common K4-like types in BSIs

(n=29 phylogroup B2, n=26 phylogroups D, F, B1). Whilst IS variants of K5-like loci and K4-like loci are rare, their existence gives us insight into how capsular diversification and gene exchange may be driven by IS moving between K-loci. There were no IS observed in the atypical loci, suggesting other mechanisms are also at play.

Recently, G3-B K-antigens have been observed in plasmids, which could act as a vehicle for capsular mobility[10]. For the majority (1719/1808) of K-typed Norweigan hybrid assemblies[27], the median distance of the K-locus from serA in the chromosome was confirmed to be 144,833bp (IQR 38,022bp). As G3-B were rare in BSIs, we observed just two isolates of G3-B K11 in multireplicon plasmids (IncFIA-IncFIB-IncY and IncFIA-IncY) and not in the chromosome. Concerningly, one of the K11 isolates was mobilised on a plasmid with multiple resistance genes: $bla_{\text{TEM-1}}$ *dfrA14, mph(A), sul2, aph(3'')-Ib and aph(6)-Id*.

Figure 5. Gene presence and absence for A) the K5-like loci and B) K4-related loci. Genes were clustered across the K-loci DB with a 70% ID threshold. Region 1 and 3 *kpsF-kpsM* (green), Region 2 containing genes determining the capsular type (purple), and putative capsular genes (pink). IS genes and other non-capsular genes are included here (other). When the capsular-specific gene name in the literature applies to a gene found in multiple capsular types across the database, the generic gene cluster name is also given. Putative and know functions are shown in brackets.




group_229
group_8
1 K4–KL29
kpsF kpsE kpsD kpsU kpsC kpsS kfoG kfoF kfoE kfoD kfoC kfoB kfoA kpsT kpsM
0 5000 10000 15000 20000
group_221
2 KL48
kpsF kpsE kpsD kpsU kpsC kpsS kfoG kfoF g229 kfoE g8 kfoD kfoC kfoB kfoA kpsT kpsM
0 5000 10000 15000 20000
g8 tir tir
3 KL107
kpsF kpsE kpsD kpsU kpsC kpsS kfoG kfoF g229 kfoE kfoD IS3_255 kfoC kfoB kfoB kfoA kpsT kpsM
0 5000 10000 15000 20000
tir g229 tir
4 KL111
kpsF kpsE kpsD kpsU kpsC kpsS kfoG kfoF kfoD IS3_255 kfoC kfoB kfoA kpsT kpsM
0 5000 10000 15000 20000
tir tir tir tir
5 KL88
kpsF kpsE kpsD kpsU kpsC kpsS kfoE g8 kfoD IS3_255 kfoC kfoB kfoA kpsT kpsM
0 5000 10000 15000 20000
position
K4–like K–loci
gene
kpsF
kpsE
kpsD
kpsU
kpsC
kpsS
kfoG/group_177
group_229 (putative glycosyltransferase)
group_221 (partial capsular biosynthesis protein)
kfoE
group_8 (putative phosphatase)
kfoA
kfoB
kfoC
kfoD
kfoF/ugd
kpsM
kpsT
IS1_316 (transposase)
IS3_255 (transposase)

Discussion

Existing phenotypic and structural data have demonstrated considerable diversity in the *E. coli* capsular antigens using clinical isolates from the latter half of the 20th century. Still, systematic cataloguing of capsular genetic diversity in contemporary disease is essential to further our understanding of K-antigens in the fight to control the increasing burden of extraintestinal pathogenic *E. coli.* Here, we have shown that the diversity of G2 and G3 K-loci found in BSIs far exceeds the number of the currently known G2 and G3 phenotypic capsular antigens, even when missing phenotypic data is accounted for. Furthermore, this diversity has major clinical implications, with large variations in the invasive potential of K-types in different genetic backgrounds.

Efforts to quantify the invasiveness of different K-types have been hindered by the scarcity of colonisation data, which we overcame by building on recent breakthroughs in high-resolution genotyping of pathogens from shotgun metagenomics[20]. Specifically, we leveraged large-scale UK neonatal cohort microbiome data[21,28] to survey asymptomatic *E. coli* carriage. This allowed us to determine that some of the most well-studied K-types, K1 and K5, were common in carriage and far less invasive lesser-studied K-types K52-like KL89 and K100 found in the global MDR lineages ST69 and ST131. This suggests they have a greater opportunistic spillover into vulnerable hosts. As most babies acquire *E. coli* from family members and their surroundings, reflected by the strain-sharing of the mother and the baby[20,29], we expect that the isolates seen in these healthy but vulnerable newborns represent the commensals circulating in the UK.

Currently, there are over twice as many recognised O-serogroups than K-types[30]. Yet, we observed a greater diversity of K-loci from G2 and G3 than O-loci in most dominant ExPEC lineages. This result was obtained despite the conservative 70% sequence identity threshold used here to define members of the same gene cluster compared other typing schemes[10,14,15]. Even greater diversity is likely to be discovered with further investigation into the genetic determinants and delineation of serogroups. As little as a single base pair difference has been seen in capsule-determining glycosyl-transferases of other species[31]. Indeed, we observe serotypes with the same capsular genes with high-sequence identities. These include K13 and K23 K-loci that are typed as KL21. They have previously been reported to belong to a serogroup

along with K20, which has replaced one gene relative to K13 and K23[32]. When comparing K13 (NCTC9022, ERR999921) and K23 (NCTC10430, ERR968281), there were only two non-synonymous changes in a region 2 gene; *vatD* that encodes an acetyltransferase. This putative determinant needs in-depth phenotypic validation to allow full genomic discrimination between them. The G3 K-types, K54 and K96, also share a set of capsular genes and are typed as KL36. Region 2 in these references (K96 MG736915, K54 CP9 ESC_JB5109AA_AS)[33], differ most notably by one non-synonymous change that resulted in missense for 9 AA and truncation of 8 AA at the end of the K96-07 encoding an alginate O-acetyltransferase. In addition to this, though, most KL36 in BSIs also had a truncation of 6 AA at the start of the K96_07 alginate O-acetyltransferase protein not seen in either the K54 or K96 phenotyped reference genomes. Other related clusters of K-loci which differ by a single region 2 gene or with gene truncations could also represent further cross-reactive serogroups.

In several cases, the K-locus corresponding to the phenotypic reference was rare or not observed in the longitudinal genomic surveys of BSIs[17–19]. Instead, closely related K-loci were more common, with a single gene difference in the K-locus gene presence-absence pattern. For example, KL189 was the only K52-like K-locus observed in the Norwegian and UK BSIs. From the literature, ST393 are known to be predominately O15:K52:H1 and were here genotyped as O15:KL189:H1. KL189 has an additional gene predicted to encode a CDP-glycerol glycerophosphotransferase not seen in the K52 phenotypic reference (O17:K52:H18 strain UMN026, CU928163) and a second divergent *kpsT* gene fragment. We hypothesise that the K52-like K-loci would be phenotyped as K52, yet represent potentially unrecognised cross-reactive serogroup diversity.

Taken together, our study provides further evidence for reinstating K-antigen typing for understanding and controlling ExPEC disease[10]. Furthermore, the *in-silico* typing performed here and subsequent use of the K-typing database in diverse settings provides a resource-efficient and targeted method as an alternative to routine or large-scale K-antigen phenotypic typing to identify novel K-antigens. This will further advance understanding of the role of K-antigens in pathogenicity and immunity.

The evolution of the *E. coli* G2 and G3 K-locus is also of considerable interest *per se*, given the fundamental biological role of the capsule and the complexity of its biosynthesis process. A specific subgroup of G2 was found to be undergoing atypical diversification of the K-locus, one

of which was the 5th most invasive K-locus showing that atypical K-loci are clinically important. The divergent *kps* genes seen in atypical K-loci may increase the opportunity for cross-species capsular gene exchange and their evolution warrants additional study. We were able to demonstrate a strong association between estimated recombination to mutation rate and the observed capsule diversity per genetic lineage, which suggests that the overall tendency of homologous recombination across the core chromosome closely reflects the rate at which capsule switches occur via horizontal gene transfer influencing the K locus. Earlier work detected the acquisition of the K1 capsule by genetic lineages across several phylogroups over centuries[34], and here we demonstrated that a similar process has influenced the spread of most of the successful and virulent capsule types. Notably, the capsule repertoire of a genetic lineage can grow rapidly by successful acquisitions of new capsule types that expand in the population, whilst the O and H antigens remain unchanged, as demonstrated by the ST131 C1 and C2 subclades. We further discovered that IS elements are common within the K loci and are likely contributing to the diversification of the polysaccharide composition of capsules by importing genes into region 2, warranting deeper functional investigation of these evolutionary processes. This appears to be a unique feature of *E. coli* G2 and G3 capsules; whilst IS-elements have been reported in other species like *Klebsiella pneumoniae*[14,15], they could be completely removed from the typing scheme as they were purely intergenic.

The development of ExPEC vaccines promises to improve public health by reducing the burden of invasive disease. Still, their potential is complicated because many *E. coli* are commensal, and some are considered beneficial to human health[35]. Therefore, vaccines must avoid targeting a major constituent of the gut microbiota and broad *E. coli* sterilising immunity. This requires a greater understanding of the differences between predominantly commensal genotypes and successful ExPECs, as well as the pathogenic potential of the latter. While ExPEC vaccines have been under development since the 1980s, only a few have been licensed. Whole-cell, O-antigen, H-antigen, K-antigen, and O-conjugates have all undergone clinical trials[30]. Although evidence of K-autoreactivity is limited, mimicry of host glycobiology may explain why anti-K antibodies are not strongly induced in vaccination or disease[30]. Nonetheless, Increased genomic data and expanded O-typing have contributed to the increased development of subunit-based vaccines rather than whole-cell vaccines in recent years[6,36].

A detailed quantification of the intrinsic virulence of particular genetic lineages and the contribution of the capsule to the risk of developing a bloodstream infection performed here are

key to further translational efforts needed to reduce the adverse outcomes of such invasive infections. As a substantial fraction of BSIs and life-threatening sepsis are community-acquired, novel rapid tests for use at the hospital emergency department would offer the possibility to screen patients infected with particularly virulent lineage-capsule combinations. This could not only alert clinicians to the higher risk associated with these bacterial infections but also allow personalised therapy options in the future. ExPEC infections display considerable geographical variation in the prevalence of genetic lineages, particularly between high-income and low-resource settings[37–40]. Systematic high-resolution surveillance and K-typing of both ExPEC carriage and bloodstream infections, combined with data on demographics and population-level risk factors, would enable the building of predictive regional risk maps for the expected incidence of invasive disease. This will facilitate the identification of hotspots and further research into drivers of high-risk areas, as well as improving public health strategy.

Methods

Data

To sample a wide diversity of G2 and G3 K-loci, we extracted *kps* K-loci from two published *E. coli* BSI genomic collections from Norway (2002-2017, n=3,254, 60% hybrid assemblies) and the UK (2001-2017 n=2,219), a collection of neonatal and mother *E. coli* carriage assemblies from the UK babybiome study (n=1,089, available from https://zenodo.org/records/14000489), and one health *E. coli* studies from the UK (n=405) and USA (n=2948/3111)[17–21,27,41–43]. Furthermore, we used a published pangenome analysis of ~75,00 high-quality *E. coli* genomes to determine that the capsular gene *kpsF* was consistently annotated and predictive of group 2 capsule presence [23]. We also confirmed that all *kps*F alleles observed in group 2 K-loci in the above screening collections had >90% ID. We subsequently screened for all *kpsF*-positive (90% ID) *E. coli* in a published searchable collection of 661-thousand bacterial assemblies (n=11,623)[22]. For G3 we screened the 661K database for all *kpsM*-positive assemblies (90% ID, n=853) using the group 3A and group 3B *kpsM* alleles from K54/96 and K11, respectively, as *kpsM* is more divergent in G3. All *kps*-positive genomes were supplemented by collections with phenotypic K-type information, including *kps*-positive *E. coli* in the NCTC3000 project (n=74, representing nine phenotypes, ENA study accession PRJEB6403)[44]. GenBank references and Enterobase[45] strains were also included when the K-type was reported (n=16, Supplementary Data T6). Several references with conflicting phenotype-genotypes were excluded, i.e., an isolate reported as K1 but with K5 K-locus genes. This amounted to 21,802 assemblies screened with phenotypic data for 29 unique G2 and G3 K-types.

K-loci

After extracting the *kps* K-loci from 21,802 assemblies, those with unknown bases were excluded, and the K-loci filtered down to 4,996 unique alleles with at least 1bp difference in their sequence. These unique K-loci were annotated with Prokka[46] and analysed with Panaroo[47] to consistently annotate K-loci gene clusters with a conservative 70% family identity threshold. The gene cluster names were updated using the capsular-specific reference annotations to reflect the literature where possible. Only one gene name was used across the database when a gene cluster was found in multiple K-types with different gene names. IS elements were

identified from the annotations and using ISEscan[48]. Unique capsular gene presence-absence patterns included in the K-loci database (IS gene did not contribute to the patterns). K-loci from complete genomes and long-read/hybrid assemblies were preferentially included where available. An IS-element-free version was detected for the 23/55 K-loci with IS that could be included in the database. For the remaining 32/55, a variant with the fewest observed IS genes was selected. The relatedness of the K-loci was assessed for G2 and G3 separately using the pairwise mash hash distance 1-(n/1000) to give greater resolution than gene presence-absence but still capture core and accessory variation[49]. These distances were used to create Neighbour-joining phylogenies in the R v4.4.1 package ape 5.8. Most K-loci do not have a known phenotypic K-type, which will be added retrospectively as phenotypes are confirmed or discovered, so the KL designations are arbitrary and do not mirror the K-type. KL8A1 (allele 1) and KL8A2 (allele 2) are used for K1 and K92, which share a gene set but are divergent enough to be typed separately by Kaptive.

The database was formatted for Kaptive[16], with IS annotations removed from the GenBank database. Given the limited K-phenotype data, K-loci belonging to a cluster of closely related K-loci containing one or more phenotypic references were described as K-like to give extra interpretability. Unpublished data from European disease (n=1,592) and carriage from underrepresented regions in Asia and Africa (n=10,085) were screened to determine that no dominant K-types had been missed. The *E. coli* G2-G3 database is available at https://github.com/rgladstone/EC-K-typing. The Norwegian and UK BSI collections and the babybiome collection assemblies were K-typed using Kaptive version 3.0.0b5 and this G2-G3 database. The assignment of chromosome or plasmid was available for the Norwegian hybrid assemblies[27]. These were used to determine if any K-loci were observed in plasmids in this BSI collection. For G3-B positive isolates with only Illumina data (n=4), PlasmidFinder and manual inspection were used to identify if the K-locus was chromosomal or plasmid-based[50,51].

E. coli carriage assemblies from metagenomic data

E. coli assemblies were extracted from metagenomic data from the babybiome UK data using a computational approach described in a previous study[40] to provide an *E. coli* gut colonisation dataset[40,43]. The sequencing read data was first pseudo-aligned against a diverse reference database constructed from the 661,000 assemblies study[22] (available from https://zenodo.org/records/7736981) with Themisto[52]. The alignments were then processed with mSWEEP and mGEMS[53,54] to assign each sequencing read to a bacterial species. Next, the

reads assigned to *E. coli* were pseudo-aligned with Themisto against an *E. coli* database (https://zenodo.org/records/12528310) and processed again with mSWEEP and mGEMS to obtain assignments of the sequencing reads to *E. coli* PopPUNK clusters[55]. The resulting assignments ('bins') were quality controlled with demix_check (https://github.com/harry-thorpe/demix_check) and bins with scores 1 or 2 were retained (n=1,402). Reads belonging to the 1,402 bins were assembled with Shovill (https://github.com/tseemann/shovill), and small contigs <5000bp were discarded before K-typing. As this collection sampled individuals within a family (mother and baby) multiple times, the isolates were deduplicated (n=1,089/1,402), allowing only one lineage representative with the same K-type per family. When both a typeable and untypeable isolate was observed within a family for a particular lineage, we preferentially selected the K-typed isolate as they are likely to be the same strain; this accounts for the fact that lower coverage assemblies limit K-type detection.

Invasiveness

The UK BSI collection (n=2,219) was filtered only to include isolates collected after ST69, and ST131 reached equilibrium population frequencies in 2003 (n=2,036) as a comparator to the UK carriage collection (n=1,089). K-loci, lineage information, and the clinical manifest of each isolate (infection/carriage) were input for modelling the invasiveness of K-types and lineages using a generalised mixed model approach[18–21]. The model included clinical manifest as a binary outcome variable, K-loci indicator variable as a fixed effect (constant for each isolate) and lineage indicator variable as a random effect. Untypeable was set as a reference category for K-loci. Data was filtered only to include K-loci found in >20 isolates, with >5 isolates of the infection and carriage groups each. This included 17 K-loci and n=2,683 isolates (920 carriage and 1,763 infection) in the analysis. The model was fitted in R using glmer-function in lme4-package version 1.1.35.5[56].

Lineage analysis

Lineages were defined based on core and accessory distances using PopPUNK[55]. The top 10 lineages where over 50% of the isolates had a G2 K-loci in Norwegian BSIs were mapped against the corresponding STs, aligned, and recombination removed with Gubbins[57,58]. The recombination over mutation ratio (r/m) per lineage was estimated with Gubbins, and the median pairwise recombination-free SNP distance for each lineage was calculated using pairSNP[59]. The Pearson correlation between r/m and the K-locus count per lineage (with and

513 without adjustment lineage diversity; count/median SNP distance) was calculated in R v4.4.1
514 base corr.test for the top 10 lineages with >95% G2-G3 capsules. The recombination-free
515 phylogenies for ST69 and ST131 were dated with the BactDating[60] ARC model with a
516 10,000,000 MCMC chain in three replicates and one with randomised tip dates. The effective
517 sampling size was >200, the Gelman statistic for convergence was ~1, and the true date model
518 was better than the random date model. Expansions were detected in the dated phylogenies
519 with CaveDive[24]. To characterise the diversity of K-loci within a lineage, we used Simpson's
520 diversity index (SDI), which measures the relationship between the number of "species" (here
521 we use either K-loci or lineages), termed richness, and individuals within each "species", termed
522 evenness. O:H-typing was performed with SRST2[6]. Differences between proportions were
523 tested with Fishers (when any n<5) or Chi-squared (when all n>5).

Acknowledgements

We thanks our collaborators forming The Norwegian E. coli BSI Study Group: Nina Handal (Akershus University Hospital), Nils Olav Hermansen (Oslo University Hospital, Ullevål), Anita Kanestrøm (Østfold Hospital), Hege Elisabeth Larsen (Nordland Hospital), Paul Christoffer Lindemann (Haukeland University Hospital), Iren Høyland Löhr (Stavanger University Hospital), Åshild Marvik (Vestfold Hospital), Einar Nilsen (Molde Hospital and Ålesund Hospital), Marcela Pino (Oslo University Hospital, Rikshospitalet), Elisabeth Sirnes (Førde Hospital), Ståle Tofteland (Sørlandet Hospital), Kyriakos Zaragkoulias (Nord-Trøndelag Hospital Trust). We also thank the Gates Foundation, Dr Kat Holt and Dr Kelly Wyres, who kindly advised us in maximising the database's utility with the Kaptive tool.

Declaration of interests

The authors declare no competing interests

Supplementary

Supplementary Data T1-T6
https://docs.google.com/spreadsheets/d/1ib2G6GFZDcsTzym60xRYv_nWgbl8p_mC/edit?usp=sharing&ouid=108936727405455208875&rtpof=true&sd=true

Supplementary Data T1. Distribution of G2 and G3 capsular loci across phylogroups in BSIs

Phylogroup	G2	G3
All	81.7% (4469/5473)	3.4% (191/5473)
A	25.0% (68/272)	4.4% (12/272)
B1	7.2% (22/306)	1.3% (4/306)
B2	94.4.% (3215/3723)	2.1% (77/3723)
C	0.0% (0/125)	0.0% (0/125)
D	85.5% (676/791)	11.8% (93/791)
E	0.0% (0/12)	0.0% (0/12)
F	95.4% (187/196)	2.0% (4/196)
G	0.0% (0/46)	0.0% (0/46)

543 Supplementary Data T2. Top K-types in by collection

Collection	Type	Collection size	Years	1st	2nd	3rd	4th	5th
Norway (NORM)	BSI	3254	2002-2017	K1 (KL8A1) 23.4%	K5 (KL6) 16.3%	K52-like (KL189) 5.1%	K14-like (KL137) 4.0%	K2 (KL3) 3.8%
UK (BSAC)	BSI	2186	2001-2017	K1 (KL8A1) 20.7%	K5 (KL6) 19.2%	K2ab (KL3) 6.0%	K52-like (KL189) 4.9%	K100 (KL2) 4.1%
BabyBiome (UK)	Carriage	1089	2014-2017	K1 (KL8A1) 14.9%	K5 (KL6) 7.6%	K13-K23 (KL21) 1.9%	Unknown (KL7) 1.7%	K14-like (KL137) 1.7%

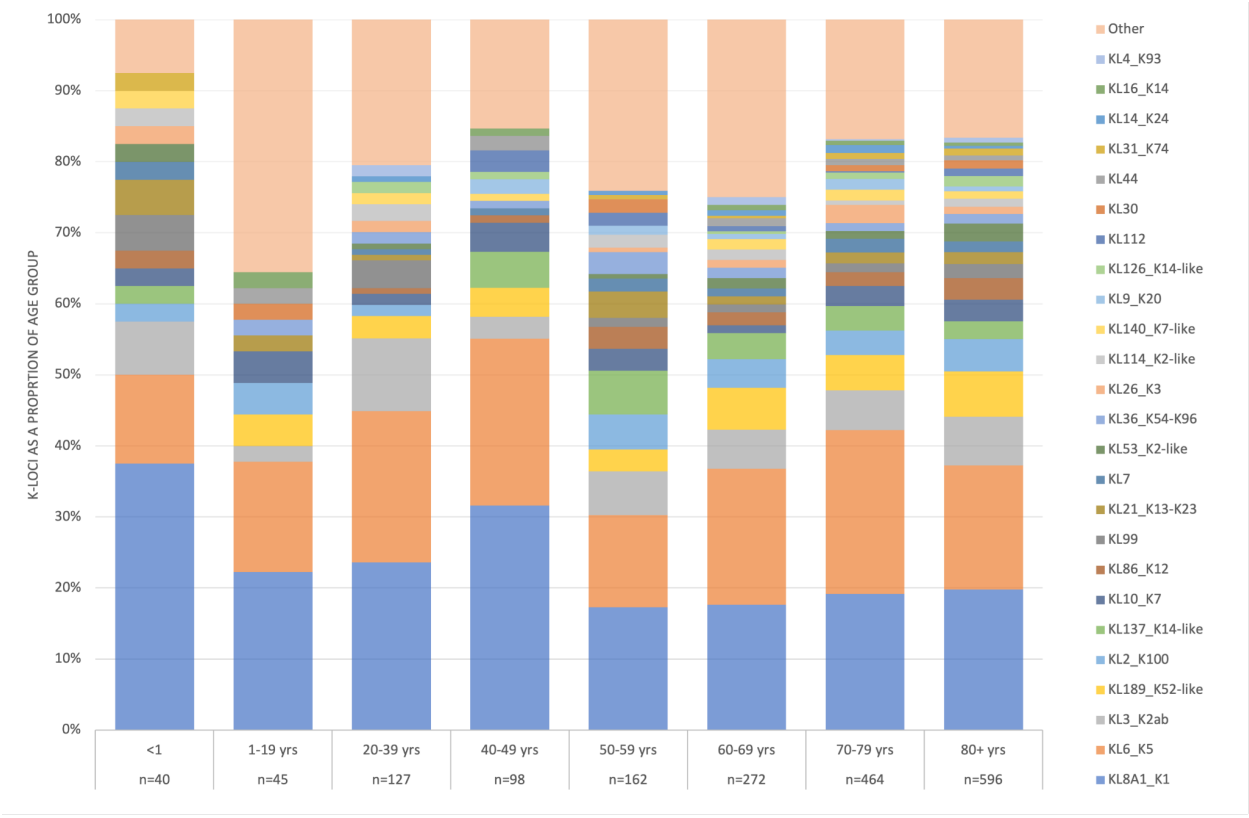
544

545 Supplementary Data T3-T5. Invasive estimates for K-type and lineages

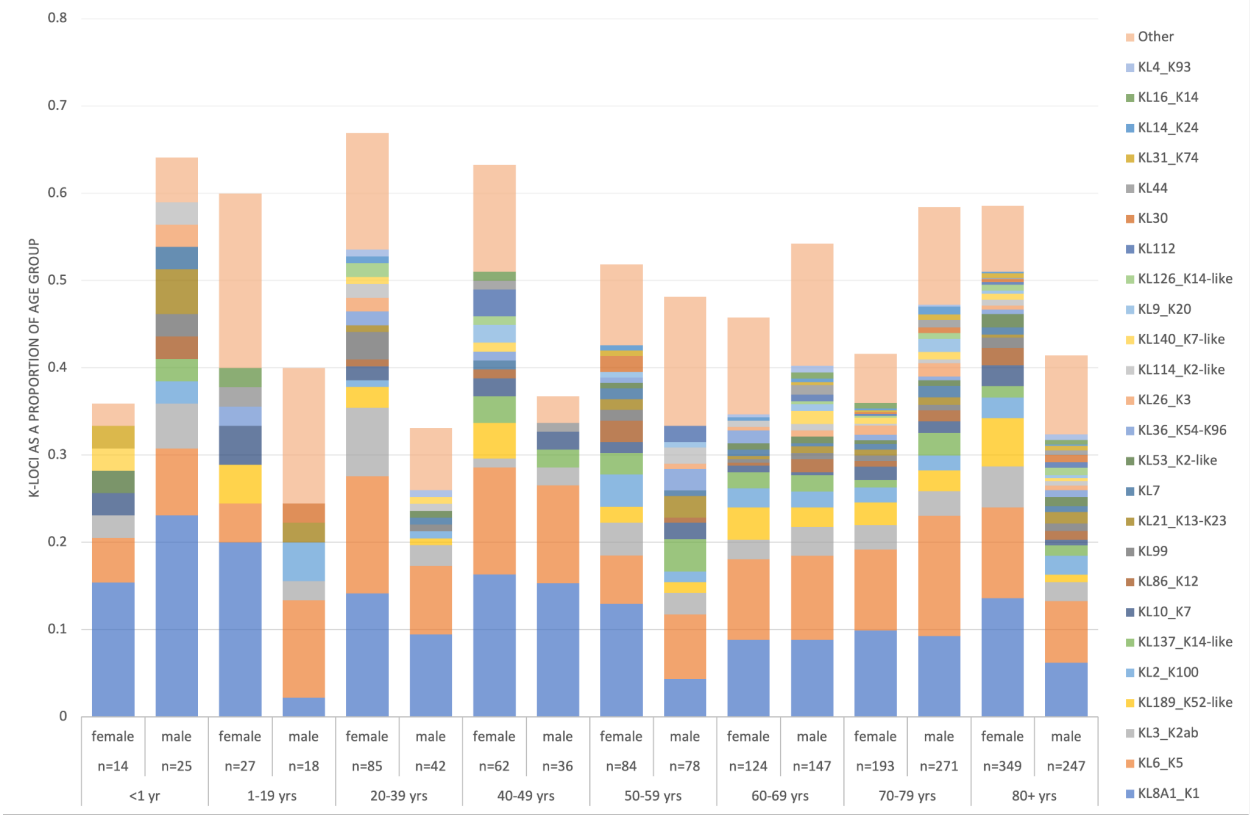
546

547 Supplementary Data T6. Isolate metadata

548 **Supplementary Figure 1. K-loci in UK BSIs A) K-loci per age group B) K-loci per age group by sex**



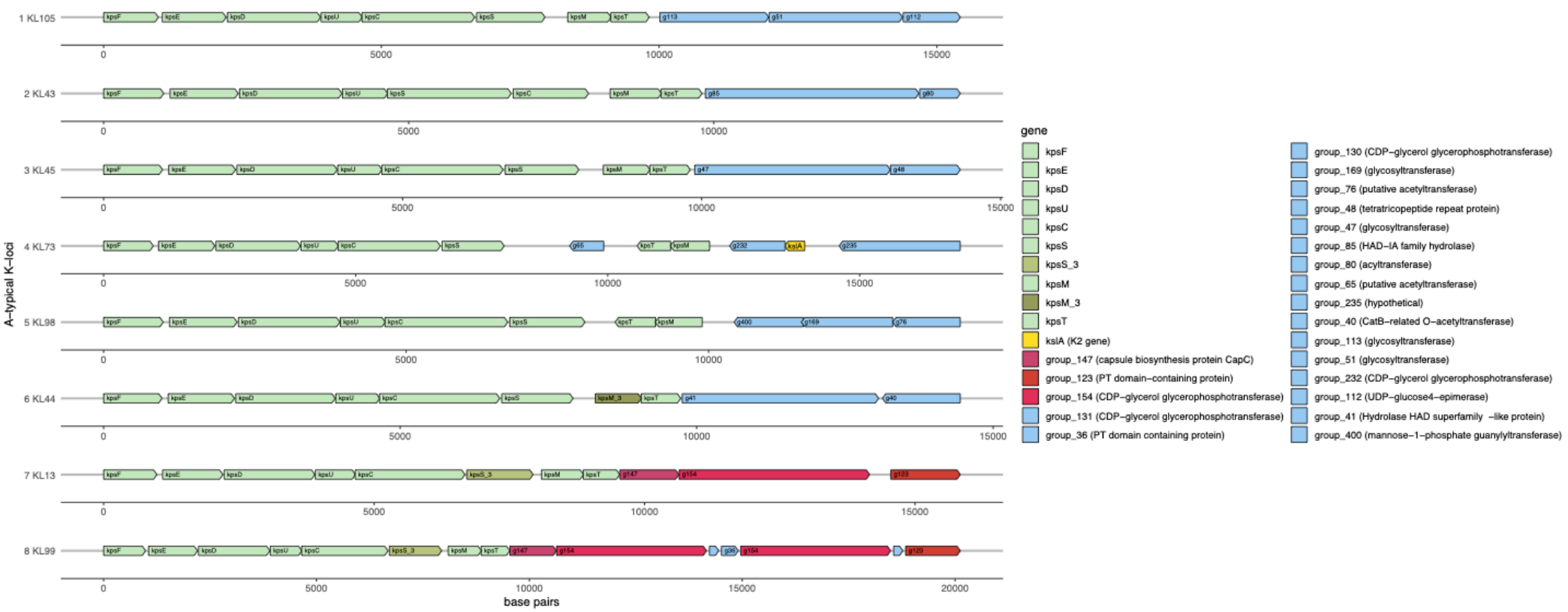
549



550

Supplementary Figure 2. Genetic architecture of atypical G2 K-loci

The conserved region 1 and 3 genes are light green, with rarer gene clusters in darker green. Region 2 genes observed in known K-types are in yellow, and those shared between atypical K-loci are in red. Genes only seen in one locus are blue. The top *E. coli* blast match is shown in brackets for unannotated gene clusters.



References

1. Mba, I. E. *et al.* Vaccine development for bacterial pathogens: Advances, challenges and prospects. *Trop. Med. Int. Health* **28**, 275–299 (2023).
2. Murray, C. J. L. *et al.* Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *Lancet* (2022) doi:10.1016/S0140-6736(21)02724-0.
3. Whitfield, C. Biosynthesis and assembly of capsular polysaccharides in Escherichia coli. *Annu. Rev. Biochem.* **75**, 39–68 (2006).
4. Sussman, M. *Escherichia Coli: Mechanisms of Virulence*. (Cambridge University Press, Cambridge, England, 1997).
5. Lin, H., Paff, M. L., Molineux, I. J. & Bull, J. J. Therapeutic Application of Phage Capsule Depolymerases against K1, K5, and K30 Capsulated E. coli in Mice. *Front. Microbiol.* **8**, 2257 (2017).
6. Ingle, D. J. *et al.* In silico serotyping of E. coli from short read data identifies limited novel O-loci but extensive diversity of O:H serotype combinations within and between pathogenic lineages. *Microb Genom* **2**, e000064 (2016).
7. Whitfield, C. Structure and Assembly of Escherichia coli Capsules. *EcoSal Plus* **3**, (2009).
8. Hong, Y., Cunneen, M. M. & Reeves, P. R. Two extremely divergent sequence forms of the genes that define Escherichia coli group 3 capsules suggest a very long history since their common ancestor. *FEMS Microbiol. Lett.* **366**, (2019).
9. Clark, J. R. & Maresso, A. M. Comparative Pathogenomics of Escherichia coli: Polyvalent Vaccine Target Identification through Virulome Analysis. *Infect. Immun.* **89**, e0011521 (2021).
10. Hong, Y., Qin, J., Forga, X. B. & Totsika, M. Extensive Diversity in Escherichia coli Group 3 Capsules Is Driven by Recombination and Plasmid Transfer from Multiple Species.

Microbiol Spectr **11**, e0143223 (2023).

11. Holt, K. E., Lassalle, F., Wyres, K. L., Wick, R. & Mostowy, R. J. Diversity and evolution of surface polysaccharide synthesis loci in Enterobacteriales. *ISME J.* **14**, 1713–1730 (2020).
12. Mostowy, R. J. & Holt, K. E. Diversity-generating machines: Genetics of bacterial sugar-coating. *Trends Microbiol.* **26**, 1008–1021 (2018).
13. Kunduru, B. R., Nair, S. A. & Rathinavelan, T. EK3D: an E. coli K antigen 3-dimensional structure database. *Nucleic Acids Res.* **44**, D675–81 (2016).
14. Wyres, K. L. *et al.* Identification of Klebsiella capsule synthesis loci from whole genome data. *Microb Genom* **2**, e000102 (2016).
15. Wyres, K. L., Cahill, S. M., Holt, K. E., Hall, R. M. & Kenyon, J. J. Identification of Acinetobacter baumannii loci for capsular polysaccharide (KL) and lipooligosaccharide outer core (OCL) synthesis in genome assemblies using curated reference databases compatible with Kaptive. *Microb Genom* **6**, (2020).
16. Lam, M. M. C., Wick, R. R., Judd, L. M., Holt, K. E. & Wyres, K. L. Kaptive 2.0: updated capsule and lipopolysaccharide locus typing for the Klebsiella pneumoniae species complex. *Microb Genom* **8**, (2022).
17. Gladstone, R. A. *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. *The Lancet Microbe* **2**, e331–e341 (2021).
18. Kallonen, T. *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.* (2017) doi:10.1101/gr.216606.116.
19. Pöntinen, A. K. *et al.* Modulation of multi-drug resistant clone success in Escherichia coli populations: a longitudinal multi-country genomic and antibiotic usage cohort study. *Lancet Microbe* (2024) doi:10.1016/S2666-5247(23)00292-6.
20. Mäklin, T. *et al.* Strong pathogen competition in neonatal gut colonisation. *Nat. Commun.*

13, 7417 (2022).

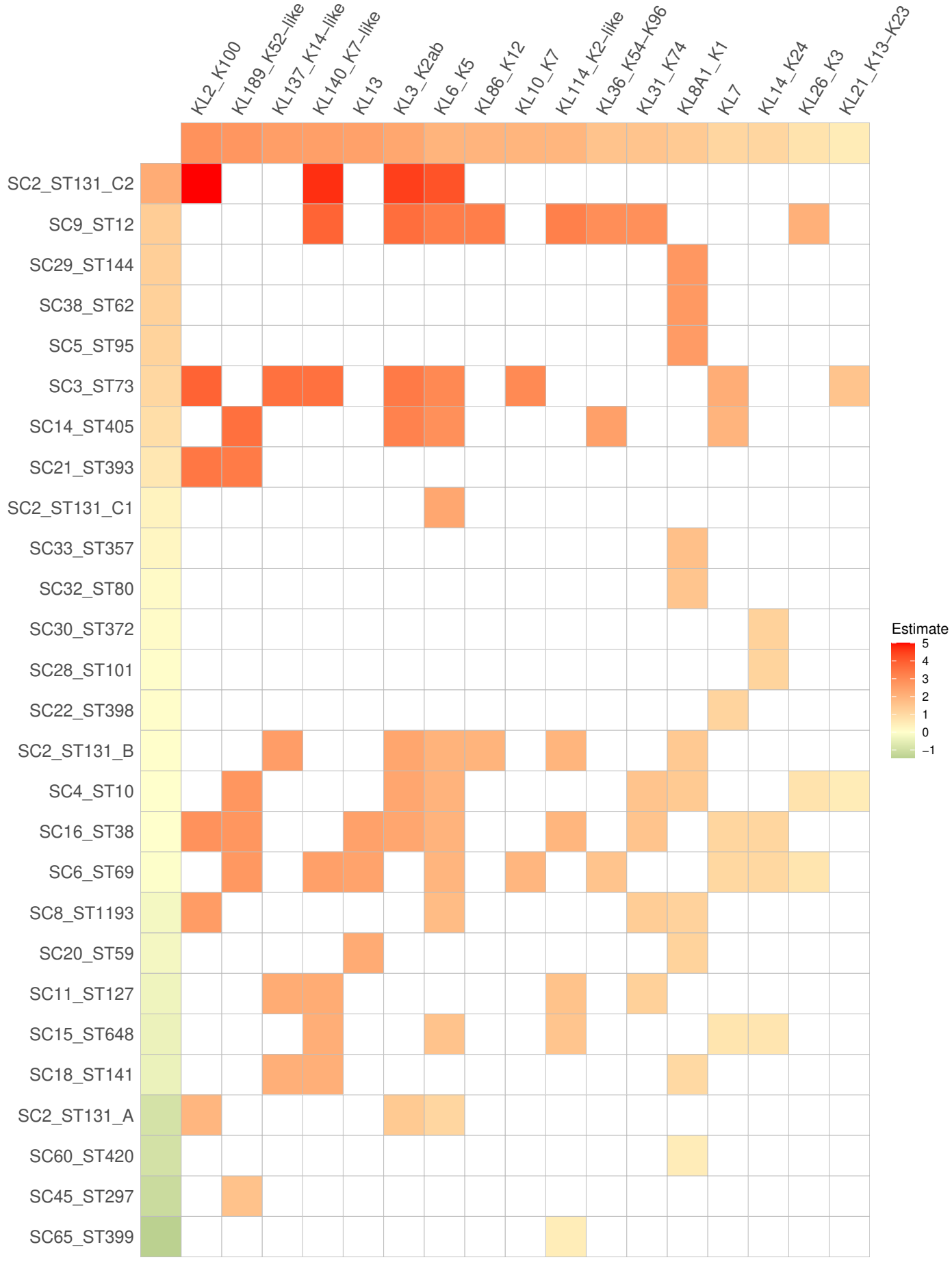
21. Shao, Y. *et al.* Primary succession of Bifidobacteria drives pathogen resistance in neonatal microbiota assembly. *Nat. Microbiol.* **9**, 2570–2582 (2024).
22. Blackwell, G. A. *et al.* Exploring bacterial diversity via a curated and searchable snapshot of archived DNA sequences. *PLoS Biol.* **19**, e3001421 (2021).
23. Horesh, G. *et al.* A comprehensive and high-quality collection of Escherichia coli genomes and their genes. *Microb Genom* **7**, (2021).
24. Helekal, D., Ledda, A., Volz, E., Wyllie, D. & Didelot, X. Bayesian inference of clonal expansions in a dated phylogeny. *Syst. Biol.* (2021) doi:10.1093/sysbio/syab095.
25. Whitfield, C. & Roberts, I. S. Structure, assembly and regulation of expression of capsules in Escherichia coli. *Mol. Microbiol.* **31**, 1307–1319 (1999).
26. Sande, C. & Whitfield, C. Capsules and Extracellular Polysaccharides in Escherichia coli and Salmonella. *EcoSal Plus* **9**, eESP00332020 (2021).
27. Arredondo-Alonso, S. *et al.* Plasmid-driven strategies for clone success in*Escherichia coli*. *bioRxiv* 2023.10.14.562336 (2023) doi:10.1101/2023.10.14.562336.
28. Shao, Y. *et al.* Stunted microbiota and opportunistic pathogen colonization in caesarean-section birth. *Nature* **574**, 117–121 (2019).
29. Tonkin-Hill, G. *et al.* Enhanced metagenomics-enabled transmission inference with TRACS. *bioRxiv* 2024.08.19.608527 (2024) doi:10.1101/2024.08.19.608527.
30. Qiu, L. *et al.* Vaccines against extraintestinal pathogenic Escherichia coli (ExPEC): progress and challenges. *Gut Microbes* **16**, 2359691 (2024).
31. Mavroidi, A. *et al.* Evolutionary genetics of the capsular locus of serogroup 6 pneumococci. *J. Bacteriol.* **186**, 8181–8192 (2004).
32. Vann, W. F. *et al.* Serological, chemical, and structural analyses of the Escherichia coli cross-reactive capsular polysaccharides K13, K20, and K23. *Infect. Immun.* **39**, 623–629 (1983).

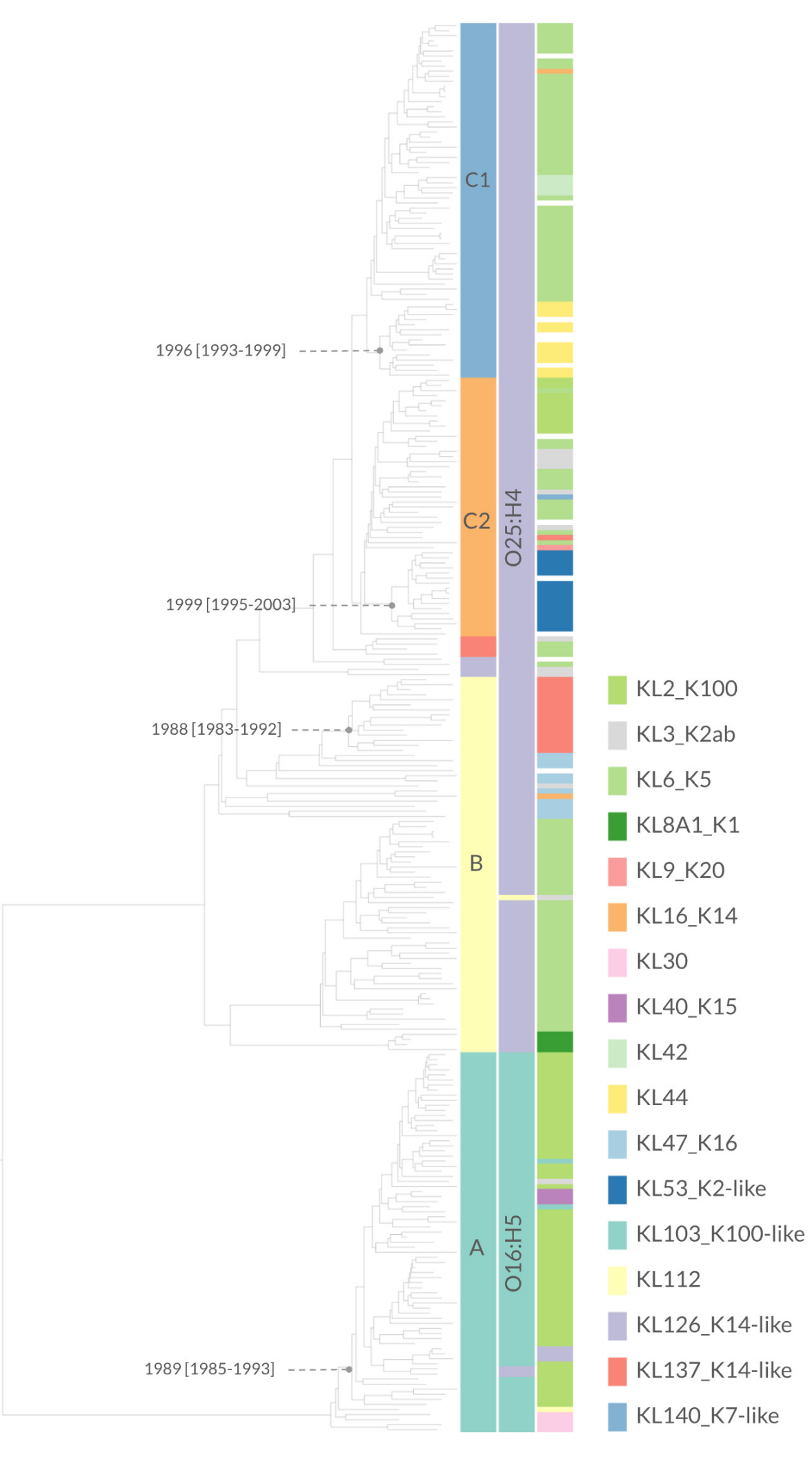
631 33. Olson, M. A. *et al.* Bile salts regulate zinc uptake and capsule synthesis in a mastitis-
632 associated extraintestinal pathogenic Escherichia coli strain. *Infect. Immun.* **89**, e0035721
633 (2021).
634 34. Arredondo-Alonso, S. *et al.* Evolutionary and functional history of the Escherichia coli K1
635 capsule. *Nat. Commun.* **14**, 3294 (2023).
636 35. Wassenaar, T. M. Insights from 100 years of research with probiotic E. coli. *Eur. J.*
637 *Microbiol. Immunol. (Bp.)* **6**, 147–161 (2016).
638 36. Weerdenburg, E. *et al.* Global distribution of O serotypes and antibiotic resistance in
639 extraintestinal pathogenic Escherichia coli collected from the blood of patients with
640 bacteremia across multiple surveillance studies. *Clin. Infect. Dis.* **76**, e1236–e1243 (2023).
641 37. Singh, S. R. *et al.* Whole genome sequencing of multidrug resistant Enterobacterales
642 identified in children and their household members within Siem Reap, Cambodia. *JAC*
643 *Antimicrob Resist* **5**, dlad067 (2023).
644 38. Muloi, D. M. *et al.* Population genomics of Escherichia coli in livestock-keeping households
645 across a rapidly developing urban landscape. *Nat. Microbiol.* **7**, 581–589 (2022).
646 39. Kantele, A. *et al.* Dynamics of intestinal multidrug-resistant bacteria colonisation contracted
647 by visitors to a high-endemic setting: a prospective, daily, real-time sampling study. *Lancet*
648 *Microbe* **2**, e151–e158 (2021).
649 40. Khawaja, T. *et al.* Deep sequencing of Escherichia coli exposes colonisation diversity and
650 impact of antibiotics in Punjab, Pakistan. *Nat. Commun.* (2024) doi:10.1038/s41467-024-
651 49591-5.
652 41. Liu, C. M. *et al.* Using source-associated mobile genetic elements to identify zoonotic
653 extraintestinal E. coli infections. *One Health* **16**, 100518 (2023).
654 42. Ludden, C. *et al.* One Health Genomic Surveillance of Escherichia coli Demonstrates
655 Distinct Lineages and Mobile Genetic Elements in Isolates from Humans versus Livestock.
656 *MBio* **10**, (2019).

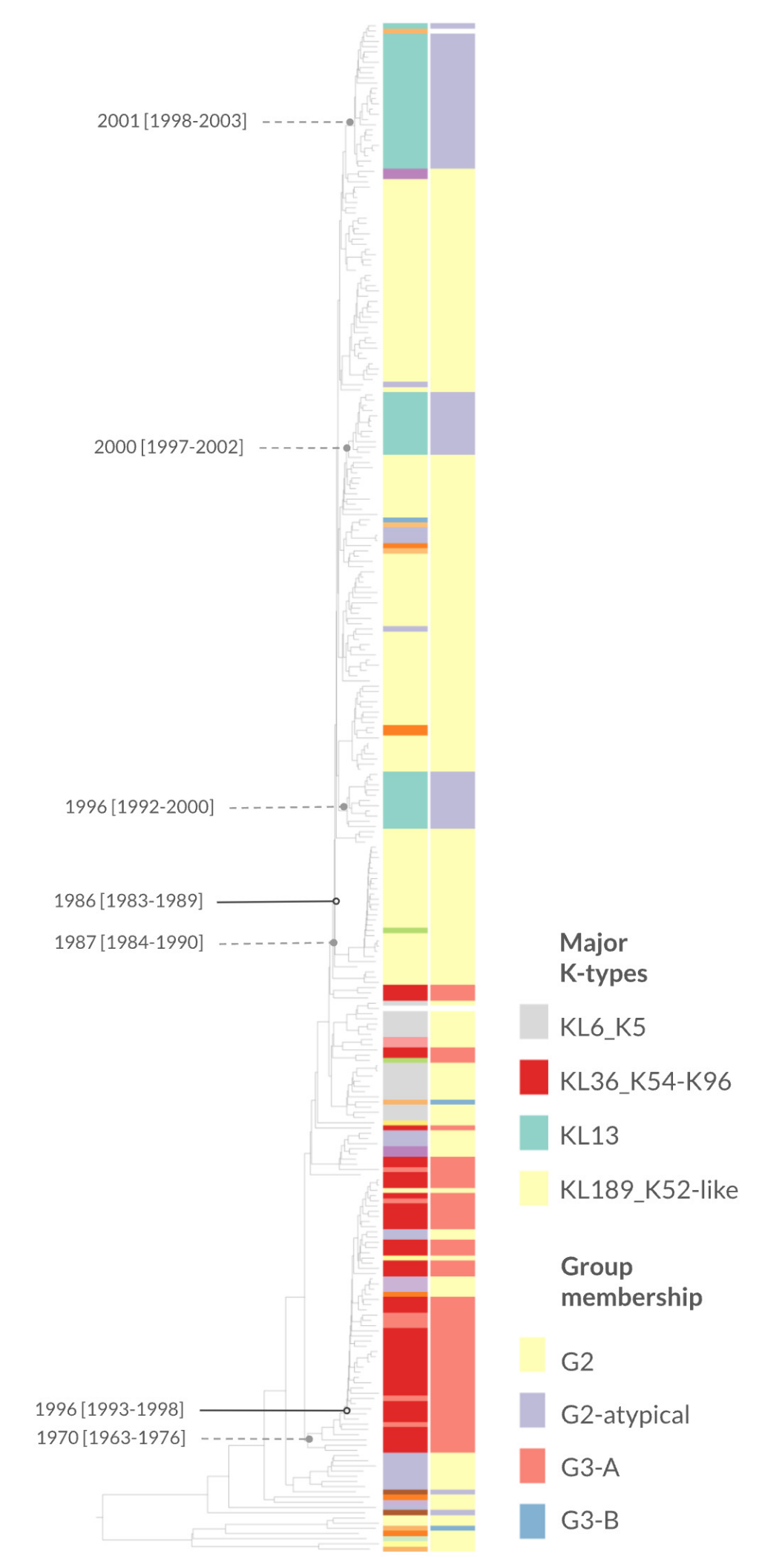
43. Mäklin, T. *et al.* Geographical variation in colorectal and urinary tract linked cancer incidence is associated with population exposure to colibactin-producing *Escherichia coli*. *Lancet Microbe* (2024) doi:10.1016/69j.lanmic.2024.101015.
44. Dicks, J. *et al.* NCTC3000: a century of bacterial strain collecting leads to a rich genomic data resource. *Microb. Genom.* **9**, mgen000976 (2023).
45. Zhou, Z., Charlesworth, J. & Achtman, M. HierCC: a multi-level clustering scheme for population assignments based on core genome MLST. *Bioinformatics* **37**, 3645–3646 (2021).
46. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* (2014).
47. Tonkin-Hill, G. *et al.* Producing polished prokaryotic pangenomes with the Panaroo pipeline. *Genome Biol.* **21**, 180 (2020).
48. Xie, Z. & Tang, H. ISEScan: automated identification of insertion sequence elements in prokaryotic genomes. *Bioinformatics* **33**, 3340–3347 (2017).
49. Ondov, B. D. *et al.* Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.* **17**, 132 (2016).
50. Carattoli, A. & Hasman, H. PlasmidFinder and in silico pMLST: Identification and typing of Plasmid replicons in whole-genome sequencing (WGS). *Methods Mol. Biol.* **2075**, 285–294 (2020).
51. Carattoli, A. *et al.* In silico detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing. *Antimicrob. Agents Chemother.* **58**, 3895–3903 (2014).
52. Alanko, J. N., Vuohtoniemi, J., Mäklin, T. & Puglisi, S. J. Themisto: a scalable colored k-mer index for sensitive pseudoalignment against hundreds of thousands of bacterial genomes. *Bioinformatics* **39**, i260–i269 (2023).
53. Mäklin, T. *et al.* High-resolution sweep metagenomics using fast probabilistic inference. *Wellcome Open Res.* **5**, 14 (2021).

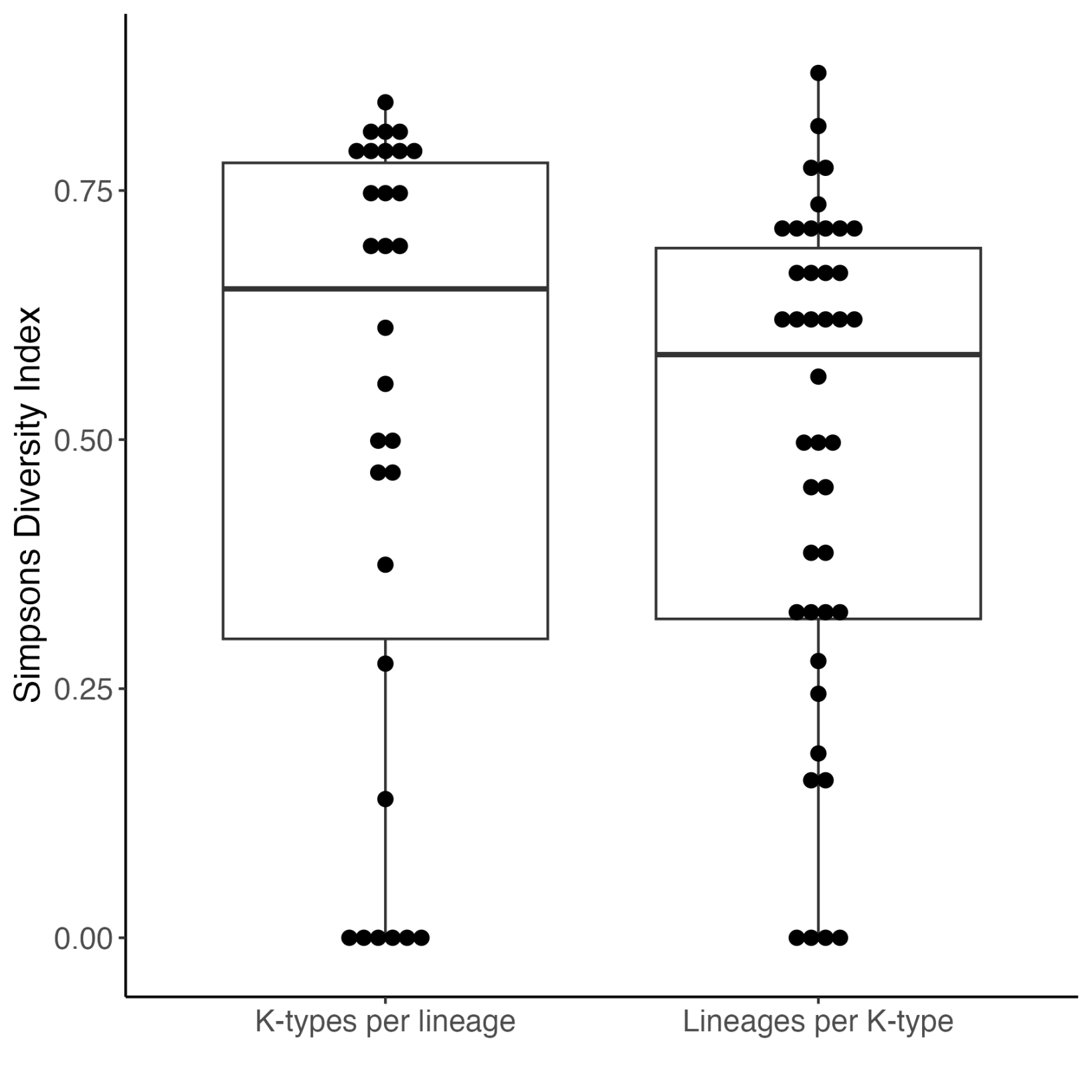
683 54. Mäklin, T. *et al.* Bacterial genomic epidemiology with mixed samples. *Microb Genom* **7**,
684 (2021).
685 55. Lees, J. A. *et al.* Fast and flexible bacterial genomic epidemiology with PopPUNK. *Genome*
686 *Res.* **29**, 304–316 (2019).
687 56. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting linear mixed-effects models
688 Usinglme4. *J. Stat. Softw.* **67**, (2015).
689 57. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler
690 transform. *Bioinformatics* **25**, 1754–1760 (2009).
691 58. Croucher, N. J. *et al.* Rapid phylogenetic analysis of large samples of recombinant bacterial
692 whole genome sequences using Gubbins. *Nucleic Acids Res.* **43**, e15 (2015).
693 59. Tonkin-Hill, G. *Pairsnp-Cpp*. (Github).
694 60. Didelot, X., Croucher, N. J., Bentley, S. D., Harris, S. R. & Wilson, D. J. Bayesian inference
695 of ancestral dates on bacterial phylogenetic trees. *Nucleic Acids Res.* **46**, e134 (2018).

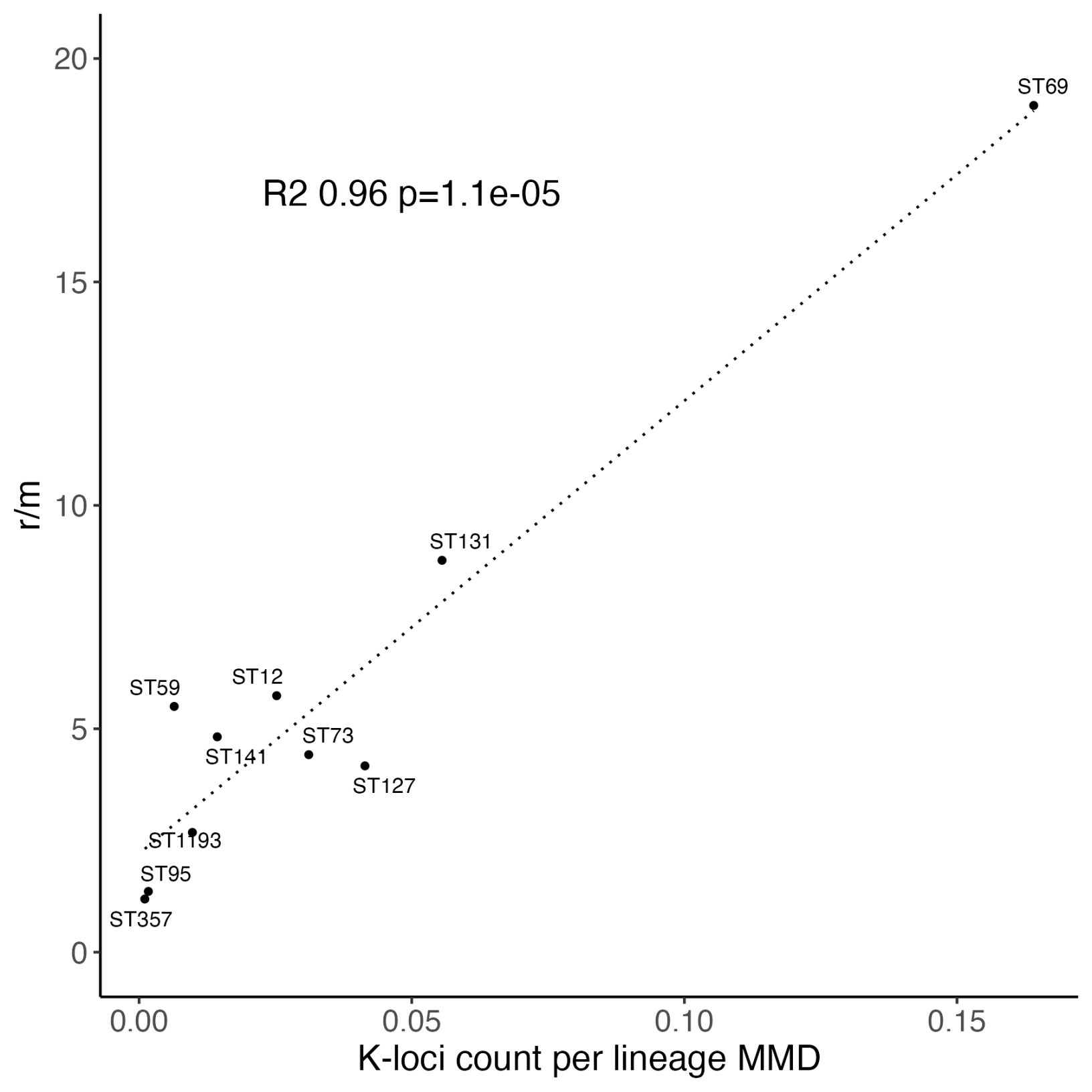
696

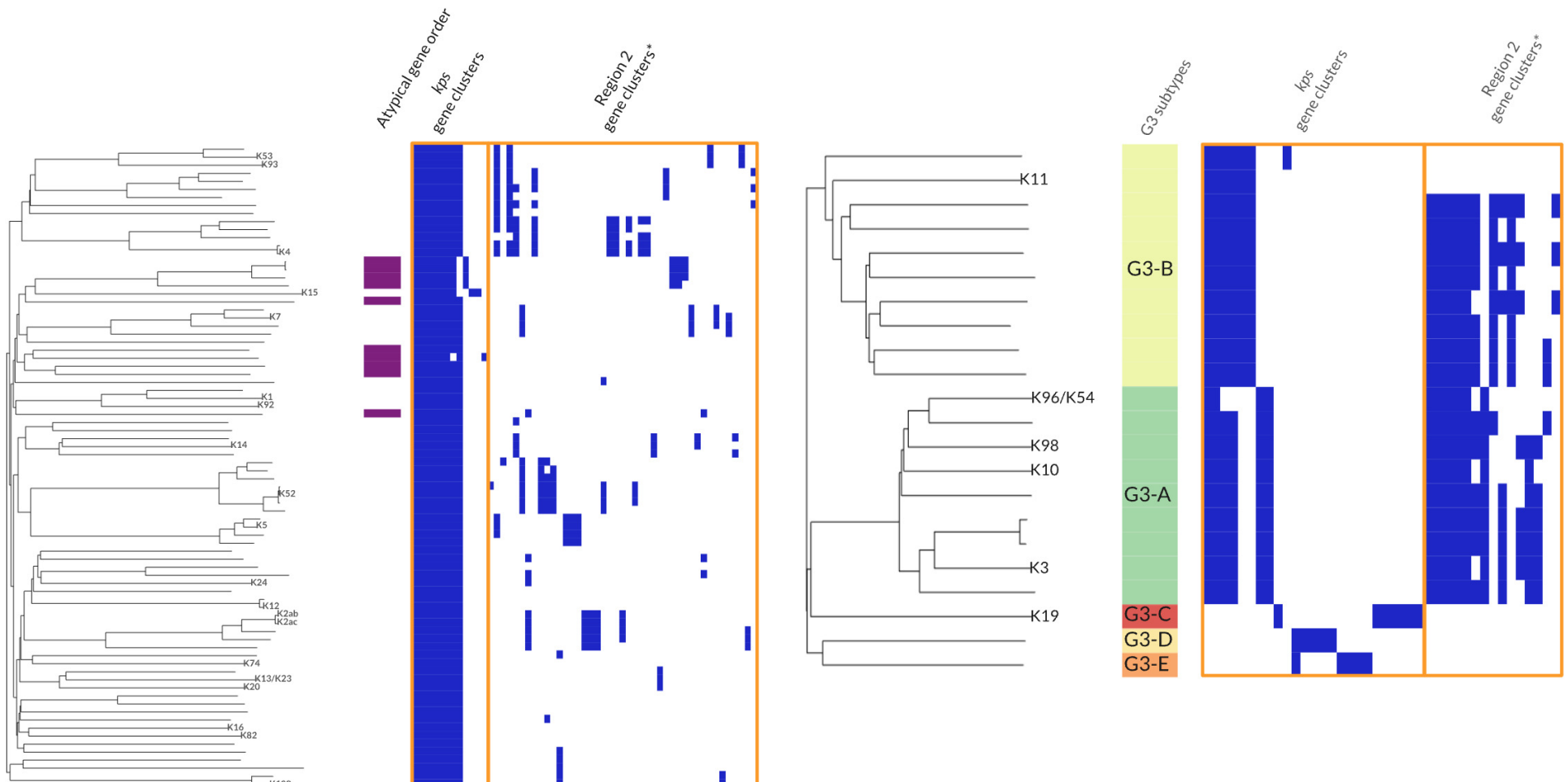



C1
C2
B
A
O25:H4
O16:H5
1996 [1993-1999]
1999 [1995-2003]
1988 [1983-1992]
1989 [1985-1993]
KL2_K100
KL3_K2ab
KL6_K5
KL8A1_K1
KL9_K20
KL16_K14
KL30
KL40_K15
KL42
KL44
KL47_K16
KL53_K2-like
KL103_K100-like
KL112
KL126_K14-like
KL137_K14-like
KL140_K7-like

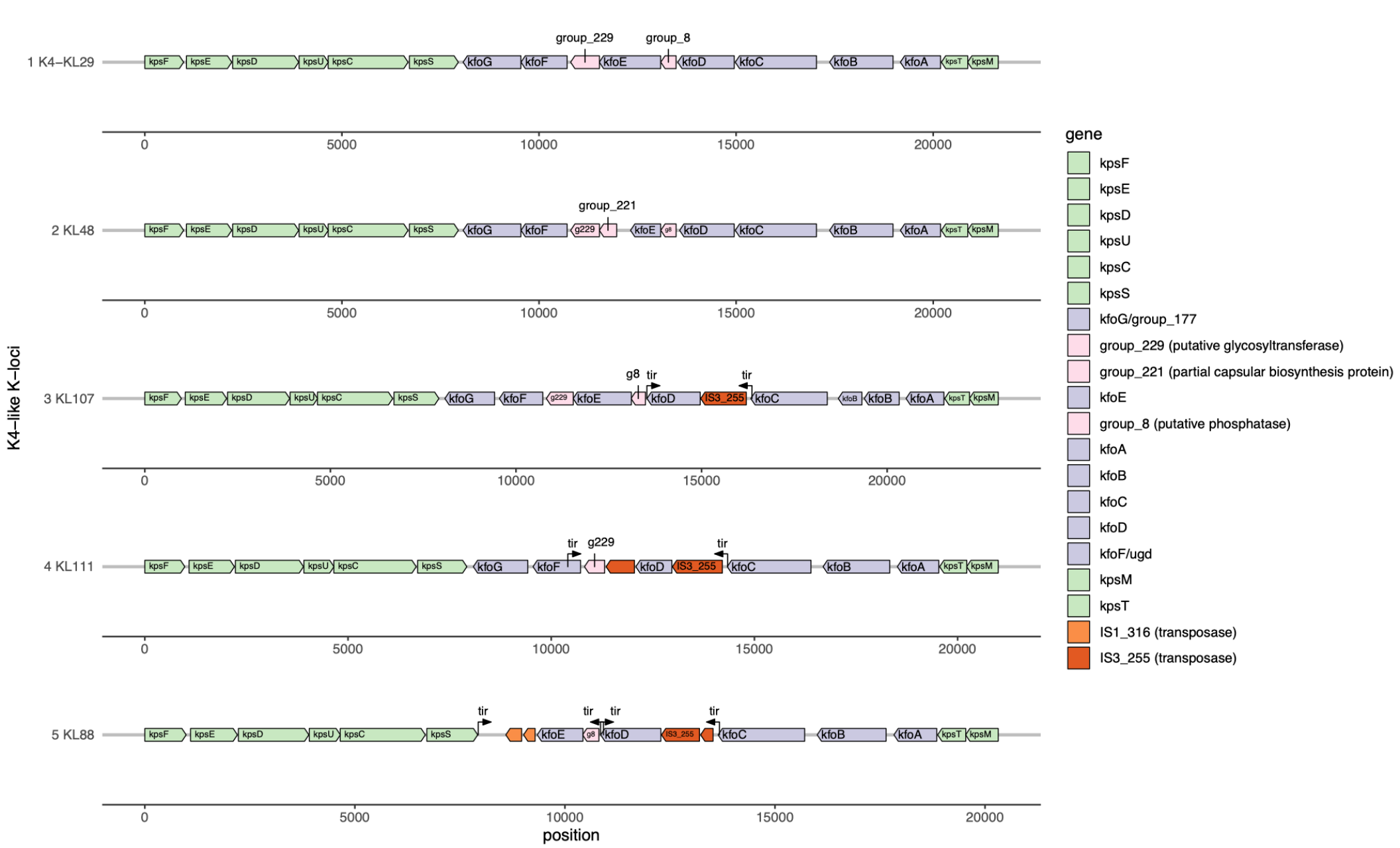

2001 [1998-2003]
2000 [1997-2002]
1996 [1992-2000]
1986 [1983-1989]
1987 [1984-1990]
1996 [1993-1998]
1970 [1963-1976]
Major K-types
KL6_K5
KL36_K54-K96
KL13
KL189_K52-like
Group membership
G2
G2-atypical
G3-A
G3-B


Simpsons Diversity Index
0.75
0.50
0.25
0.00
K-types per lineage
Lineages per K-type


R2 0.96 p=1.1e-05
r/m
K-loci count per lineage MMD
0
5
10
15
20
0.00
0.05
0.10
0.15
ST69
ST131
ST12
ST59
ST73
ST141
ST127
ST1193
ST95
ST357


Atypical gene order
kps gene clusters
Region 2 gene clusters*
G3 subtypes
kps gene clusters
Region 2 gene clusters*
K53
K93
K4
K15
K7
K1
K92
K14
K52
K5
K24
K12
K2ab
K2ac
K74
K13/K23
K20
K16
K82
K100
K11
K96/K54
K98
K10
K3
K19
G3-B
G3-A
G3-C
G3-D
G3-E


1 K5 KL6
2 KL177
3 KL91
4 KL6−1
5 KL6−2
6 KL6−3
7 KL6−4
K5−like K−loci
position
0
5000
10000
15000
kpsF
kpsE
kpsD
kpsU
kpsC
kpsS
kfiD
kfiC
kfiB
kfiA
kpsT
kpsM
group_145
tir
IS5_222
gene
kpsF
kpsE
kpsD
kpsU
kpsC
kpsS
group_145 (putative gylcosyl−transferase)
kfiA
kfiB
kfiC
kfiD/ugd
kpsM
kpsT
partial CTX−M−1
group_174 (putative transposase)
group_240 (putative transposase)
IS1_316 (transposase)
IS3_422 (transposase)
IS5_222 (transposase)


1 K4–KL29
2 KL48
3 KL107
4 KL111
5 KL88
K4–like K–loci
position
0
5000
10000
15000
20000
kpsF
kpsE
kpsD
kpsU
kpsC
kpsS
kfoG
kfoF
kfoE
kfoD
kfoC
kfoB
kfoA
kpsT
kpsM
group_229
group_8
group_221
g229
g8
tir
IS3_255
gene
kpsF
kpsE
kpsD
kpsU
kpsC
kpsS
kfoG/group_177
group_229 (putative glycosyltransferase)
group_221 (partial capsular biosynthesis protein)
kfoE
group_8 (putative phosphatase)
kfoA
kfoB
kfoC
kfoD
kfoF/ugd
kpsM
kpsT
IS1_316 (transposase)
IS3_255 (transposase)