Group 2 and 3 ABC-transporter-dependent capsular K-loci contribute significantly to variation in the estimated 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], Pål J. Johnsen[10], Ørjan Samuelsen[11], Nicholas R. Thomson[4,12], 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 Division of Population Health and Genomics, School of Medicine, University of Dundee, 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 Doherty Institute for Infection and Immunity, Melbourne, Australia
10 Department of Pharmacy, Faculty of Health Sciences, UiT The Arctic University of Norway, Tromsø, Norway
11 Norwegian National Advisory Unit on Detection of Antimicrobial Resistance, Department of Microbiology and Infection Control, University Hospital of North Norway, Tromsø, Norway
12 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.). This work was supported, in part, by the Bill & Melinda Gates Foundation INV-025304. The conclusions and opinions expressed in this work are those of the author(s) alone and shall not be attributed to the Foundation. Under the grant conditions of the Foundation, a Creative Commons Attribution 4.0 License has already been assigned to the Author Accepted Manuscript version that might arise from this submission. Please note works submitted as a preprint have not undergone a peer review process.

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), however, phenotypic capsular (K) typing has fallen out of use. Recent systematic genomic surveys have provided an opportunity to compare colonisation and bloodstream infection frequencies to quantify the relative invasive potential of capsules across different ExPEC lineages. We assembled and curated an *in silico* capsular locus typing database for group 2 and group 3 K-loci from >20,000 genomes and applied it to carriage and disease cohorts to investigate capsule epidemiology. Several widely circulating K-loci have estimated odds ratios of >7 for being found in bloodstream infections compared with carriage. The relative invasive potential differed markedly between lineages, and notably, even between subclades of the multidrug-resistant ST131 clone. We determined that insertion sequence elements contribute to the evolutionary dynamics of group 2 and group 3 K-loci by importing new capsular genes. The level of capsule diversity was positively correlated with a higher rate of homologous recombination per lineage, likely a consequence of faster adaptation of their antigenic repertoire. Our investigation highlights several capsules and lineages that contribute disproportionately to invasive ExPEC disease, some of which are also associated with high levels of AMR. These results have generated new insights into capsule epidemiology and 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 impact DNA exchange, phage predation, and plasmid uptake, all of which have evolutionary underpinnings. Bacterial capsular polysaccharide antigens are often well-studied and used as a central component of 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 found 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 translational research to develop new antimicrobials, phage therapy and vaccines. Although some *E. coli* capsular (K)-types mimic vertebrate glycoconjugates, hindering vaccine development,[3,4] their true prevalence in contemporary disease has not been determined in large, unbiased disease cohorts. There have also been early *in vivo* successes demonstrating the potential of phage therapy in invasive infections[5].

Over 80 *E. coli* K-types can be identified with traditional phenotypic capsular typing using a set of antisera. However, this method is labour-intensive and no longer in general use. Unlike other important *E. coli* antigens, such as O and H-types[6], no genotypic method exists for typing the capsular loci encoding the K-types, and a lack of paired capsular phenotype and genotype data has hindered its development. *E. coli* capsular K-loci have been classified into four groups based on whether they are *wzy*-dependent (groups 1 and 4) or *kps*-ABC-transporter dependent (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 the four 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 less common in invasive disease[9]. The ExPEC-associated G2 and group 3 (G3) K-loci have conserved capsular polysaccharide (*kps)* genes found in Region 1 involved in assembly and export, and Region 3 involved in transport. These genes flank the capsular-determining genes

93 in Region 2 and encode, for example, glycosyl transferases and sugar synthetases. G2 and G3
94 K-loci are found near *serA* in the chromosome; however, recently, a subgroup of G3 K-loci have
95 also been observed on plasmids[10]. The conserved gene content and organisation of G2 and G3
96 loci make them an ideal candidate for *in silico* typing. *In silico* methods for the broader
97 assessment of capsule loci across bacterial families and species have already increased our
98 understanding of these important antigens, even in the absence of comprehensive phenotypic
99 data, by taking a capsular gene-based approach [11–13]. A G2 and G3 *E. coli* typing scheme will
100 complement the considerable previous work on phenotypes and structures to stimulate further
101 study of *E. coli* K-types[14], as has been demonstrated for typing schemes in other species[13,15,16].
102
103 Systematically collected data on capsule prevalence can allow the estimation of their relative
104 invasive potential. This has been instrumental in identifying those capsules with the greatest
105 potential to cause invasive disease in multiple species, especially the pneumococcus, providing
106 a ranking for prioritising inclusion in vaccines and modelling optimal polyvalent vaccine
107 formulations.[17–22] This epidemiological approach leverages measured frequencies in
108 colonisation and observed disease cases to calculate an odds ratio that indirectly captures the
109 relationship between exposure and the subsequent manifestation of disease cases in a
110 population[23,24], which has been further employed in the estimation of R0 for *E. coli* strains[25].
111 Here, we catalogue the substantial diversity of *E. coli* G2 and G3 capsular loci, present their
112 prevalence in BSIs, UTIs and carriage cohorts, and quantify the relative invasive potential
113 associated with them. Furthermore, we investigate the links between capsule diversity and DNA
114 exchange in homologous recombination, insertion sequence elements and plasmids. Our
115 results highlight key K-loci encoding capsular antigens that require renewed scrutiny and
116 attention in translational efforts to reduce the future public health burden of *E. coli* invasive
117 disease.

Results

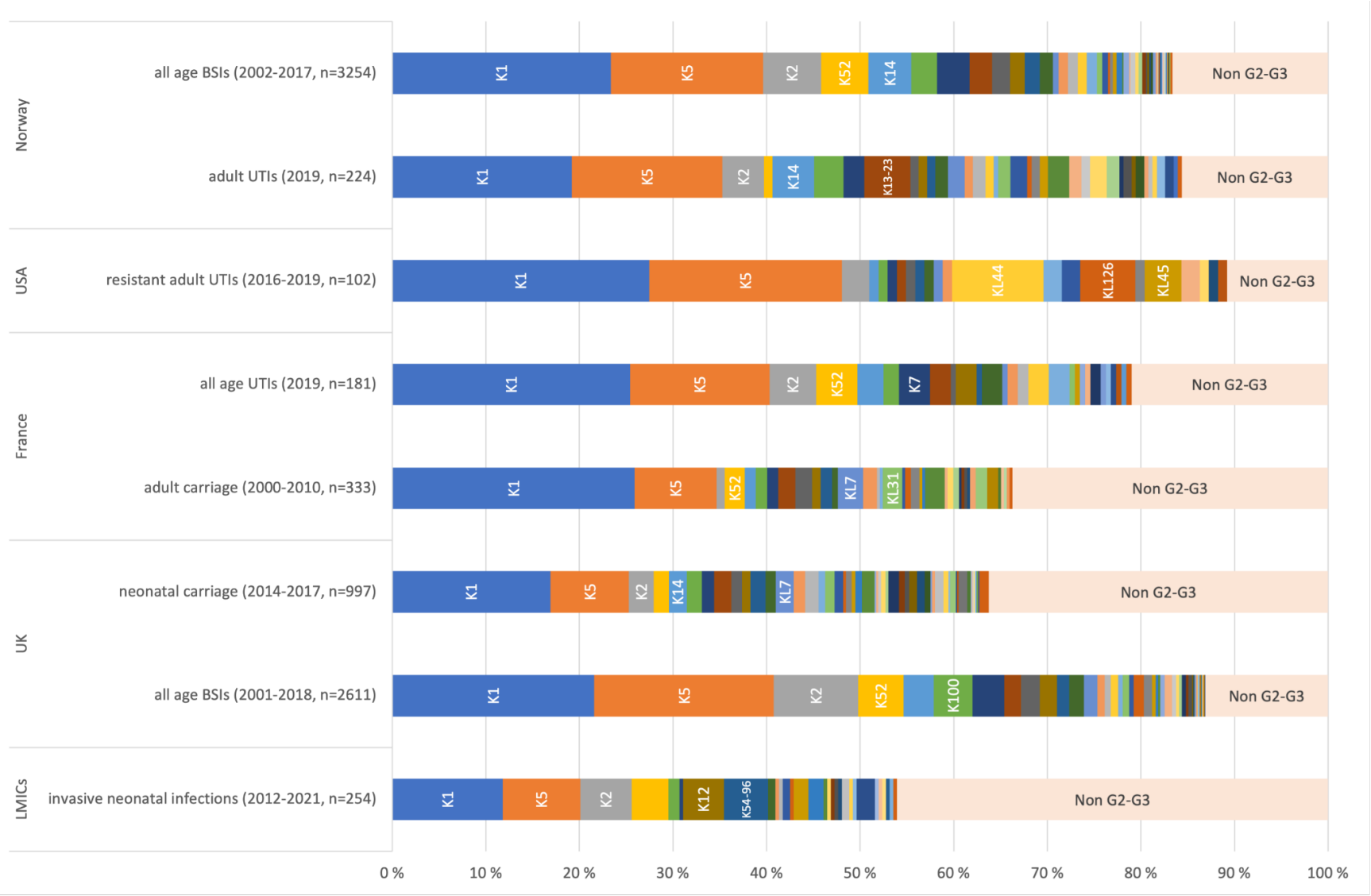
By investigating a collection of >20,000 high-quality *E. coli* genomes, we observed 90 unique G2 (n=68) and G3 (n=22) ABC-transporter-dependent capsular (K) loci, based on capsular locus gene presence-absence patterns. Eighty percent of the known G2 and G3 phenotypic K-types are represented here by phenotype-genotype paired data, yet known phenotypes only represent 34% of the 90 K-loci observed to have unique capsular gene sets. Furthermore, for 11 K-loci without a known phenotype that were observed in Norwegian BSIs we were able to confirm a K+ phenotype with a positive precipitate with Cetavlon and that they were negative for known K antigens representing putatively novel K-types. The 90 curated K-loci were used to create a G2 and G3 *in silico E. coli* K-typing reference database available at https://github.com/rgladstone/EC-K-typing, compatible with the species agnostic bacterial capsular loci typing tool Kaptive[26]. We K-locus typed assemblies from several systematic studies of infections, including: 1) large longitudinal genomic surveys of bloodstream infections (BSIs) from Norway[27] and the UK[24,28,29], 2) two urinary tract infection (UTI) surveys from Norway and France[30], 3) UTIs with specific resistance profiles from the USA[31], and 4) two collections of invasive isolates from neonates collected in low and middle-income countries (LMICs).[32,33] In addition, K-locus typing of assemblies from infant-mother gut metagenomic surveys in the UK[23,34] and from traditional culture picks (France[35]) provided information on K-types in asymptomatic colonisation. As over 65% of K-loci did not pair with the known phenotypic K-types described in the previous century, phenotypes can be added retrospectively to each K-locus in the public database as data become available. The KL designations reported here are determined algorithmically and consequently do not mirror the phenotypic K-type numbering. Throughout this manuscript, we will report the inferred phenotypic K-type (e.g, K1) for a given K-locus (KL) genotype when available. Where no phenotypic K-type currently exists for a K-locus, a numerical KL designation (e.g. KL5) will be used instead.

K-locus epidemiology

European BSIs[24,27–29] are predominantly caused by strains with a G2 or G3 capsule (81.3%); among these, G2 dominated, and only a minority had G3 loci (3.5%). The vast majority of BSI isolates in phylogroups B2 (94.2%), D (84.1%), and F (94.6%) had G2 K-loci, and isolates with G3 K-loci were mostly found in phylogroups B2, D and A. The top five K-types in European BSIs were K1, K5, K52, K2, and K14 (Fig. 1). These common K-types accounted for over 50% of

149 BSIs (UK[24,28,29], Norway[27]) and UTIs (Norway[30], France[30]). Whilst G2 and G3 K-loci were far less
150 common in asymptomatic colonisation in Europe (55.4%)[23,34,35], K1 and K5 were still amongst
151 the most common colonising G2-G3 K-types. Notably, G2 and G3 loci were much less common
152 (53.9%) in invasive neonatal infections in LMICs[32,33] (Fig. 1), a stark contrast to the 95% of UK
153 BSIs in children <1 year old that were G2 or G3.

154 **Figure 1. K-type prevalence across different collections.** Published genomic collections were K-typed [23,24,27–35]. The top 5 K-types
155 in each collection are annotated. Isolates that were *kps* negative and do not contain group 2 or 3 K-loci are classified as non-G2-G3.

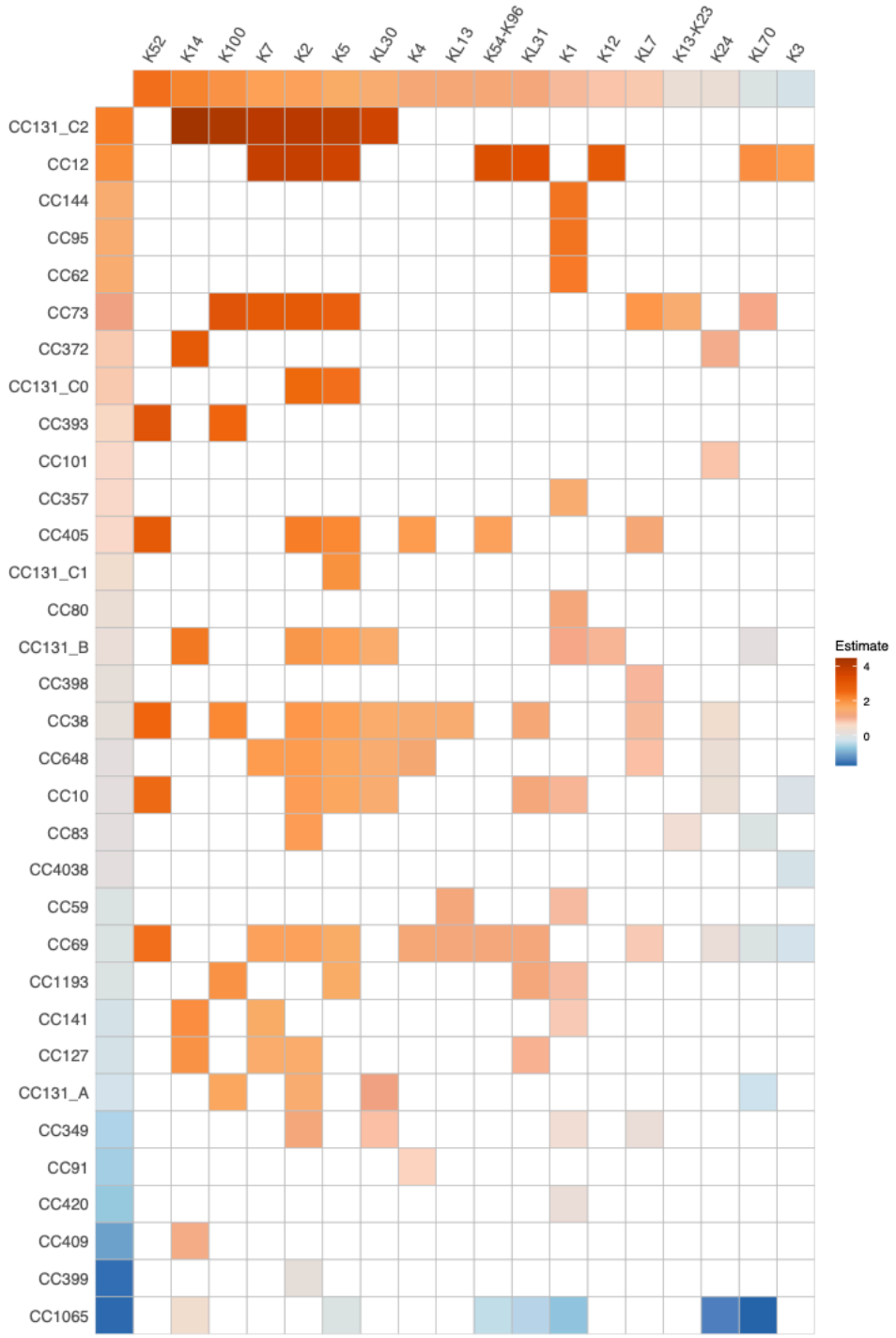


156

157 ## The estimated relative invasive potential of K-types

158 Using a mixed-effect model for UK isolates from carriage and BSIs, we estimated the
159 contributions of each K-locus and lineage to the relative invasive potential (Figure 2). We
160 determined that K52 had the highest invasive potential (OR=11.6, 95% CI: (5.3-25.4), p<0.0001)
161 among the 18 most common K-loci in comparison to the average untypeable isolate (*i.e.*
162 genomes without the G2-G3 *kps* genes). K14 and K100 had the second and third highest
163 invasive potentials, respectively (OR=8.9 [3.7-21.7], OR=7.2 [2.6-20.0]). The lineage-specific
164 odds of K100 being found in BSIs were estimated to vary within CC131 between 5.42 (CC131
165 clade A) and 69.80 (CC131 clade C2). There was also substantial variation in the lineage
166 random effect estimates between CC131 subclades B and C despite sharing the same O-H
167 type. The odds ratios for the widely studied K-types K1, K2, and K5 were ranked 12th, 5th, and
168 6th, respectively. These three types nevertheless had an OR significantly greater than one,
169 demonstrating their greater invasive potential compared to untypeables. Conversely, K3 had the
170 lowest invasive potential of 0.82 [0.2-2.7]. These findings suggest considerable variability in the
171 propensity of K-type and lineage to cause infection, similar to previously established
172 epidemiological findings for the pneumococcus.
173
174 We considered the burden of *E. coli* BSI by age and sex due to immunosenescence in the
175 elderly and a higher burden of UTIs in younger women. Patient age group and sex data were
176 available for n=1802 UK BSI isolates, we observed in an exploratory analysis (Supplementary
177 Figure 1) that K1 was overrepresented in the <1 year (p=0.0090) and 40-49 years (p=0.0072)
178 age groups compared to the collection as a whole, although after adjusting for multiple testing
179 on all common age-K-type combinations these were no longer significant. We observed some
180 sex differences in the proportion of BSIs. CC95-K1, but not K1 overall, was significantly
181 enriched in females (adjusted p=0.013), and K52, predominantly found in the uropathogenic
182 CC69, was observed in females in every decade of life but only occurred in older men aged
183 ≥50, making it significantly overrepresented in females (adjusted p<0.0001), presumably due to
184 disease progression from urinary tract infections. This may drive the general observation that
185 the proportion of BSIs in the 1-59 age group was significantly higher for females (60%) than
186 males (40%, p<0.0001), but was approximately even across the entire life course (females 52%,
187 males 48%).
188

Figure 2. 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 manifestation (infection/carriage) as a binary outcome, K-loci (KL) as a fixed effect and lineages as random effects. Lineages are denoted as Clonal Complexes (CC) derived from the representative sequence type for each popPUNK lineage. Red tones are associated with higher, and blue tones with lower estimates of invasive potential. The reference category for K-loci was untypeable (*i.e.* genomes without G2-G3 *kps* genes). The rows and columns are sorted by the relative invasive potential estimates from highest to lowest. The CC131 lineage is split into its major clades (A, B, C1 and C2). The corresponding K-type, if known, is presented for a given K-locus.



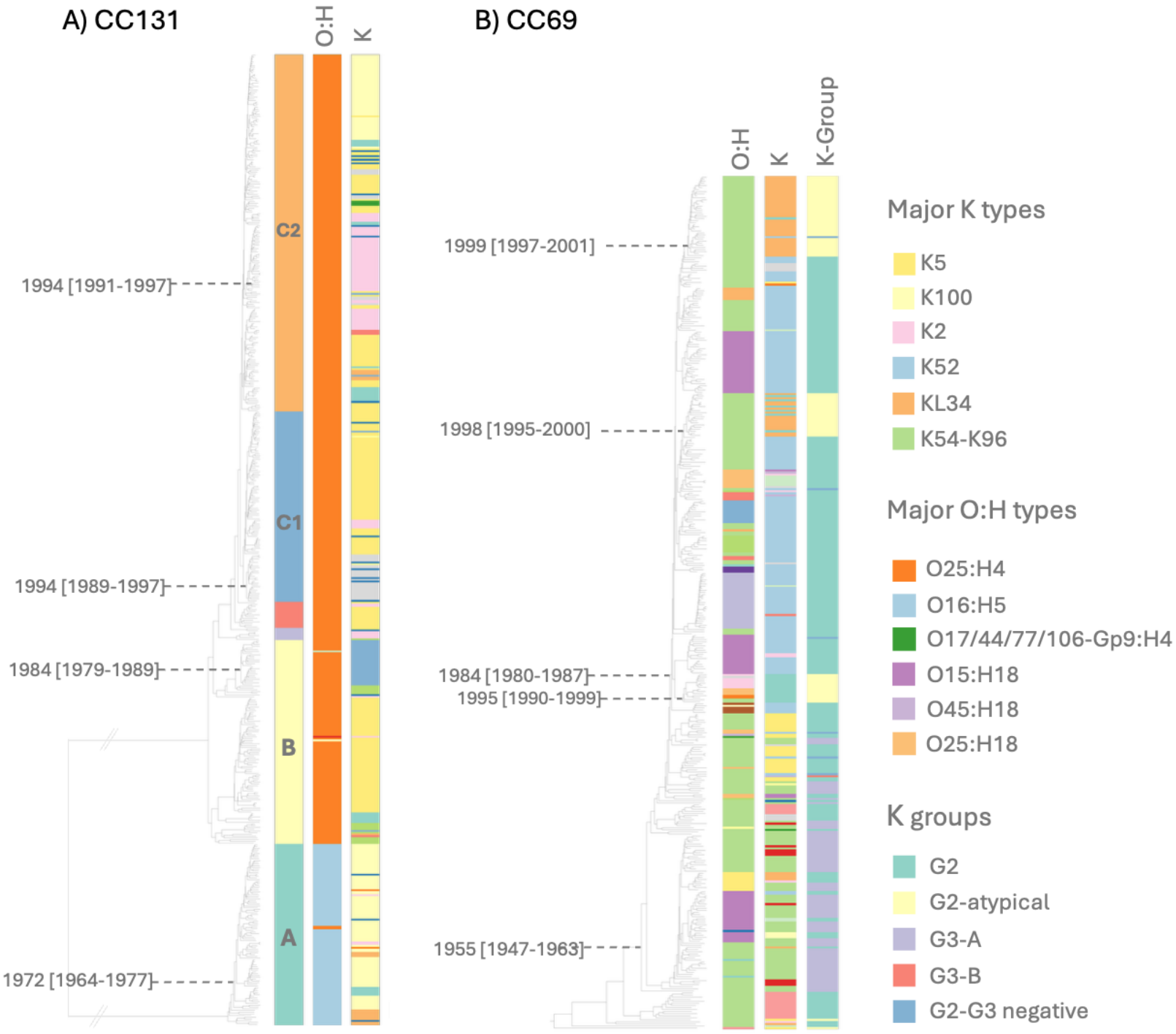
211 ## Group 2 K-loci association with *E.coli* and the ExPEC pathotype

212 *E. coli* G2 Region 1 and 2 *kps* genes are not often found in other species. In a published
213 collection of 661,000 assemblies[36], only 11,737 were positive for the highly conserved and
214 essential *kpsF* gene (1kb) at 90% kmer identity. These were nearly exclusively *E. coli* (99.0%,
215 11623/11737). Outside *E. coli*, *kpsF* was most commonly observed in *Salmonella enterica*
216 *subspecies enterica* (n=69), *Klebsiella pneumoniae* (n=17) and *Staphylococcus aureus* (n=11).
217 Using a published pangenome of ~7,500 *E. coli*[37], we assessed the pathotype association for
218 G2. ExPECs accounted for 85.8% (1590/1853) of the *kpsF*-positive isolates with pathotype
219 information, and non-ExPEC pathotypes accounted for 97.5% (3647/3741) of the *kpsF*-negative
220 isolates. Shiga toxin-producing *E. coli* (STEC) accounted for the majority of non-ExPEC
221 pathotypes positive for *kpsF* (90.5%, 238/263), and these STEC were found in multiple
222 lineages, including sequence types (ST)442, ST10, ST25, ST504 and ST675.

223 ## K-loci in major MDR lineages

224 The globally disseminated multidrug-resistant (MDR) lineage corresponding to the clonal
225 complex (CC)131 is known to have two dominant O-H types. However, we observed at least 17
226 different K-loci that have been introduced into CC131, highlighting the rapid capsule
227 diversification in this lineage (Fig. 3A). The ancestral K-type for the B/C clades is K5
228 (n=207/453) and K100 in clade A (n=80/104). Other K-loci found with considerable prevalence
229 in CC131 BSIs include: K2 (n=66/557), K14 (n=30/557) and KL44 (n=20/557), all of which were
230 estimated to have been acquired between 1980 and 2000. In total, K2 was acquired eight times
231 independently across the lineage. Interestingly, KL44 in CC131 is an example of a K-locus with
232 an atypical architecture, where Region 1 is immediately followed by Region 3, and Region 2 is
233 at the end of the locus. Its acquisition was often followed by subsequent fragmentation and
234 complete G2 loss events.
235
236 In another globally distributed MDR ExPEC lineage, CC69, we observed 11 O-types, 9 H-types
237 and 23 K-types. O diversity in CC69 was higher than in CC131 (Simpson's Diversity Index (SDI)
238 CC69: 0.66, CC131: 0.39), but is still less than CC69 K-locus diversity (SDI: 0.76). K-loci
239 belonging to G2, atypical-G2 and G3 are all found in this lineage (Fig. 3B). The most common
240 K-locus was K52 (39.5%, 163/413), which had the highest estimated invasive potential (Fig. 2).
241 Switches from K52 to the unrelated atypical KL13 occurred three separate times between 1992
242 and 2003.

Figure 3. Diversity of K-loci across the clades of CC131 and CC69 from Norwegian and UK BSIs. Major capsule switch events are indicated with a dashed line from the relevant nodes with the time to the most recent common ancestor and confidence intervals. A) The four major CC131 clades are labelled in column one. The major O-H types are displayed in column 2 and the K-type in column 3. B). The major CC69 O-H types are colour-coded in column 1, K-type in column 2 and the K-group in column 3. The adjacent key displays the major O-H and K-types. Where the K-phenotype is unknown, the K-locus (KL) is presented. Fully annotated interactive phylogenies are available at https://microreact.org/project/cc131-bsac-norm and https://microreact.org/project/cc69-bsac-norm

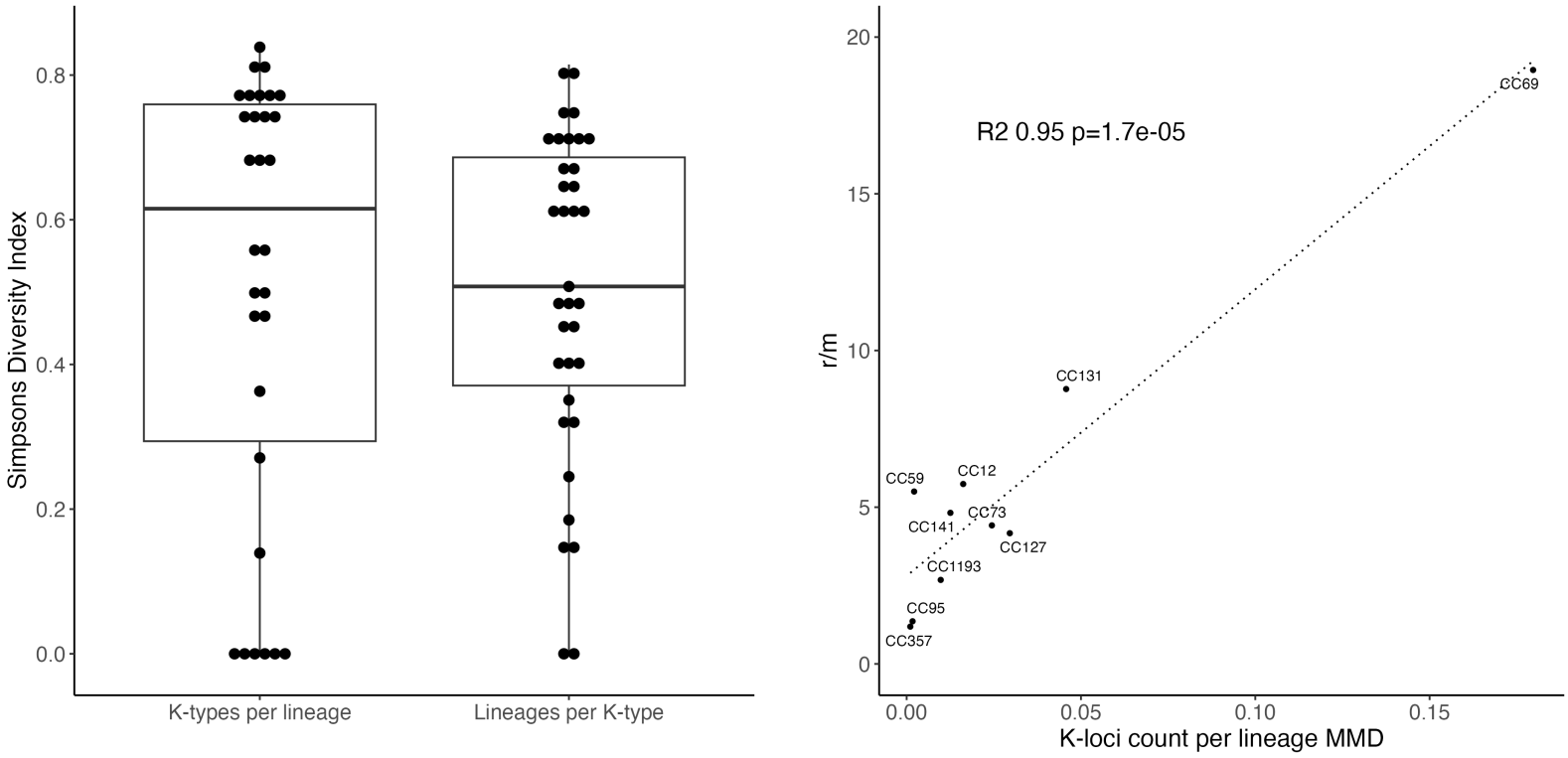


K-locus evolution

Whilst K diversity measured by Simpson's Diversity index was higher than O or H diversity in over half of the lineages observed in BSIs, lineages varied greatly in their K-type diversity (Fig. 4A), suggesting differing evolutionary processes. Indeed, the number of K-loci in a lineage was correlated with the r/m (recombination to mutation ratio) of that lineage, even when controlling for the lineage diversity by using the recombination-free median mutational pairwise distance (MMD, R2=0.95, p<0.0001, Fig. 4B). There was no correlation between r/m and MMD (R2=-0.35, p=0.3). The trend remained significant with the exclusion of CC69 (R2=0.75, p=0.02), but not after excluding the two most recombinogenic lineages CC69 and CC131, though the association remained positive (R2=0.47, p=0.2). In the top four most common lineages in BSIs (CC73, CC95, CC69 and CC131), the *kps*-locus is a clear recombination hotspot, with the exception of CC95, which only expresses K1. Highly recombinogenic lineages could contribute to the discovered changes in the K-locus genetic architecture, including the observed atypical K-loci.

Figure 4. A) Simpson's diversity index (1-D) 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)

A) **B)**



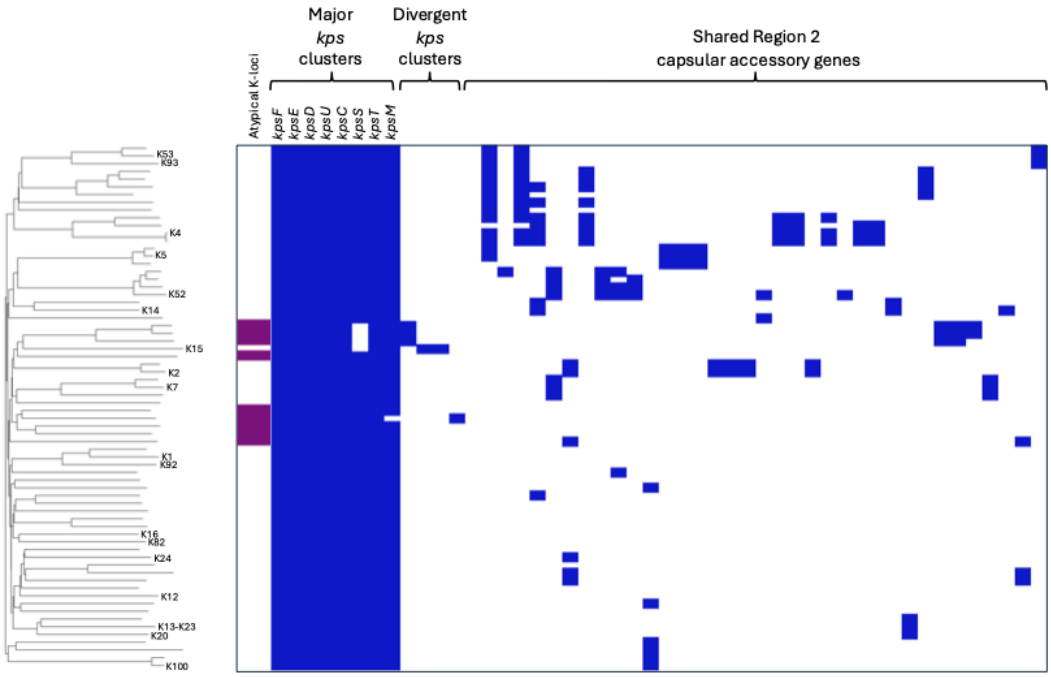
K-loci structural variation

Known K-phenotypes account for only 34% of the numerous K-loci identified in this study, with these phenotypes sporadically distributed across the K-loci phylogeny (Fig. 5). Many K-loci correspond to deep ancestral branches that likely represent distinct K-types. In total, there were 229 region 2 capsular gene clusters (>70% amino acid AA sequence identity within clusters), of which 48% were annotated as “hypothetical protein”, demonstrating considerable unexplored capsular-determining gene variation. The eight *kps* genes in G2 K-loci formed one major gene cluster each. Only six of the G2 K-loci contained a *kps* gene in a separate divergent cluster for at least one of the *kpsM, kpsT, kpsS* and/or *kpsC* genes, which are important in forming the biosynthesis-export complex[38]. Nine K-loci feature an atypical G2 locus structure that has not been previously reported, where region 2 is outside of regions 1 and 3 (Supplementary Figure 3). Although the atypical K-locus organisations are relatively rare in BSIs (<3.5%), they were spread across the species phylogeny and found in phylogroups A, B2, D and F. There were also sizable clusters of atypical loci in major *E. coli* BSI lineages: CC69 (KL13, n=84/449), CC131 (KL44, n=25/631), CC73 (KL44 n=20/981), CC59 (KL13 n=14/93). A Norwegian BSI isolate with the atypical K-locus KL73 was confirmed to be K+ but did not react with known K-antisera.

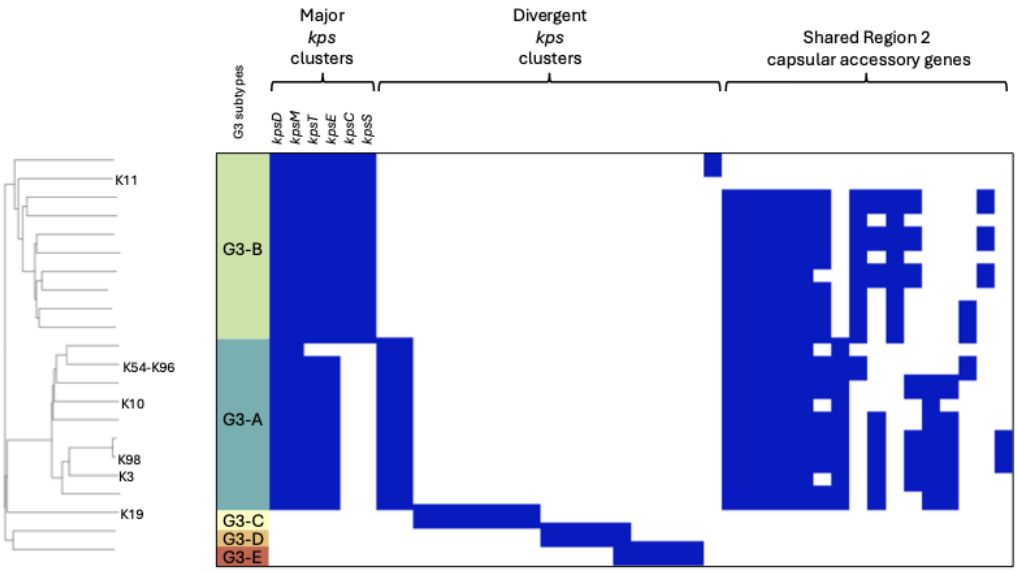
Unlike in G2, the diverse *kps* genes in G3 form two major gene clusters delineating the G3-A and G3-B subtypes (Fig. 5B). No G3 *kps* genes clustered with G2 *kps* genes. G3-A and most of G3-B share the five region 2 gene clusters *wsaE* and *rmlBDAC*. Meanwhile, K11, KL66, K19, KL65 and KL75 loci also have divergent and unique region 2 gene sets. 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 and shares no *kps* gene clusters with G2, KL66 has previously been classified as G3, and G3 has been suggested to be divergent G2 K-loci elsewhere[10,39]. Therefore, we denote these three loci as G3-C to G3-E.

Figure 5. Neighbour-joining trees of G2 (5A) and G3 (5B) K-loci based on core and accessory sequence presence/absence, annotated with known K-phenotypes at the tips. The first metadata column denotes the K-loci with atypical gene organisation (5A) and subtypes of G3 (5B). The gene cluster presence (blue) and absence (white) used a 70% AA identity threshold. Three highlighted blocks show the dominant *kps* gene clusters in the order they are most often observed in the K-locus, divergent kps gene clusters, and the most common region 2 gene clusters ordered by the number of KL they were observed. *only gene-clusters present in >2 K-loci are presented. These figures can be viewed interactively with full taxa and gene sets and annotation on phandango.net using the input files provided on https://github.com/rgladstone/EC-K-typing/tree/main/phandango

A)



B)



341

342 ## K-loci diversification and gene mobility

343 The genomic survey of G2 and G3 K-loci allowed us to interrogate the mechanisms that may
344 have generated this considerable diversity. The proportion of unique K-loci sequences (at least
345 1bp difference) with one or more insertion sequences (IS) was 28.2% (1411/4996). While K1
346 was only observed with an IS once (n=1/692), K5 always had at least one IS (n=706). IS1 and
347 IS3 were the most common of the ten IS families observed in the K-loci collection. Importantly,
348 IS were observed overlapping with capsular coding sequences (CDS) in 17 K-loci and in nine of
349 these, the IS carried a putative capsule gene in Region 2 as cargo.
350
351 The K-locus KL6 (K-type K5) is one of three closely related K-loci based on gene presence-
352 absence patterns (after exclusion of the IS elements). The KL177 locus differs from KL6 by the
353 loss of *kfiD*; it was rare and only observed in carriage (n=3). In KL91, an IS1 316 element had a
354 putative glycosyl-transferase family two protein (group 145) and a truncated *kfiC* within its
355 terminal inverted repeats (TIRs, Supplementary Figure 2A). This locus was observed four times
356 in BSIs and five times in carriage. An IS variant of the KL6 locus (KL6-1) seen in a single ST131
357 isolate had an additional putative transposase (group 240) and a fragment of the extended-
358 spectrum β-lactamase gene bla_{CTX-M} near the IS3 remnant seen in all the K5-like loci. The
359 bla_{CTX-M} and transposase gene were observed together in multiple plasmid types within CC131
360 (100% identity), suggesting that both genes may have moved into this K-locus from a plasmid
361 via homologous recombination with the existing IS in these K-loci. There were three more
362 examples of IS in a KL6 background without any alteration to the K-locus-related gene content,
363 each observed once in the collection used to populate the database, suggesting IS movement in
364 and out of KL6 is happening frequently.
365
366 There were four different K-loci related to K4 (KL29) with differing gene presence-absence
367 patterns when IS genes were excluded (Supplementary Figure 2B). The *kfoA, kfoB, kfoC,* and
368 *kfoD* genes are present in all five, and kfoF (gene-cluster name *udg*) and *kfoE* were only absent
369 in KL88 and KL111, respectively. KL29-1 is an IS variant of KL29 where *kfoB* is split into two
370 fragments and kfoD is contained within an IS element. These IS elements were all observed
371 between *kpsS* and *kfoC* and likely contributed to the different patterns of capsular-specific
372 genes within region 2 for this cluster of related K-loci. K4 (KL29) was the most common of these

373 loci in BSIs (n=54). Whilst IS variants of K5-like loci and K4-like loci are rare, their existence
374 gives us insight into how capsular diversification and gene exchange may be driven by IS
375 moving between K-loci. There were no IS observed in the atypical loci, suggesting other
376 mechanisms are also at play.
377
378 The median distance of the K-locus from serA in the chromosome was confirmed to be
379 144,833bp (IQR 38,022bp). Recently, G3-B K-loci have been observed in plasmids, which could
380 act as a vehicle for capsular mobility[10]. As G3-B are rare in BSIs (n=11), we observed just two
381 Norwegian isolates with G3-B K11 in multireplicon plasmids (IncFIA-IncFIB-IncY and IncFIA-
382 IncY), and not in the chromosome. One of the isolates had K11 mobilised on a plasmid with
383 multiple resistance genes: *bla*$_{\text{TEM-1}}$ *dfrA14, mph(A), sul2, aph(3")-Ib and aph(6)-Id*. Additionally,
384 a single *E. coli* isolate from Oxfordshire (n=1/549) was typed as G3-B KL139 on the plasmid
385 and not in the chromosome.

Discussion

Extensive phenotypic and structural data generated over several decades have demonstrated considerable diversity in the *E. coli* capsular antigens using clinical isolates. Nevertheless, systematic cataloguing of capsular genetic diversity in contemporary disease is essential to further our understanding of the epidemiology of K-types, in particular with the increasing infection 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 types. Furthermore, this diversity has major clinical implications, with observed differences in K-type prevalence by isolation source and geographical location, in addition to variation in the estimated invasive potential of K-types across different genetic backgrounds. We determined that G2 and G3 capsules are present in the majority of ExPEC infections across diverse settings. Specifically, K1 and K5 are the most common causes of UTIs and BSIs. Although these two types are also the most common G2/G3 K-types in LMIC neonatal disease, a significant fraction of these infections are caused by non-G2-G3, unlike in high-resource settings. These particular findings have implications for the development of preventive strategies. Of note, *in silico* typing of the *wzy*-dependent group 1 and group 4 capsules is a critical remaining knowledge gap to understand ExPEC disease in resource-limited settings. There is a paucity of published ExPEC genomic data from systematic epidemiological studies conducted in resource-limited regions of the world that sequencing efforts have neglected. Although at present the existing data represent only a couple of LMICs, our results clearly suggest that capsular epidemiology differs significantly from high-resource settings for at least vulnerable neonates, suggesting prioritisation of future sampling efforts to generate region-specific representative data similar to that for pneumococcal disease [40–42]. However, by screening a published database of assemblies created from all data deposited in the sequence read archive (SRA) pre-2018[36], we were able to represent K-loci in the database from 69 countries, including 31 countries from Africa, Asia and Oceania. The G2-G3 K-loci repertoire observed in LMICs was well represented by the G2-G3 database, even if the exact prevalence of G2-G3 K-types varies.

The role of capsules in pathogenesis has long been appreciated across bacterial species[1,43]. Whilst the study of *E. coli* capsules has stalled more recently, previous research efforts have already experimentally demonstrated the virulence associated with most of the common *E. coli* capsular types (e.g. K1, K2, K5, K52, K92 and K100) using either human immunological assays

419 or murine models[44–49]. Estimating the relative invasive potential allowed us to put these
420 experimental findings into a wider context by ranking the capsular types in terms of their
421 propensity to be found in BSIs, given the exposure frequency in asymptomatic colonisation. The
422 results suggest that the high frequency of the widely studied K1 and K5 types in disease is, at
423 least in part, a product of greater opportunistic spillover into vulnerable hosts. However, our
424 estimates additionally indicate that there are multiple K-types that have a greater ability to cause
425 an infection, given their ranked invasiveness. For example, K52 and K14, which are found in
426 multiple lineages, including the global MDR lineages CC69 and CC131, had the highest
427 estimates of invasiveness and subsequently ranked 3rd and 4th in BSI prevalence despite
428 rarely being seen in carriage samples in the UK and France (<2.5%). K1 has previously been
429 reported to be found across multiple lineages[46], and whilst the estimated invasive potential of K1
430 was not top-ranked, the ST95 lineage ranked 4th in the lineage random effects estimates,
431 suggesting the lineage is playing a key role in progression to bloodstream infections; this is
432 concordant with the many reported virulence factors found in this lineage[50,51]. In summary, our
433 results provide in this regard impetus for reinstating phenotypic capsular typing for this species
434 and renewed interest for experimental studies of virulence that could further disentangle the
435 contributions of capsules versus lineages towards invasiveness.
436
437 The development of ExPEC vaccines promises to improve public health by reducing the burden
438 of invasive disease. Still, their potential is complicated because many *E. coli* are commensal,
439 and some are considered beneficial to human health[52]. Therefore, vaccines must avoid
440 targeting a major constituent of the gut microbiota and broad *E. coli* sterilising immunity. This
441 requires a greater understanding of the differences between predominantly commensal
442 genotypes and successful ExPECs, as well as the pathogenic potential of the latter. K-types
443 with high estimated invasive potential are obvious vaccine targets due to both their capacity to
444 cause disease and because the removal of such types from the gut flora is less likely to affect
445 key *E. coli* commensals or have the potential for large replacement in carriage and disease as
446 other colonisers fill the vacated niche. While ExPEC vaccines have been under development
447 since the 1980s, only a few have been licensed. Whole-cell, O, H, K, and O-conjugates have all
448 undergone clinical trials[53]. Although evidence of K-autoreactivity is limited, mimicry of host
449 glycobiology may explain why anti-K antibodies are not strongly induced in vaccination or
450 disease[53]. Nonetheless, increased genomic data and expanded O-typing have contributed to
451 the increased development of subunit-based vaccines rather than whole-cell vaccines in recent
452 years[6,54].

453

454 Efforts to quantify the invasive potential of different K-types have been hindered by the scarcity
455 of *E. coli* colonisation data, which we overcame by building on recent breakthroughs in high-
456 resolution genotyping of pathogens from shotgun metagenomics[23]. Specifically, we leveraged
457 microbiome data from a large-scale UK infant-mother cohort[34,55] to survey asymptomatic *E. coli*
458 carriage. Using infant carriage to represent general colonisation is one limitation of this
459 approach. However, as healthy full-term babies acquire *E. coli* from family members and their
460 surroundings during their first year of life, reflected by the strain-sharing of the mother and the
461 baby[23,56], this suggests that the isolates seen in the infant population likely represent the
462 commensals circulating in the population more generally. Furthermore, we found no significant
463 differences in the overall K-type composition between the UK infant-mother collection and
464 French adults. K1 and CC10 were the only K-type and lineage with a significantly higher
465 prevalence in the French collection. If these two differences in carriage were due to the different
466 age-groups sampled, the relative invasive potential could be overestimated for K1 and CC10 in
467 our analysis; this, in fact, would not alter our conclusion that high K1 prevalence in carriage
468 influences the high prevalence in disease and that other lesser studied capsular types top the
469 invasiveness rankings. Despite differences in the sampling time period (2014-2017 vs. 2010),
470 sequencing approach (metagenomic vs. single colony sequencing) and geographical location,
471 the two carriage populations are similar, suggesting that there is limited age structure to human
472 *E. coli* colonisation in this geographical region. Whilst here we are currently limited to UK data
473 for which representative snapshots of both carriage and disease have been sampled and
474 sequenced, it has been shown that estimates of invasive potential have temporal and
475 geographical stability for other species, as this is considered a fixed trait of the capsule structure
476 independent of when and where matched datasets were isolated.[57]
477

478 Given that we observed some age and sex differences in epidemiology, the invasive potential
479 may vary in specific patient groups. The estimates reported in this manuscript represent the
480 average estimated relative invasive potential in BSIs across all age groups. However, 74% of
481 BSIs are in elderly patients of greater than 59 years, and so the estimates are likely to better
482 represent this important high-risk age group. Limiting the data to only elderly adults aged 59
483 years or older had a negligible effect on the rank order but, as expected, increased uncertainty
484 in the confidence intervals of all the K-type estimates. There was insufficient data from elderly
485 adults alone to estimate the relative invasive potential for four capsular types (KL30, K4, KL13,
486 KL70). The only other notable difference was that K12 moved up four places in the rank order

487 for the estimates using only elderly adults. It may thus represent a K-type that either requires a
488 more vulnerable population or circulates more commonly in older demographics, which were not
489 captured by the infant-mother carriage cohort. This further highlights how systematic population-
490 based K-type screening paired with high-quality metadata can advance understanding of any
491 potential age and sex-specific epidemiology of ExPEC infections. As further colonisation data is
492 generated alongside additional BSI metadata, there is potential for future meta-analyses across
493 countries that would increase the power to estimate the relative invasive potential of K-types in
494 different groups and reduce uncertainty around the point estimates.
495
496 Hard-to-treat infections could also conceivably be overrepresented in BSI isolates due to an
497 increased opportunity to progress to a systemic infection and influence estimates of invasive
498 potential. To account for this, we tested the association between the lineage random effects and
499 the proportion of UK BSI isolates per lineage that were either phenotypically MDR (EUCAST
500 v15) or $bla_{CTX\text{-}M}$ positive. There was no correlation with the MDR proportion ($R^2 = 0.029$, $p =$
501 0.79), and when the CC131 C2 outlier was excluded, there was also no correlation with the
502 $bla_{CTX\text{-}M}$ proportion ($R^2 = 0.036$, $p = 0.74$). Furthermore, when including the $bla_{CTX\text{-}M}$ proportion
503 per lineage in an alternative mixed model, this covariate was not significant, and the only
504 resulting difference was that the CC131 C2 lineage ranking fell one position (behind CC12).
505 Finally, to further assess its robustness, we applied our model to a pneumococcal collection that
506 had previously been used to assess invasive potential [58] and reproduced the general rank order
507 of invasiveness by others, which includes capsules since long established to be the most
508 virulent for this bacterium[17,57,58].
509
510 Currently, there are over twice as many recognised O-serogroups as K-types[53]. Yet, we
511 observed a greater diversity of K-loci from G2 and G3 than O-loci in most dominant ExPEC
512 lineages. This result was obtained despite the conservative 70% sequence identity threshold
513 used here to define members of the same gene cluster compared to other typing schemes
514 [10,13,15]. Even greater diversity is likely to be discovered with further investigation into the genetic
515 determinants and delineation of serogroups. As little as a single base pair difference has been
516 seen in capsule-determining glycosyl-transferases of other species[59]. Indeed, we observe
517 multiple serotypes with the same capsular genes with high-sequence identities. These include
518 the K13 and K23 reference phenotypes that are both typed as KL21. They have previously been
519 reported to belong to a serogroup, along with K20, which has replaced one gene relative to K13
520 and K23[60]. When comparing K13 (NCTC9022, ERR999921) and K23 (NCTC10430,

521 ERR968281), there were only two non-synonymous changes in a region 2 gene, *vatD,* that
522 encodes an acetyltransferase. This putative determinant needs an in-depth phenotypic
523 validation to allow full genomic discrimination between them. The G3 phenotypes K54 and K96,
524 also share a set of capsular genes and are typed as KL36. Region 2 in these references (K96
525 MG736915, K54 CP9 ESC_JB5109AA_AS)[61], differs most notably by one non-synonymous
526 change that resulted in missense for 9 AA and truncation of 8 AA at the end of the K96-07
527 encoding an alginate O-acetyltransferase. In addition to this, though, most KL36 in BSIs also
528 had a truncation of 6 AA at the start of the K96_07 alginate O-acetyltransferase protein, not
529 seen in either the K54 or K96 phenotyped reference genomes. Other related clusters of K-loci
530 which differ by a single region 2 gene or with gene truncations could also represent further
531 cross-reactive serogroups. Alternatives to traditional K-phenotyping are in development that
532 could discriminate between subtly different K-phenotypes and allow novel K-types to be
533 assessed more readily.[62]
534
535 The evolution of the *E. coli* G2 and G3 K-locus is also of considerable interest *per se*, given the
536 fundamental biological role of the capsule and the complexity of its biosynthesis process. We
537 observed that a specific subgroup of G2 K-loci were found to be undergoing atypical
538 diversification of the K-locus, one of which had an invasiveness OR of significantly above one,
539 showing that atypical K-loci can also be of clinical importance. The divergent *kps* genes seen in
540 atypical K-loci may increase the opportunity for cross-species capsular gene exchange, and
541 their evolution warrants additional study. We were able to demonstrate an association between
542 estimated recombination to mutation rate and the observed capsule diversity per genetic
543 lineage, which suggests that the overall tendency of homologous recombination across the core
544 chromosome closely reflects the rate at which capsule switches occur via horizontal gene
545 transfer, influencing the K locus. Whilst earlier work detected the acquisition of the K1 capsule
546 by genetic lineages across several phylogroups over centuries[46], here we demonstrate that a
547 similar process has influenced the spread of most of the successful and invasive capsule types.
548 Notably, the capsule repertoire of a genetic lineage can grow rapidly by successful acquisitions
549 of new capsule types that expand in the population, whilst the O and H loci remain unchanged,
550 as demonstrated by the CC131 C1 and C2 subclades. The ability to diversify the capsule locus
551 for two major MDR lineages (CC69 and CC131) may have facilitated their rapid expansion in
552 BSIs in the 21st century.[24,27,28] We further discovered that IS elements are common within the K
553 loci and are likely contributing to the diversification of the polysaccharide composition of
554 capsules by importing genes into region 2, warranting deeper functional investigation of these

555 evolutionary processes. This appears to be a unique feature of *E. coli* G2 and G3 capsules;
556 whilst IS-elements have been reported in other species like *Klebsiella pneumoniae*[13,15], they
557 could be completely removed from the typing scheme as they were less common and purely
558 intergenic.
559
560 Taken together, our study provides further evidence for the need for improved understanding of
561 K-types and renewed experimental investigation to allow the development of strategies to
562 effectively control ExPEC disease.

Methods

Data

We used a published pangenome analysis of ~7,500 high-quality *E. coli* genomes to determine that the capsular gene *kpsF* was consistently annotated and predictive of G2 capsule presence [37]. We subsequently downloaded all *kpsF*-positive (0.9 kmer-ID threshold) *E. coli* in a published searchable collection of 661-thousand bacterial assemblies (n=11,623)[36]. For G3, we additionally screened the 661K database for all *kpsM*-positive assemblies (0.9 kmer-ID threshold, n=853) using the G3-A and G3-B *kpsM* alleles from K96 and K11, respectively, as *kpsM* is more divergent in G3. We supplemented these genomes with data from large published *E. coli* BSI longitudinal genomic collections from Norway (2002-2017, n=3,254, 60% hybrid assemblies)[27,63] and the UK (2001-2018 n=2,219)[24], and a collection of infant and mother *E. coli* carriage assemblies from the UK babybiome study (n=997, available from https://zenodo.org/records/14000489)[23,34], and two one-health *E. coli* studies from the UK (n=405) and USA (n=2948)[64,65]. Genomes with a known K phenotype were sourced for 24 different K-types from GenBank, the NCTC project[66] and Enterobase[67]. The NCTC reference strains (accession PRJEB6403) are maintained as preserved strains that can be obtained from the National Collection of Type Cultures, UK. K-loci were initially extracted from 21,802 assemblies using an *in-silico* PCR (https://github.com/simonrharris/in_silico_pcr) for known K-loci boundaries, i.e. *kpsF-kpsM* or *kpsM/D-kpsS*. PCR-negative K-loci were then manually extracted from assemblies that had been annotated with one or more *kps* genes.

K-loci

Extracted K-loci with unknown bases were excluded, and the K-loci were filtered down to 4,996 unique sequences with at least 1bp difference. These unique K-loci were annotated with Prokka[68] and analysed with Panaroo[69] 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. This resulted in an initial 225 gene absence patterns, excluding insertion elements (IS), which are arbitrarily numbered KL1-225. IS elements were identified from the annotations using ISEscan[70]. These 225 sequences were pruned to remove K-loci with redundant patterns due to misannotation,

592 paralogs, non-capsular genes, and incomplete K-locus remnants, leaving 90 curated K-loci in
593 the final K-loci database. K-loci from complete genomes, long-read/hybrid assemblies and
594 sequences containing the fewest available IS were preferentially included as the database
595 reference. The database was formatted for Kaptive3 [26], with IS annotations removed.
596 Unpublished data from European disease (n=1,592) and carriage from underrepresented
597 regions in Asia and Africa (n=10,085) were screened to determine that no dominant K-types had
598 been missed.
599
600 Assemblies (n=8,473) from published genomic collections [23,24,28,30–32,34,35,71,72] were K-typed
601 using Kaptive version 3.0.0b5, and this G2-G3 database, which is available at
602 https://github.com/rgladstone/EC-K-typing. Untypeables with three or fewer essential *kps* genes
603 were designated absent for G2-G3 K-loci. The assignment of chromosome or plasmid was
604 available for hybrid assemblies from Norway and Oxfordshire[29,63]. These were used to
605 determine if any K-loci were observed in plasmids in these BSI collections.
606
607 The relatedness of K-loci was assessed for G2 and G3 separately using the pairwise mash
608 hash distance of the K-locus sequences converted to a proportion 1-(n/1000), which gives
609 greater resolution than gene presence-absence but still captures core and accessory variation[73].
610 These distances were used to create Neighbour-joining phylogenies in the R v4.4.1 package
611 ape 5.8 and visualised in Phandango.net[74] using the following input files
612 https://github.com/rgladstone/EC-K-typing/tree/main/phandango.

613 Phenotypic confirmation

614 One hundred and fifty Norwegian BSI isolates were selected for phenotypic K-typing; these
615 were selected to represent imperfect matches to K-loci with known phenotypes (n=71), K-loci
616 where a phenotypic reference genome was not available (n=40) and isolates which were kps-
617 negative (n=39) to distinguish between G1-G4 and potentially acapsular isolates. K-phenotyping
618 was carried out by the Staten Serum Institute using reference capsular antisera. These data
619 were used to confirm phenotypes assigned in the K-typing database and to determine if K-loci
620 without a known K-phenotype expressed a putatively novel capsule using a positive precipitate
621 reaction with Cetavlon but negative with known K antigens.

622 ## *E. coli* carriage assemblies from metagenomic data

623 *E. coli* assemblies were extracted from metagenomic data from the UK babybiome data using a
624 computational approach described in a previous study[72] to provide an *E. coli* gut colonisation
625 dataset[34,75]. The sequencing read data was first pseudo-aligned against a diverse reference
626 database constructed from the 661,000 assemblies study[36] (available from
627 https://zenodo.org/records/7736981) with Themisto[76]. The alignments were then processed with
628 mSWEEP and mGEMS[77,78] to assign each sequencing read to a bacterial species. Next, the
629 reads assigned to *E. coli* were pseudo-aligned with Themisto against an *E. coli* database
630 (https://zenodo.org/records/12528310) and processed again with mSWEEP and mGEMS to
631 obtain assignments of the sequencing reads to *E. coli* PopPUNK lineages[79]. The resulting
632 assignments ('bins') were quality controlled with demix_check (https://github.com/harry-
633 thorpe/demix_check) and bins with scores 1 or 2 were retained (n=1,402). Reads belonging to
634 the 1,402 bins were assembled with Shovill (https://github.com/tseemann/shovill), and small
635 contigs <5000bp were discarded before K-typing. The assemblies are available at
636 https://zenodo.org/records/14000489. As this collection sampled individuals within a family
637 (mother and infant) multiple times during the first year of life, the isolates were deduplicated
638 (n=997/1,402), allowing only one lineage representative with the same K-type per family.
639 Mothers accounted for 67/997 of the isolates and infants between 6 months and 1 year old
640 accounted for one third of the deduplicated set. When both a typeable and untypeable isolate
641 were observed within a family for a particular lineage, we preferentially selected the K-typed
642 isolate, as they are likely to be of the same strain but with a higher quality assembly; this
643 accounts for the fact that lower coverage assemblies limit K-type detection.

644 ## Relative invasive potential

645 Lineages were defined based on core and accessory distances using PopPUNK[79]. Each lineage
646 is named after the representative sequence type for each popPUNK lineage and described as a
647 clonal Complex (CC), as each lineage contains multiple STs. To avoid a temporal bias whilst
648 estimating the relative invasive potential, due to differences in collection years between BSI
649 (2001-2017) and carriage (2014-2017), we excluded BSIs from 2001-2002 from the analysis to
650 remove a known expansion of the CC69 and CC131 lineages, which later reached a stable
651 equilibrium population frequency by 2003. Furthermore, we tested for a difference in proportions
652 for each lineage and K-type between the 2003-2013 and 2014-2017 periods using two-sided
653 Fisher's exact tests and adjusted for multiple testing. Only CC393 exhibited a significant

654 difference between the periods that could introduce bias. The K-loci, lineage information, and
655 the source of each isolate (infection/carriage) were input for modelling the relative invasive
656 potential of K-types and lineages using a generalised mixed model approach[23,24,28,34]. The model
657 included the isolation source as a binary outcome variable, the K-loci indicator variable as a
658 fixed effect (constant for each isolate) and the lineage indicator variable as a random effect.
659 Untypeable (*kps*-negative genomes or rare loci) was set as the reference category for K-loci.
660 Data were filtered to include only K-loci found in more than 20 isolates, with at least 5 isolates in
661 each infection and carriage group. The model was fitted in R using the glmer-function in the
662 lme4-package version 1.1.35.5[80]. The age of BSI patients, the likelihood that a strain was CTX-
663 M positive, and an increase in BSI prevalence of CC393 between 2002-2013 and 2014-2017
664 were possible confounders, the model was run four additional times: 1) BSAC restricted to
665 adults only n=1550/2036, 2) BSAC elderly adults only n=1202/2036, 3) BSAC with CC393
666 subsampled in 2003-2013 down to the same proportion as 2014-2017 of 0.02 n=2008/2036, 4)
667 adjusting for the CTX-M prevalence of each lineage in BSIs. We further assessed whether the
668 infant-mother metagenomic carriage data had significantly different K-epidemiology to a healthy
669 adult carriage whole genome collection [35] sampled in France in 2010, using a nonparametric
670 test of independence in R package coin v1.4-3 for the K-loci and lineages included in the
671 invasive potential model, and a comparison of proportions for each individual K-type and lineage
672 included in our model between these two carriage collections with the "BH" method for adjusting
673 for multiple testing. The model was also applied to a previously published pneumococcal
674 dataset for assessing invasive potential[17,58], leveraging the availability of paired carriage and
675 disease collections from two countries included in the model as a fixed effect. This multi-country
676 approach increased the sample size to 1,415 carriage isolates and 1,081 disease isolates.
677 https://github.com/rgladstone/EC-K-typing/tree/main/invasiveness/pneumo_example.

678 Lineage analysis

679 The top 10 PopPUNK[79] lineages were mapped against a corresponding ST reference, aligned,
680 and recombination removed with Gubbins[81,82]. The recombination over mutation ratio (r/m) per
681 lineage was estimated with Gubbins, and the median pairwise recombination-free mutational
682 distance (MMD) for each lineage was calculated using pairSNP[83]. The Pearson correlation
683 between r/m and the K-locus count per lineage, with adjustment for lineage diversity, using the
684 MMD, was calculated in R v4.4.1 base corr.test for the top 10 lineages with >95% G2-G3
685 capsules. For dating analysis, recombination-free phylogenies for CC69 and CC131 using
686 Norwegian and UK BSI data[24,27] were created with Gubbins v3.4[82] using SKA2[84] and dated with

687 the BactDating[85] ARC model with a 10,000,000 MCMC chain in three replicates and one with
688 randomised tip dates. The effective sampling size was >200, the Gelman statistic for
689 convergence was ~1, and the true date models were better than the random date models. The
690 resultant phylogenies were visualised in microreact[86]. To characterise the diversity of K-loci
691 within a lineage, we used Simpson's diversity index (SDI 1-D), which measures the relationship
692 between the number of "species" (here we use either K-loci or lineages), termed richness, and
693 individuals within each "species", termed evenness. O:H-typing was performed with SRST2[6].
694 Differences between proportions were tested with Fisher's (when any n<5) or Chi-squared
695 (when all n>5).

Acknowledgements

We thank 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 Dr Kat Holt and Dr Kelly Wyres, who kindly advised us in maximising the database's utility with the Kaptive tool. Finally, we thank Susanne Jespersen and Marie Vilborg Jacobsen for their efforts in K-typing 150 isolates supported by the Gates Foundation.

Declaration of interests

The authors declare no competing interests

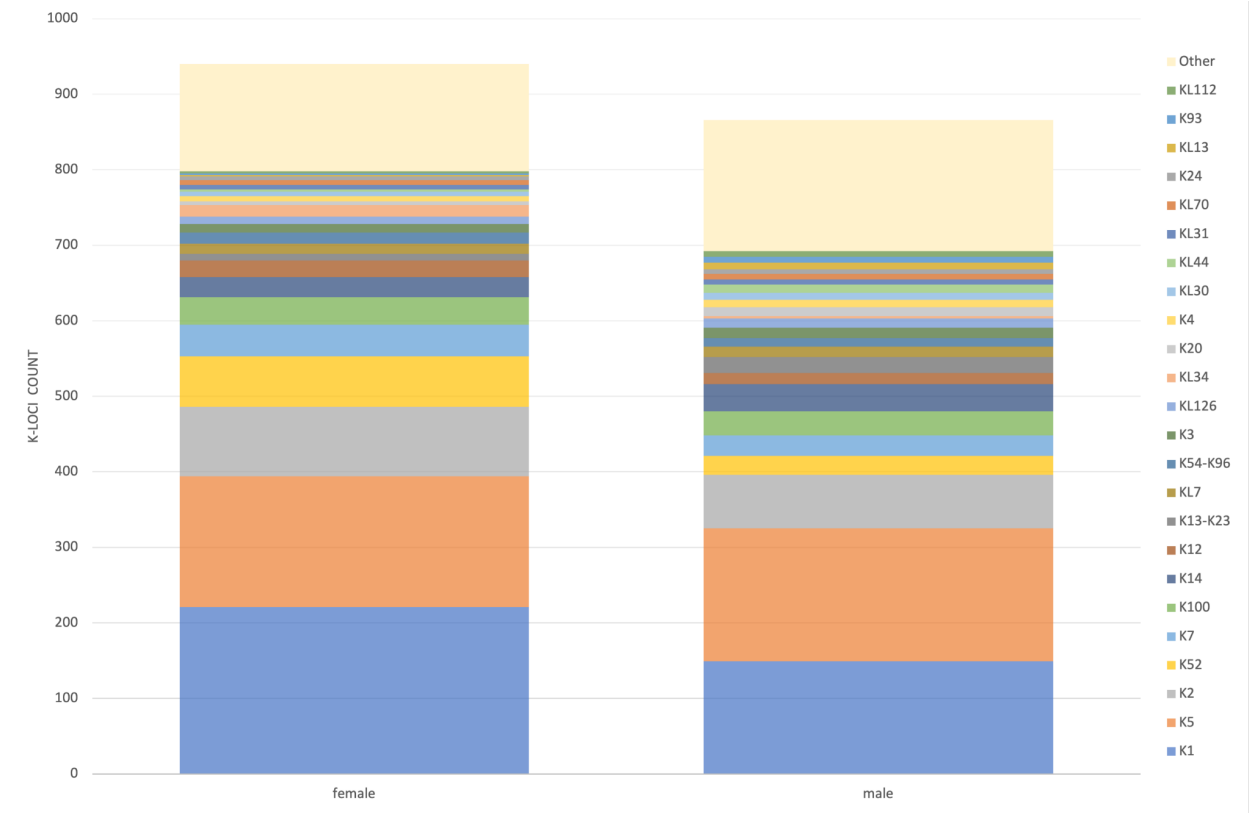
Inclusion and ethics statement

This publication used published genomic collections and metadata.

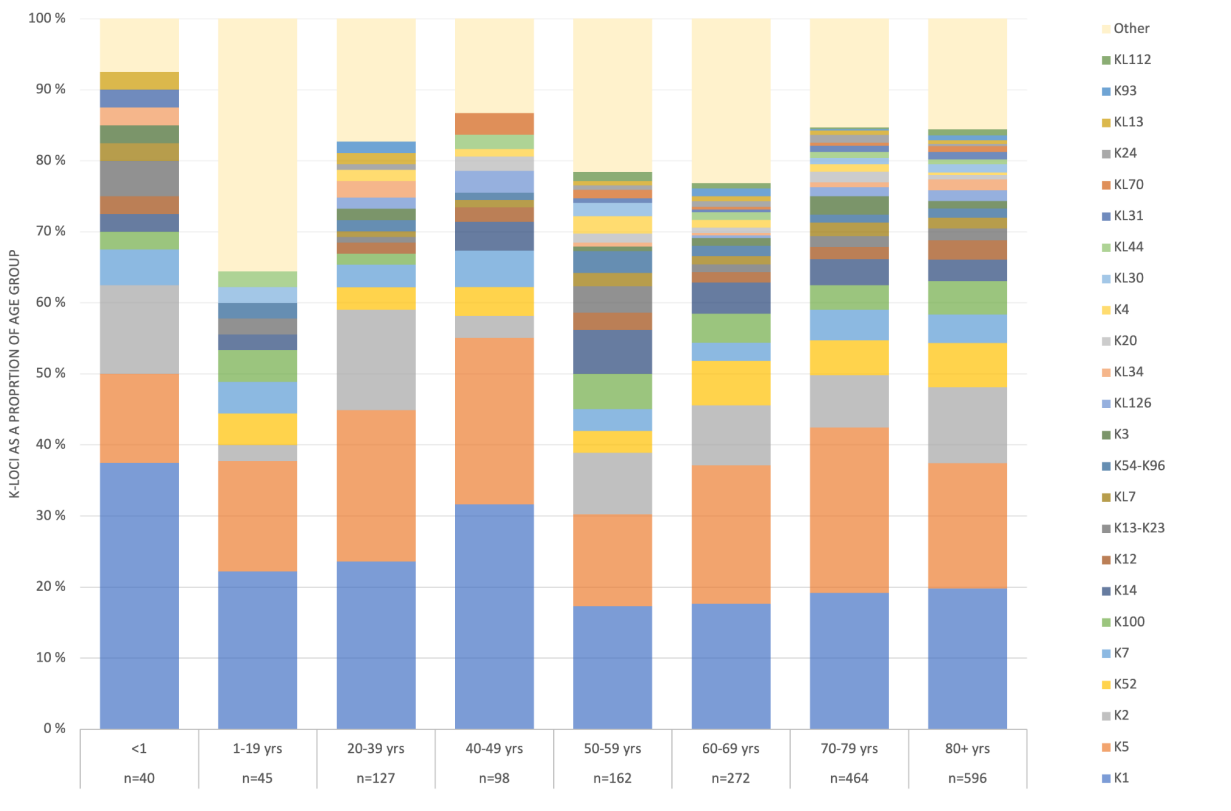
712 # Supplementary

713 Supplementary Figure 1. K-loci in UK BSIs A) K-loci by sex B) K-loci per age group C) K-loci per
714 age group by sex
715 Supplementary Figure 2. Gene presence and absence for A) the K5-like loci and B) K4-related
716 loci
717 Supplementary Figure 3. Genetic architecture of atypical G2 K-loci
718 Supplementary Data - isolate metadata, database details, phenotypic data

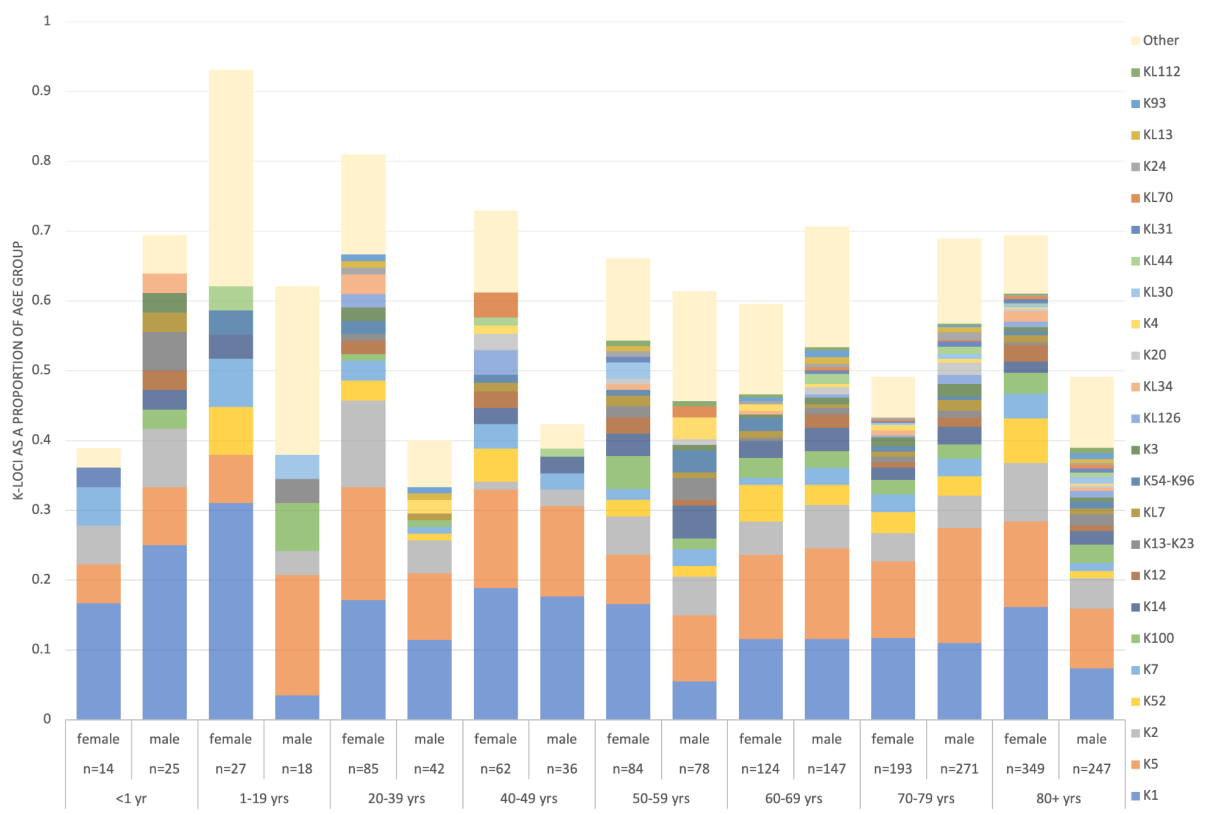
719 **Supplementary Figure 1. K-loci in UK BSIs A) K-loci by sex B) K-loci per age group C) K-loci per**
720 **age group by sex**



721


K-LOCI AS A PROPRTION OF AGE GROUP
100 %
90 %
80 %
70 %
60 %
50 %
40 %
30 %
20 %
10 %
0 %
<1
n=40
1-19 yrs
n=45
20-39 yrs
n=127
40-49 yrs
n=98
50-59 yrs
n=162
60-69 yrs
n=272
70-79 yrs
n=464
80+ yrs
n=596
Other
KL112
K93
KL13
K24
KL70
KL31
KL44
KL30
K4
K20
KL34
KL126
K3
K54-K96
KL7
K13-K23
K12
K14
K100
K7
K52
K2
K5
K1

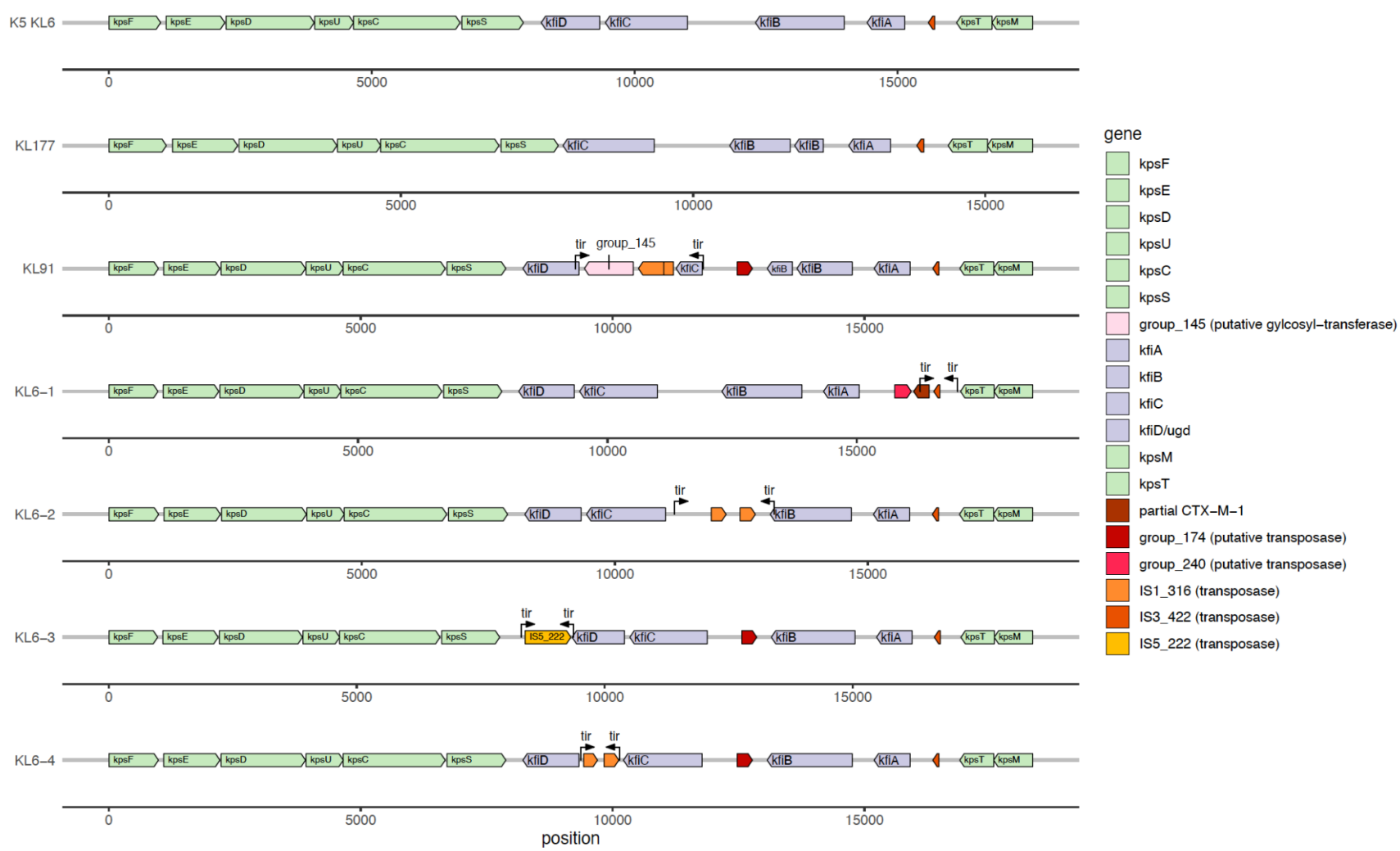
722


K-LOCI AS A PROPORTION OF AGE GROUP
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
female
n=14
male
n=25
<1 yr
female
n=27
male
n=18
1-19 yrs
female
n=85
male
n=42
20-39 yrs
female
n=62
male
n=36
40-49 yrs
female
n=84
male
n=78
50-59 yrs
female
n=124
male
n=147
60-69 yrs
female
n=193
male
n=271
70-79 yrs
female
n=349
male
n=247
80+ yrs
Other
KL112
K93
KL13
K24
KL70
KL31
KL44
KL30
K4
K20
KL34
KL126
K3
K54-K96
KL7
K13-K23
K12
K14
K100
K7
K52
K2
K5
K1

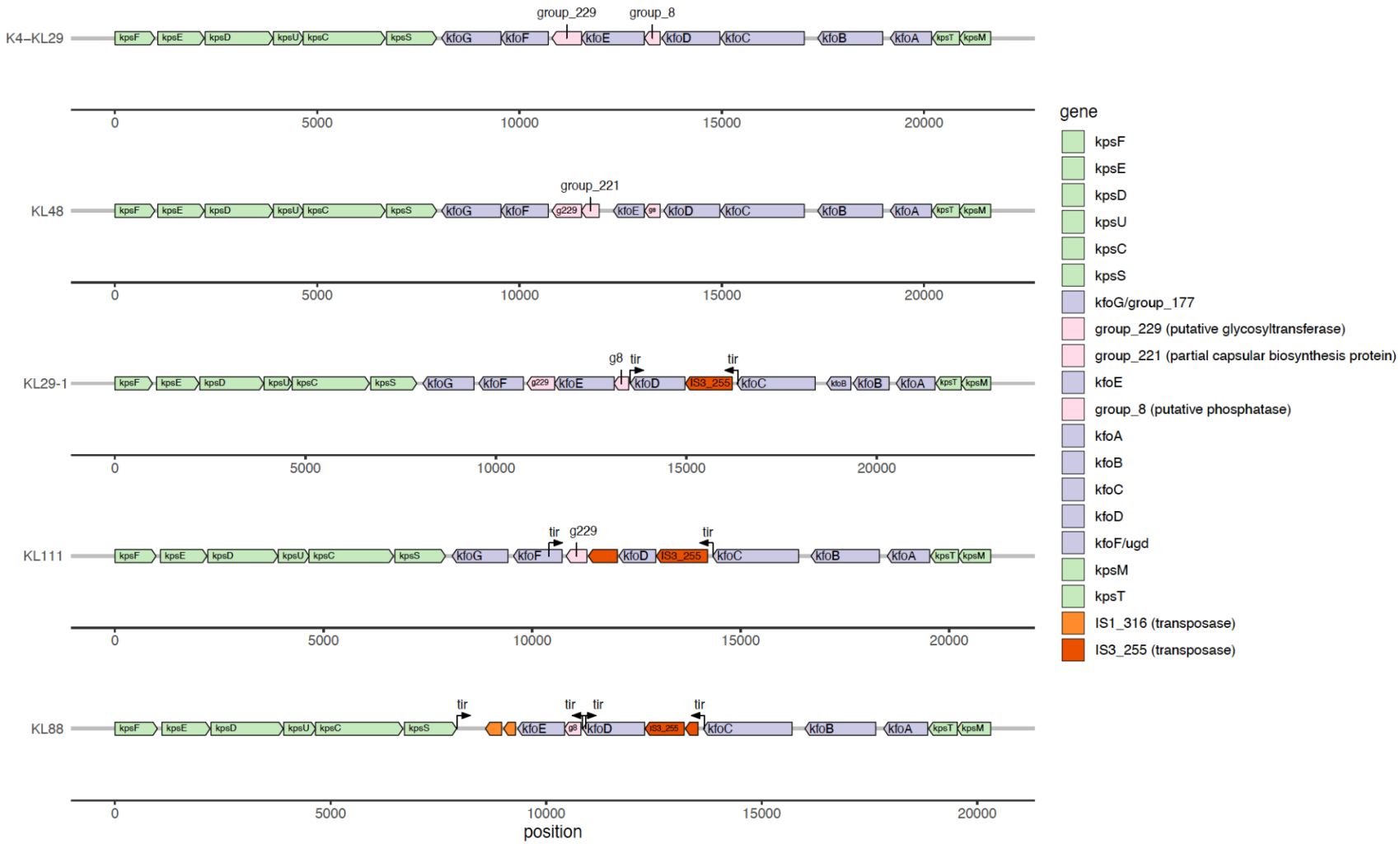
723

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

A)

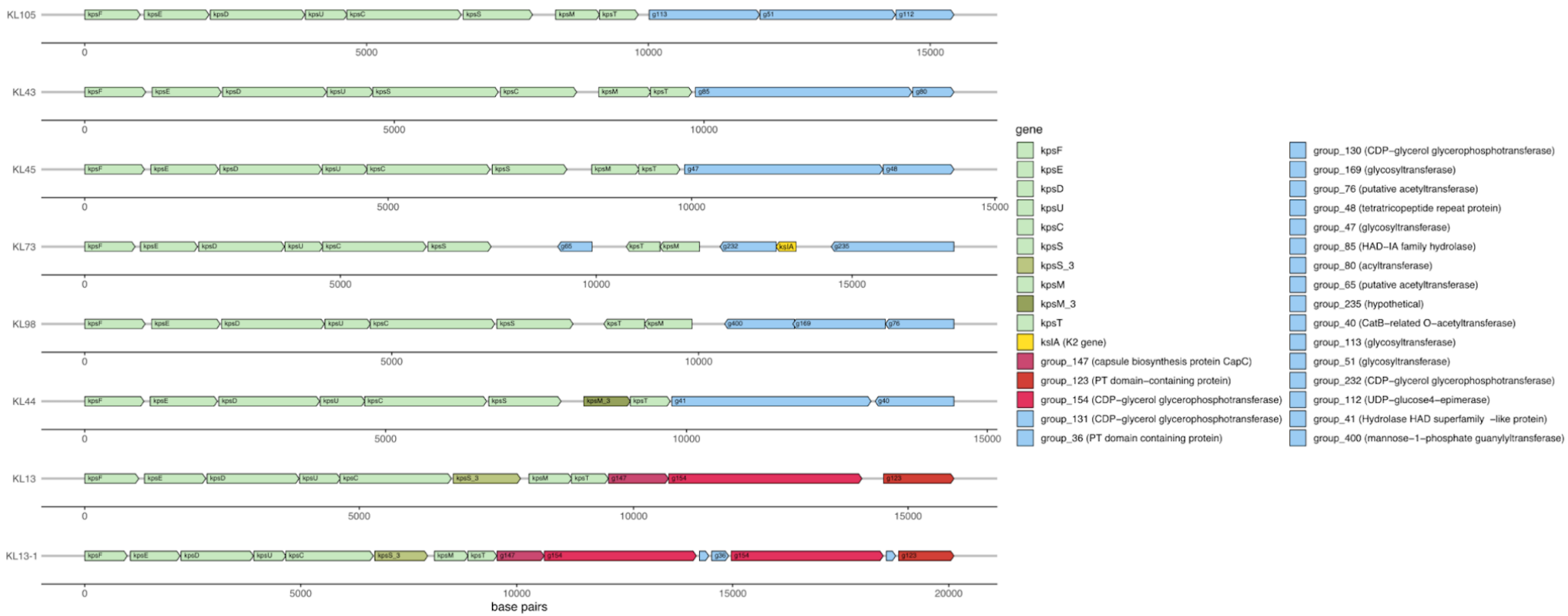

K5 KL6
KL177
KL91
KL6–1
KL6–2
KL6–3
KL6–4
kpsF
kpsE
kpsD
kpsU
kpsC
kpsS
kfiD
kfiC
kfiB
kfiA
kpsT
kpsM
tir
group_145
IS5_222
0
5000
10000
15000
position
gene
kpsF
kpsE
kpsD
kpsU
kpsC
kpsS
group_145 (putative gylcosyl–transferase)
kfiA
kfiB
kfiC
kfiD/ugd
kpsM
kpsT
partial CTX–M–1
group_174 (putative transposase)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

B)



Supplementary Figure 3. Genetic architecture of atypical G2 K-loci

735 **The conserved Regions 1 and 3 genes are light green, with rarer gene clusters in darker green. Region 2 genes observed in known K-**
736 **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**
737 **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.

762 *Microbiol Spectr* **11**, e0143223 (2023).
763 11. Holt, K. E., Lassalle, F., Wyres, K. L., Wick, R. & Mostowy, R. J. Diversity and evolution of
764 surface polysaccharide synthesis loci in Enterobacteriales. *ISME J.* **14**, 1713–1730 (2020).
765 12. Mostowy, R. J. & Holt, K. E. Diversity-generating machines: Genetics of bacterial sugar-
766 coating. *Trends Microbiol.* **26**, 1008–1021 (2018).
767 13. Wyres, K. L., Cahill, S. M., Holt, K. E., Hall, R. M. & Kenyon, J. J. Identification of
768 Acinetobacter baumannii loci for capsular polysaccharide (KL) and lipooligosaccharide
769 outer core (OCL) synthesis in genome assemblies using curated reference databases
770 compatible with Kaptive. *Microb Genom* **6**, (2020).
771 14. Kunduru, B. R., Nair, S. A. & Rathinavelan, T. EK3D: an E. coli K antigen 3-dimensional
772 structure database. *Nucleic Acids Res.* **44**, D675–81 (2016).
773 15. Wyres, K. L. *et al.* Identification of Klebsiella capsule synthesis loci from whole genome
774 data. *Microb Genom* **2**, e000102 (2016).
775 16. Yang, S. *et al.* Genetic diversity of K-antigen gene clusters of Escherichia coli and their
776 molecular typing using a suspension array. *Can. J. Microbiol.* **64**, 231–241 (2018).
777 17. Løchen, A., Truscott, J. E. & Croucher, N. J. Analysing pneumococcal invasiveness using
778 Bayesian models of pathogen progression rates. *PLoS Comput. Biol.* **18**, e1009389 (2022).
779 18. Corander, J. *et al.* Frequency-dependent selection in vaccine-associated pneumococcal
780 population dynamics. *Nat Ecol Evol* **1**, 1950–1960 (2017).
781 19. Colijn, C., Corander, J. & Croucher, N. J. Designing ecologically optimized pneumococcal
782 vaccines using population genomics. *Nat Microbiol* **5**, 473–485 (2020).
783 20. Sleeman, K. L. *et al.* Capsular serotype-specific attack rates and duration of carriage of
784 Streptococcus pneumoniae in a population of children. *J. Infect. Dis.* **194**, 682–688 (2006).
785 21. Jounio, U. *et al.* Genotypic and phenotypic characterization of carriage and invasive
786 disease isolates of Neisseria meningitidis in Finland. *J. Clin. Microbiol.* **50**, 264–273 (2012).
787 22. Yazdankhah, S. P. *et al.* Distribution of serogroups and genotypes among disease-

788 associated and carried isolates of Neisseria meningitidis from the Czech Republic, Greece,
789 and Norway. *J. Clin. Microbiol.* **42**, 5146–5153 (2004).
790 23. Mäklin, T. *et al.* Strong pathogen competition in neonatal gut colonisation. *Nat. Commun.*
791 **13**, 7417 (2022).
792 24. Pöntinen, A. K. *et al.* Modulation of multi-drug resistant clone success in Escherichia coli
793 populations: a longitudinal multi-country genomic and antibiotic usage cohort study. *Lancet*
794 *Microbe* (2024) doi:10.1016/S2666-5247(23)00292-6.
795 25. Ojala, F. *et al.* Basic reproduction number for pandemic *Escherichia coli* clones varies
796 markedly and can be comparable to pandemic influenza viruses. *bioRxiv*
797 2024.05.08.593267 (2024) doi:10.1101/2024.05.08.593267.
798 26. Lam, M. M. C., Wick, R. R., Judd, L. M., Holt, K. E. & Wyres, K. L. Kaptive 2.0: updated
799 capsule and lipopolysaccharide locus typing for the Klebsiella pneumoniae species
800 complex. *Microb Genom* **8**, (2022).
801 27. Gladstone, R. A. *et al.* Emergence and dissemination of antimicrobial resistance in
802 Escherichia coli causing bloodstream infections in Norway in 2002–17: a nationwide,
803 longitudinal, microbial population genomic study. *The Lancet Microbe* **2**, e331–e341 (2021).
804 28. Kallonen, T. *et al.* Systematic longitudinal survey of invasive Escherichia coli in England
805 demonstrates a stable population structure only transiently disturbed by the emergence of
806 ST131. *Genome Res.* (2017) doi:10.1101/gr.216606.116.
807 29. Lipworth, S. *et al.* The plasmidome associated with Gram-negative bloodstream infections:
808 A large-scale observational study using complete plasmid assemblies. *Nat. Commun.* **15**,
809 1612 (2024).
810 30. Handal, N. *et al.* A comparative study of the molecular characteristics of human
811 uropathogenic Escherichia coli collected from two hospitals in Norway and France in 2019.
812 *J. Antimicrob. Chemother.* dkaf130 (2025).
813 31. Thänert, R. *et al.* Persisting uropathogenic Escherichia coli lineages show signatures of

814 niche-specific within-host adaptation mediated by mobile genetic elements. *Cell Host*
815 *Microbe* **0**, (2022).
816 32. Sands, K. *et al.* Characterization of antimicrobial-resistant Gram-negative bacteria that
817 cause neonatal sepsis in seven low- and middle-income countries. *Nat Microbiol* **6**, 512–
818 523 (2021).
819 33. Pearse, O. *et al.* High diversity of Escherichia coli causing invasive disease in neonates in
820 Malawi poses challenges for O-antigen based vaccine approach. *Commun. Med. (Lond.)* **5**,
821 298 (2025).
822 34. Shao, Y. *et al.* Primary succession of Bifidobacteria drives pathogen resistance in neonatal
823 microbiota assembly. *Nat. Microbiol.* **9**, 2570–2582 (2024).
824 35. Marin, J. *et al.* The Population Genomics of Increased Virulence and Antibiotic Resistance
825 in Human Commensal Escherichia coli over 30 Years in France. *Appl. Environ. Microbiol.*
826 **88**, e0066422 (2022).
827 36. Blackwell, G. A. *et al.* Exploring bacterial diversity via a curated and searchable snapshot of
828 archived DNA sequences. *PLoS Biol.* **19**, e3001421 (2021).
829 37. Horesh, G. *et al.* A comprehensive and high-quality collection of Escherichia coli genomes
830 and their genes. *Microb Genom* **7**, (2021).
831 38. Whitfield, C. & Roberts, I. S. Structure, assembly and regulation of expression of capsules
832 in Escherichia coli. *Mol. Microbiol.* **31**, 1307–1319 (1999).
833 39. Sande, C. & Whitfield, C. Capsules and Extracellular Polysaccharides in Escherichia coli
834 and Salmonella. *EcoSal Plus* **9**, eESP00332020 (2021).
835 40. Lo, S. W. *et al.* Pneumococcal lineages associated with serotype replacement and
836 antibiotic resistance in childhood invasive pneumococcal disease in the post-PCV13 era: an
837 international whole-genome sequencing study. *Lancet Infect. Dis.* **19**, 759–769 (2019).
838 41. Garcia Quesada, M. *et al.* Serotype distribution of remaining invasive pneumococcal
839 disease after extensive use of ten-valent and 13-valent pneumococcal conjugate vaccines

840 (the PSERENADE project): a global surveillance analysis. *Lancet Infect. Dis.* **25**, 445–456
841 (2025).
842 42. Maeda, H. & Morimoto, K. Global distribution and characteristics of pneumococcal
843 serotypes in adults. *Hum. Vaccin. Immunother.* **21**, 2469424 (2025).
844 43. Moxon, E. R. & Kroll, J. S. The role of bacterial polysaccharide capsules as virulence
845 factors. *Curr. Top. Microbiol. Immunol.* **150**, 65–85 (1990).
846 44. Buckles, E. L. *et al.* Role of the K2 capsule in Escherichia coli urinary tract infection and
847 serum resistance. *J. Infect. Dis.* **199**, 1689–1697 (2009).
848 45. Cross, A. S., Kim, K. S., Wright, D. C., Sadoff, J. C. & Gemski, P. Role of
849 lipopolysaccharide and capsule in the serum resistance of bacteremic strains of
850 Escherichia coli. *J. Infect. Dis.* **154**, 497–503 (1986).
851 46. Arredondo-Alonso, S. *et al.* Evolutionary and functional history of the Escherichia coli K1
852 capsule. *Nat. Commun.* **14**, 3294 (2023).
853 47. Merino, I. *et al.* Molecularly defined extraintestinal pathogenic Escherichia coli status
854 predicts virulence in a murine sepsis model better than does virotype, individual virulence
855 genes, or clonal subset among E. coli ST131 isolates. *Virulence* **11**, 327–336 (2020).
856 48. Mostafavi, S. K. S., Najar-Peerayeh, S., Mobarez, A. M. & Parizi, M. K. Characterization of
857 uropathogenic E. coli O25b-B2-ST131, O15:K52:H1, and CGA: Neutrophils apoptosis,
858 serum bactericidal assay, biofilm formation, and virulence typing. *J. Cell. Physiol.* **234**,
859 18272–18282 (2019).
860 49. Suerbaum, S., Friedrich, S., Leying, H. & Opferkuch, W. Expression of capsular
861 polysaccharide determines serum resistance in Escherichia coli K92. *Zentralbl. Bakteriol.*
862 **281**, 146–157 (1994).
863 50. Weissman, S. J. *et al.* Clonal analysis reveals high rate of structural mutations in fimbrial
864 adhesins of extraintestinal pathogenic Escherichia coli. *Mol. Microbiol.* **59**, 975–988 (2006).
865 51. Johnson, T. J. *et al.* Comparison of extraintestinal pathogenic Escherichia coli strains from

human and avian sources reveals a mixed subset representing potential zoonotic pathogens. *Appl. Environ. Microbiol.* **74**, 7043–7050 (2008).

52. Wassenaar, T. M. Insights from 100 years of research with probiotic E. coli. *Eur. J. Microbiol. Immunol. (Bp.)* **6**, 147–161 (2016).
53. Qiu, L. *et al.* Vaccines against extraintestinal pathogenic Escherichia coli (ExPEC): progress and challenges. *Gut Microbes* **16**, 2359691 (2024).
54. Weerdenburg, E. *et al.* Global distribution of O serotypes and antibiotic resistance in extraintestinal pathogenic Escherichia coli collected from the blood of patients with bacteremia across multiple surveillance studies. *Clin. Infect. Dis.* **76**, e1236–e1243 (2023).
55. Shao, Y. *et al.* Stunted microbiota and opportunistic pathogen colonization in caesarean-section birth. *Nature* **574**, 117–121 (2019).
56. 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.
57. Brueggemann, A. B. *et al.* Clonal relationships between invasive and carriage Streptococcus pneumoniae and serotype- and clone-specific differences in invasive disease potential. *J. Infect. Dis.* **187**, 1424–1432 (2003).
58. Gladstone, R. A. *et al.* International genomic definition of pneumococcal lineages, to contextualise disease, antibiotic resistance and vaccine impact. *EBioMedicine* **43**, 338–346 (2019).
59. Mavroidi, A. *et al.* Evolutionary genetics of the capsular locus of serogroup 6 pneumococci. *J. Bacteriol.* **186**, 8181–8192 (2004).
60. 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).
61. Olson, M. A. *et al.* Bile salts regulate zinc uptake and capsule synthesis in a mastitis-associated extraintestinal pathogenic Escherichia coli strain. *Infect. Immun.* **89**, e0035721

892 (2021).
893 62. Rutschmann, C. *et al.* DMB labelling for detection and analysis of capsular
894 polysaccharides. *Biochemistry* (2025).
895 63. Arredondo-Alonso, S. *et al.* Plasmid-driven strategies for clone success in Escherichia coli.
896 *Nat. Commun.* **16**, 2921 (2025).
897 64. Liu, C. M. *et al.* Using source-associated mobile genetic elements to identify zoonotic
898 extraintestinal E. coli infections. *One Health* **16**, 100518 (2023).
899 65. Ludden, C. *et al.* One Health Genomic Surveillance of Escherichia coli Demonstrates
900 Distinct Lineages and Mobile Genetic Elements in Isolates from Humans versus Livestock.
901 *MBio* **10**, (2019).
902 66. Dicks, J. *et al.* NCTC3000: a century of bacterial strain collecting leads to a rich genomic
903 data resource. *Microb. Genom.* **9**, mgen000976 (2023).
904 67. Zhou, Z., Charlesworth, J. & Achtman, M. HierCC: a multi-level clustering scheme for
905 population assignments based on core genome MLST. *Bioinformatics* **37**, 3645–3646
906 (2021).
907 68. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* (2014).
908 69. Tonkin-Hill, G. *et al.* Producing polished prokaryotic pangenomes with the Panaroo
909 pipeline. *Genome Biol.* **21**, 180 (2020).
910 70. Xie, Z. & Tang, H. ISEScan: automated identification of insertion sequence elements in
911 prokaryotic genomes. *Bioinformatics* **33**, 3340–3347 (2017).
912 71. Gladstone, R. A. *et al.* Emergence and Dissemination of Antimicrobial Resistance in
913 *Escherichia Coli* Causing Bloodstream Infections: A Nationwide Longitudinal Microbial
914 Population Genomic Cohort Study in Norway between 2002-2017. (2020)
915 doi:10.2139/ssrn.3645193.
916 72. Khawaja, T. *et al.* Deep sequencing of Escherichia coli exposes colonisation diversity and
917 impact of antibiotics in Punjab, Pakistan. *Nat. Commun.* (2024) doi:10.1038/s41467-024-

49591-5.

73. Ondov, B. D. *et al.* Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.* **17**, 132 (2016).

74. Hadfield, J. *et al.* Phandango: an interactive viewer for bacterial population genomics. *Bioinformatics* **34**, 292–293 (2018).

75. 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.

76. 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).

77. Mäklin, T. *et al.* High-resolution sweep metagenomics using fast probabilistic inference. *Wellcome Open Res.* **5**, 14 (2021).

78. Mäklin, T. *et al.* Bacterial genomic epidemiology with mixed samples. *Microb Genom* **7**, (2021).

79. Lees, J. A. *et al.* Fast and flexible bacterial genomic epidemiology with PopPUNK. *Genome Res.* **29**, 304–316 (2019).

80. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting linear mixed-effects models Usinglme4. *J. Stat. Softw.* **67**, (2015).

81. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).

82. Croucher, N. J. *et al.* Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res.* **43**, e15 (2015).

83. Tonkin-Hill, G. *Pairsnp-Cpp*. (Github).

84. Derelle, R. *et al.* Seamless, rapid, and accurate analyses of outbreak genomic data using split k-mer analysis. *Genome Res.* **34**, 1661–1673 (2024).

85. Didelot, X., Croucher, N. J., Bentley, S. D., Harris, S. R. & Wilson, D. J. Bayesian inference of ancestral dates on bacterial phylogenetic trees. *Nucleic Acids Res.* **46**, e134 (2018).

86. Argimón, S. *et al.* Microreact: visualizing and sharing data for genomic epidemiology and phylogeography. *Microb Genom* **2**, e000093 (2016).