# The natural diversity of *E. coli* transporter-dependent capsules

**Authors**: Carine Roese Mores[1, 14], Samuel Miravet-Verde[1, 14, *], Elisabetta Cacace[1, 14, *], Christoph Rutschmann[2, 12], Chia-Wei Lin[6], Hans J. Ruscheweyh[1], Aline Cuenod[8], Elisa Cappio Barazzone[2], Enora Marrec[1], Kateryna Vershynina[2], Raffael Schumann[2, 13], Dan J. Bower[5, 11], Mario Schubert[7], Adrian Egli[8], Timm Fiebig[9, 10], Emma Slack[2, 3, 4, *], Shinichi Sunagawa[1, *], Timothy G. Keys[2, 12, 15, *]

**Affiliations**

[1] Department of Biology, Institute of Microbiology and Swiss Institute of Bioinformatics, ETH Zurich, Zurich, Switzerland
[2] Department of Health Sciences and Technology, Institute of Food, Nutrition and Health, ETH Zurich, Zurich, Switzerland
[3] Basel Research Center for Child Health, Basel, Switzerland
[4] Sir William Dunn School of Pathology, University of Oxford, United Kingdom
[5] QuadraticMed LLC, Baar, Switzerland
[6] Functional Genomics Centre Zurich, Zurich, Switzerland
[7] Department of Biology, Chemistry and Pharmacy, Free University of Berlin, Berlin, Germany
[8] Institute for Medical Microbiology, University of Zurich, Zurich, Switzerland
[9] German Center for Infection Research (DZIF), Partner Site Hannover-Braunschweig, Hannover, Germany
[10] Institute of Clinical Biochemistry, Hannover Medical School, Hannover, Germany
[11] Department of Earth and Planetary Sciences, ETH Zurich, Switzerland.
[12] Baxiva AG, Schlieren, Switzerland.
[13] Present address: MRC Laboratory of Molecular Biology and University of Cambridge, United Kingdom
[14] These authors contributed equally.
[15] Lead contact.
* Correspondence: smiravet@biol.ethz.ch (S.M.V.), ecacace@biol.ethz.ch (E.C.), emma.slack@immune.engineering (E.S.), ssunagawa@ethz.ch (S.S.), keyst@ethz.ch (T.G.K).

## Summary

Serotyping of bacteria using genomic information (*in silico* serotyping) has increasingly replaced serology. However, the *E. coli* capsule serotyping system has been largely abandoned since the 1990s, leaving gaps in our knowledge of capsule genetics, diversity, distribution, and epidemiology. To address this, we established a definitive genotype-serotype map for 35 serologically identified and structurally characterized transporter-dependent capsules. We then surveyed >37,000 *E. coli* genomes, cataloging 85 transporter-dependent capsule types (K-types), including 55 novel ones. We leveraged this catalog to develop an *in silico* serotyping tool, kTYPr, and applied it to curated sets of >25,000 *E. coli* genomes and metagenome-assembled genomes spanning diverse environmental and clinical sources. We found novel K-types enriched in under-sampled environments and associated with *E. coli* disease. This research expands our understanding of *E. coli* surface structures, supporting efforts for precision targeting with phage therapy or vaccines.

**Keywords (max10)** Capsular polysaccharide, K antigen, *E. coli*, serotype, HMM, AlphaFold, CAZy, glycosyltransferase, *in silico* serotyping, sequencing.

## Introduction

Antigens exposed on the bacterial surface are vulnerable to recognition by phages and host immune factors, driving extensive structural variation across bacterial populations. While the repertoire of glycans making up the O antigen of lipopolysaccharide, as well as large surface proteins such as adhesins and flagellin are well-studied,[1,2] we have little information on the repertoire of capsular polysaccharides produced by *E. coli.*[3] Capsular polysaccharides are considered virulence factors, owing to their ability to inhibit antibody binding to all other surface structures (including O antigens and proteins), as well as to block the action of innate immune mechanisms, e.g. opsonophagocytosis[4–6] As the dominant exposed structure on the *E. coli* surface, capsular polysaccharides are therefore critical binding targets for protective antibodies and therapeutic bacteriophage, or even potential drug-targets.[5,7]

The constituents, linkages, and chemical modifications that make up capsular polysaccharides are highly diverse. During the last century, classical serology was used to identify 80 different capsule serotypes in *E. coli*.[8] However, capsular polysaccharides are only weakly immunogenic, likely due to weak association to the bacterial surface.[9,10] Correspondingly, antisera generated by vaccination with whole bacteria are dominated by other antibody specificities that cannot be completely removed by cross-absorption and agglutination is non-informative.[8] Instead, counter-current immunoelectrophoresis (a precipitin-line-based approach) is required for capsule detection, which is slow, labor-intensive and expensive.[11] With the advent of PCR-based *E. coli* typing of O- and H- antigens, research into capsule serology stalled, and the most recent new capsule serotype was described in 1977.[12]

*E. coli's* capsular polysaccharides can be divided into four groups, based on their biosynthesis machinery and genetics.[13] Group 1 and 4 capsules are known as Wzx/Wzy-dependent polysaccharides, these are synthesized in a manner analogous to O antigens. In contrast, group 2 and 3 capsules are ABC-transporter-dependent. These latter capsules are polymerized onto a phosphatidylglycerol oligo-β-3-deoxy-d-*manno*-octulosonic acid (Kdo) structure in the inner-leaflet of the inner-membrane and exported intact to the cell surface.[14–17] The proteins required for oligo-β-Kdo synthesis (encoded by *kpsC* and *kpsS*)[14,16,17] and capsule export (encoded by *kpsE*, *kpsD, kpsM,* and, *kpsT*) are well-conserved (Fig. 1A).[18] K-type-specific genes extend the conserved oligo-β-Kdo, producing the serotype-specific part of the K antigen. Within this context, a dogma established in the field, based largely on knowledge of a few specific capsule serotypes such as K1 (associated with neonatal meningitis), states that group 2 and 3 capsules dominate extraintestinal pathovars, while group 1 and 4 capsules dominate in intestinal pathovars of *E. coli.* However, the lack of appropriate experimental and computational tools to identify capsule serotypes have hindered a comprehensive evaluation of these associations.[19]

With the increasing accessibility of whole-genome sequencing technologies and the escalating antibiotics resistance crisis, there is an urgent need for reliable methods to identify capsular polysaccharides from genetic data. Diverse computational approaches for *E. coli* serotype mapping from genomic data have been proposed, relying either on BLAST-based search[20–22] or on more sensitive and robust Hidden Markov Models (HMMs).[23] However, the absence of a

comprehensive genotype-phenotype map limited the number of loci that could be linked with established capsule serotypes and prevented the assignment of novel K-types,[24–27] leaving a critical knowledge gap.

Here we report major progress in closing this gap for the transporter-dependent class of capsular polysaccharides. We sequenced the *E. coli* capsule reference strain collection,[8] confirming the transporter-dependent nature and establishing the genotype-to-serotype map for 35 known group 2 and 3 K antigens. We complemented this by cataloguing the diversity of capsule biosynthesis loci from 37,723 *E. coli* genomes, identifying 85 distinct gene clusters, including 55 putative novel transporter-dependent capsule types. We functionally annotated capsule biosynthesis genes by predicting protein structural similarities, allowing polysaccharide structural groups to be inferred for many novel K-types. For K-types with published polysaccharide structures, predicted gene functions aligned well with the expected biosynthetic requirements. However, we suggest revision of the K6 structure based on gene content and confirmative mass spectrometric analysis. The newly established catalog of genetically defined K-types was used to develop an HMM-based capsule typing tool, kTYPr. We demonstrated its utility by analyzing a globally distributed, non-redundant collection of 24,031 genomes and 2,763 stool-derived metagenome assembled genomes (MAGs). Our findings revealed a substantial overlap in K-type distribution between gut-colonizing and invasive *E. coli* strains, and uncovered associations between novel K-types and both human disease and understudied environments and hosts. We envision that the updated genotype-serotype reference and bioinformatic tool presented here will enable epidemiologists, infectiologists, vaccinologists and microbiologists to explore the role of *E. coli* capsules in infection, transmission and protection from disease.

## Results

### A genotype-serotype map for established capsule types

To establish a definitive K antigen genotype-serotype map we recently sequenced the reference strain collection established at the World Health Organization's Collaborative Centre for Research and Reference on Escherichia.[8] (Roese Mores et al., Genome Announcements, 2025) Genetic information for four further transporter-dependent K serotypes was obtained from publicly available sequence databases (K15 and K74) or from strains from the Culture Collection at the University of Gothenburg (K2a and K22). We identified a *kps* locus, including all essential genes (*kpsEDCSMT,* Fig. 1A), in 35 of the 70 strains, confirming their designation as ABC-transporter-dependent capsule types (groups 2 and 3).[3,13,18,28] In contrast, these genes were not identified in strains designated as producers of Wzx/Wzy-dependent capsule types (groups 1 and 4).

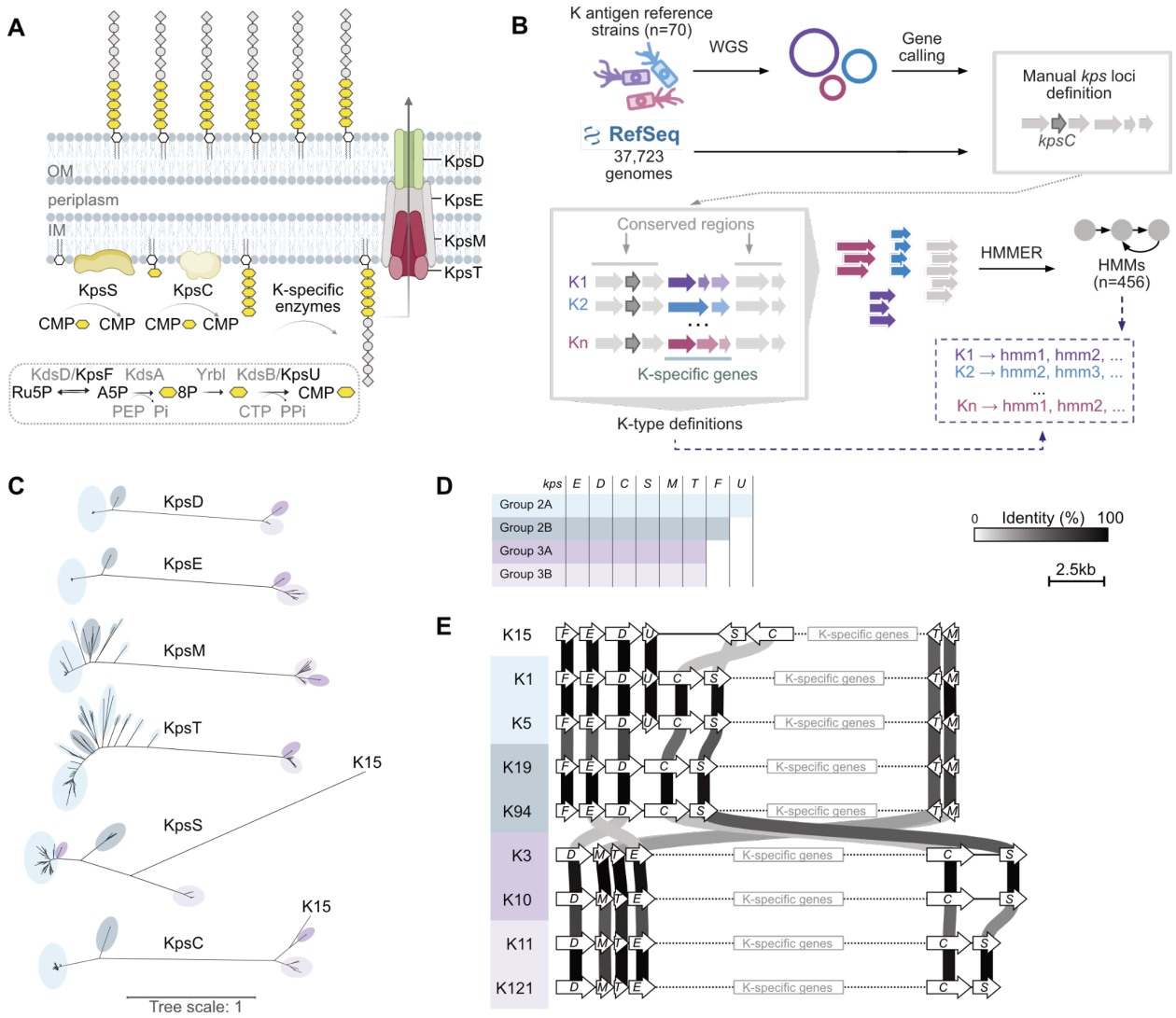



**Figure 1. Overview of capsule biosynthesis, cataloging procedure, and *kps* lineages.**

(A) Schematic representation of transporter-dependent capsule biosynthesis and export, highlighting the roles of the conserved Kps proteins.

(B) Construction of a comprehensive capsule gene cluster catalog by integrating K antigen reference strains (n=35) and RefSeq (n=37,723). The workflow includes whole genome sequencing (WGS) of reference strains, locus extraction, ORF identification, sequence clustering, and K-specific HMM generation (**STAR Methods**).

(C) Unrooted phylogenetic trees for the essential proteins KpsEDCSMT involved in capsule biogenesis (STAR Methods). Leaves are colored according to the assigned lineage 2A, 2B, 3A, and 3B as in (D).

(D) Presence/absence of conserved capsule biosynthesis genes in the four *kps* lineages. Colored bars indicate the presence of the corresponding gene.

(E) Arrangement of conserved genes that is characteristic of each lineage. Two exemplary gene clusters from each lineage and the exceptional K15 cluster are shown. White arrows representing the conserved genes are drawn to scale. K-specific genes are indicated within each cluster (not to scale). Vertical greyscale lines indicate the percentage identity between the respective protein products.

134 In some reference strains we observed identical or near-identical *kps* loci (Fig. S1). These
135 correspond to cross-reactive K serotypes,[8] that consist of identical polysaccharide backbones
136 with variable non-stoichiometric modifications.[29–34] For example, K2ab is an acetylated version of
137 the K2a backbone[32]. Accordingly, K2a and K2ab coding sequences differed by a single base
138 insertion disrupting a putative acetyltransferase gene in K2a (Fig. S1A). Similarly, two missense
139 mutations differentiate the candidate acetyltransferase genes in K13 and K23 (Fig. S1B),
140 potentially reflecting differing acetylation of the respective polysaccharides.[35] In some cases, the
141 genes relevant to serotype differentiation are likely to be found outside of the capsule gene cluster
142 (Fig. S1C-D).[36] We have taken a cautious approach to assigning these closely related K-types
143 where a robust genetic differentiation is not yet possible, grouping them under shared
144 designations as follows: K2a and K2ab are assigned K2; K13 and K23 are assigned K13_K23;
145 K18a, K18ab, and K22 are assigned K18_K22; K54 and K96 are assigned K96. Accordingly, we
146 define a total of 30 K-types corresponding to the 35 reference transporter-dependent serotypes.

147 **An extended catalog of transporter-dependent capsules**

148 Although previous studies have shown that there are more than the 35 established *kps* loci in *E.*
149 *coli* genomes,[7,21,36,37] the lack of a complete genotype-serotype map has prevented the clear
150 identification of novel K-types. With this map now established, we set out to systematically
151 characterize the diversity of transporter-dependent kps loci in *E. coli*.

152 To this end, we cataloged them in 37,723 *E. coli* genomes from the National Center for
153 Biotechnology Information Reference Sequence Database (NCBI RefSeq; Fig. 1B, STAR
154 Methods) by first using *kpsC* as a marker gene due to its role in biosynthesis of the unique oligo-
155 β-Kdo that is characteristic of the glycolipid anchor common to this class of capsules.[14,16,17] We
156 then dereplicated sequences 30 kb up- and down-stream of *kpsC*, filtered for the presence of
157 essential genes (*kpsEDCSMT*), and progressively trimmed the sequences to a presumed locus
158 based on the organization of conserved genes, functional annotations, and identification of
159 common flanking genes. The set of unique *kps* locus sequences was further refined using a
160 protein catalog (excluding transposase-related ORFs), with each set of closely related ORFs
161 (≥90% for both identity and coverage) used to build a protein HMM. In an iterative process, these
162 HMMs were used to identify and extract the *kps* locus from *kpsC* flanking sequences (±30 kb),
163 providing additional sequence diversity to build HMMs. A final set of 85 distinct *kps* loci was
164 defined based on unique gene content, making our HMM-based K-type definitions robust to
165 genetic rearrangements, and insertions or deletions in non-coding sequences. Comparison of the
166 catalog with the reference strains led us to assign 55 putative novel K-types, named between
167 K104 and K163, resulting in an updated catalog of 85 transporter-dependent K-types (STAR
168 Methods).

169 **Defining a fourth lineage of transporter-dependent capsules**

170 The historically designated group 2 and 3 capsules represent two distinct genetic lineages.[38] A
171 third group of plasmid-borne *kps* clusters (designated as 3B) was recently identified and proposed
172 as a recent acquisition in *E. coli.*[7,36] To investigate the evolutionary relationships across the 85
173 transporter-dependent K-types in our *kps* catalog, we constructed phylogenetic trees based on
174 the conserved capsule biosynthesis proteins encoded by the *kpsEDCSMT* genes (Fig. 1C). Trees
175 for the two most conserved proteins, the outer membrane pore (KpsD) and the periplasmic

adapter (KpsE), revealed four distinct lineages. In addition to the three known lineages, we identified a previously unrecognized group which we designated 2B. These four lineages are further differentiated by the presence of *kpsF* and *kpsU*, which encode enzymes involved in CMP-Kdo biosynthesis and are redundant in the *E. coli* genome (Fig. 1A, D), and by the organization of conserved genes (Fig. 1E). Phylogenetic trees of KpsC and KpsS share the KpsDE-architecture with two known exceptions; the *kpsS* genes in the 3A lineage appear to derive from the 2A lineage;[36] furthermore, KpsC and KpsS from K15 form distinct branches that are incongruent with the other lineages, suggesting a possible origin outside *E. coli.*[39,40] Phylogenetic trees of KpsM and KpsT revealed higher diversity and evidence of exchanges between the 2A and 2B lineages, consistent with frequent rearrangements of *kpsM*, *kpsT* and the K-specific genes in these lineages, as observed in K7, K51, K52, K104, K105, K106, and K108 (Fig. 2).

**Functional annotation of K-specific genes is supported by predictions of protein structures**

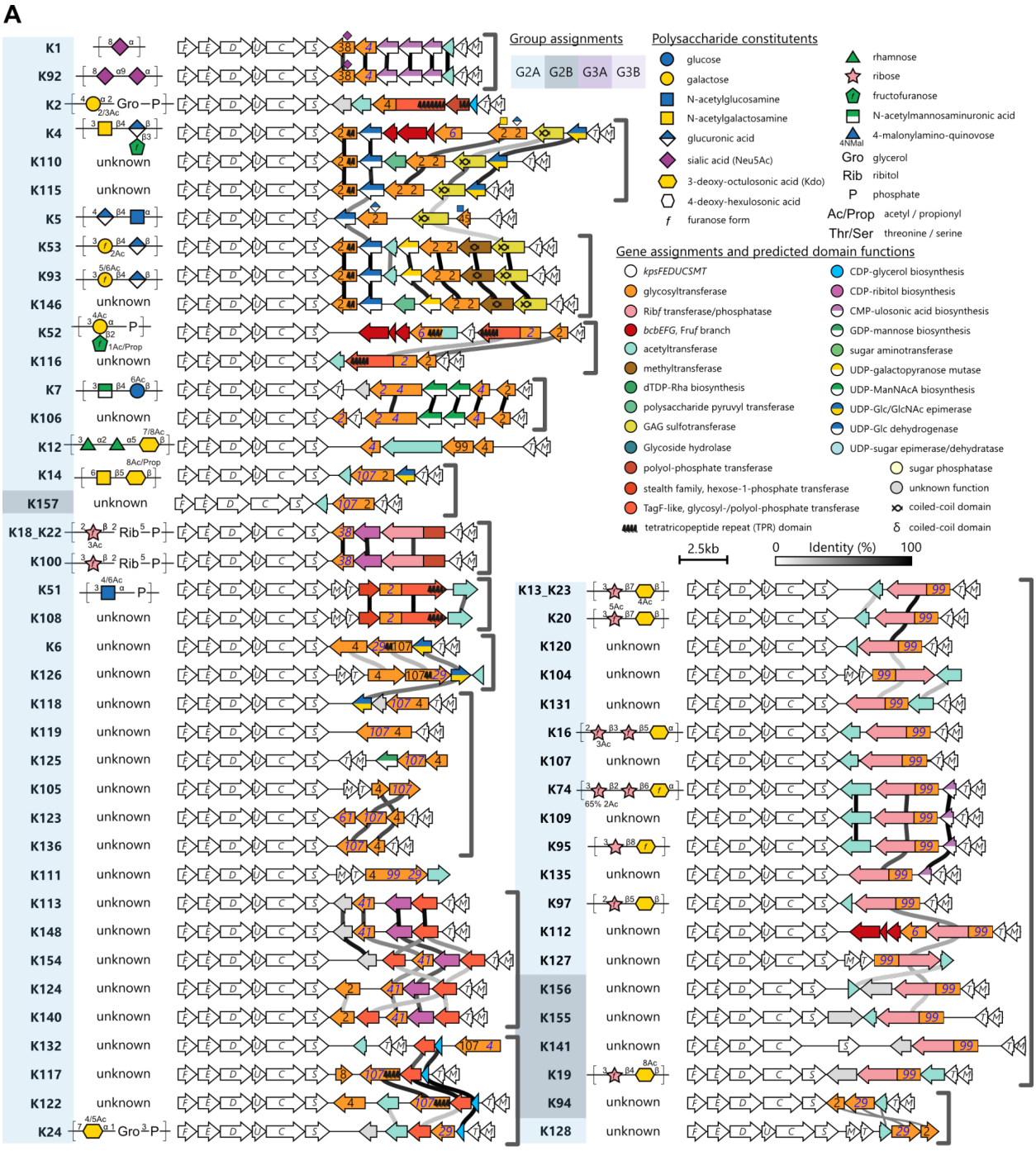
To explore the functional repertoire within the catalog, we functionally annotated serotype-specific genes. We identified conserved domains and glycosyltransferase (GT) families according to the Carbohydrate-Active enZYmes Database (CAZy, STAR Methods).[41–45] We also predicted structural similarities with proteins in the Protein Data Bank (PDB), using AlphaFold2[46] and Foldseek.[47] Altogether these analyses assigned putative functions to 421 of 456 ORFs (Fig. 2). Sequence similarity allowed us to assign a CAZY family to 50.5% (137 of 271) of the GT domains. The remaining GTs were assigned based on predicted structural similarity with characterized CAZy family members in the PDB, providing functional insights in the absence of clear sequence similarity. For example, 29 gene clusters encode a domain with predicted structural similarity to the related GT99 and GT107 families,[14] whose characterized members are retaining Kdo transferases (KdoT). This aligns with the presence of Kdo in the polysaccharide repeat units of clusters with known structures (namely, K13_K23, K14, K16, K19, K20, K74, K95 and K97), and suggests a role for GT99/107-like domains in catalyzing the respective Kdo linkage.

Grouping *kps* clusters with functionally related K-specific genes revealed 18 sets of K-types predicted to encode polysaccharides with related structural features, e.g. similar backbone composition (Fig. 2). Thirteen of these sets include at least one cluster with a known polysaccharide structure, facilitating structural inferences for novel polysaccharides. A striking example is the set of K-types (K13_K23 to K19, Fig. 2A) sharing a large ORF (>1100 amino acids) with little detectable sequence similarity but a consistent set of three domains identified by predicted structural similarity: a GT99-like KdoT, and a dual domain ribofuranosyltransferase/phosphatase, recently shown to be the biosynthetic origin of ribofuranose (Rib*f*) in bacterial polysaccharides.[48,49] These domains are consistent with Kdo- and Rib*f*-containing polysaccharides published for seven members of the group. This example illustrates the power of predicted structural similarity (as opposed to sequence similarity) to support functional annotation of carbohydrate active enzymes by identifying functional relationships obscured by sequence divergence, and to provide structural hypotheses for novel capsular polysaccharides.

216 **Structural analysis of polysaccharides clarifies inconsistencies between functional gene**
217 **content and published structures**

218 Comparison of the predicted functions in each gene cluster with published polysaccharide
219 structures revealed broadly consistent associations across the catalog (Fig. 2), with two notable
220 exceptions. In the first case, the K6 gene cluster lacks predicted Rib*f*T domains but is reported to
221 produce a Rib*f*-Rib*f*-Kdo repeat unit.[50] To clarify this inconsistency, we purified capsular
222 polysaccharide from seven reference strains reported to produce various Rib*f*-Kdo
223 polysaccharides and analyzed repeat unit composition by MALDI-MS of permethylated glycans
224 (STAR Methods, Fig. S2). Consistent with our expectations based on the gene clusters, we
225 observed Rib*f*-Kdo repeat units in all except the K6 polysaccharide. There we observed m/z
226 species consistent with a homopolymer of Kdo. Notably, a putative poly-Kdo structure is in
227 agreement with the presence of an ORF encoding a dual GT29-GT107 domain protein, both of
228 which are known for transfer of ulosonic acids, such as Kdo, from CMP-activated donor
229 substrates. We account for these results by assigning K6 as a serotype with an unknown
230 polysaccharide structure (Fig. 2) pending further structural investigation.

231 On the other hand, K16 is reported to produce a Rib*f*-Rib*f*-Kdo trisaccharide repeat unit[51] despite
232 having an identical set of predicted domains to several clusters reported to produce a Rib*f*-Kdo
233 disaccharide repeat (e.g. K13, K19, K20, etc., Fig. 2). MALDI-MS analysis of permethylated
234 glycans revealed the expected disaccharide repeats for K13, K19, K20, K23, and K97 and,
235 contrary to our expectations, the Rib*f*-Rib*f*-Kdo trisaccharide repeat unit was observed for K16
236 (Fig. S2). The presence of this trisaccharide repeat in K16 indicates a gap in our biochemical
237 understanding. It may be explained by either a bifunctional Rib*f*T capable of transferring Rib*f* to
238 both Kdo and Rib*f* acceptors, or the involvement of an additional Rib*f*T encoded outside of the
239 capsule gene cluster. Notably, the presence of single-domain bifunctional glycosyltransferases
240 involved in biosynthesis of other capsules favors the former hypothesis. These include a GlcNAc
241 transferase that transfers two consecutive GlcNAc moieties in forming the *Neisseria meningitidis*
242 serogroup L capsule[52] and the *E. coli* K92 polysialyltransferase which transfers successive sialic
243 acid residues with alternating linkages.[53]


A
Group assignments
G2A G2B G3A G3B
Polysaccharide constitutents
glucose
galactose
N-acetylglucosamine
N-acetylgalactosamine
glucuronic acid
sialic acid (Neu5Ac)
3-deoxy-octulosonic acid (Kdo)
4-deoxy-hexulosonic acid
f furanose form
rhamnose
ribose
fructofuranose
N-acetylmannosaminuronic acid
4NMal 4-malonylamino-quinovose
Gro glycerol
Rib ribitol
P phosphate
Ac/Prop acetyl / propionyl
Thr/Ser threonine / serine
Gene assignments and predicted domain functions
kpsFEDUCSMT
glycosyltransferase
Ribf transferase/phosphatase
bcbEFG, Fruf branch
acetyltransferase
methyltransferase
dTDP-Rha biosynthesis
polysaccharide pyruvyl transferase
GAG sulfotransferase
Glycoside hydrolase
polyol-phosphate transferase
stealth family, hexose-1-phosphate transferase
TagF-like, glycosyl-/polyol-phosphate transferase
tetratricopeptide repeat (TPR) domain
CDP-glycerol biosynthesis
CDP-ribitol biosynthesis
CMP-ulosonic acid biosynthesis
GDP-mannose biosynthesis
sugar aminotransferase
UDP-galactopyranose mutase
UDP-ManNAcA biosynthesis
UDP-Glc/GlcNAc epimerase
UDP-Glc dehydrogenase
UDP-sugar epimerase/dehydratase
sugar phosphatase
unknown function
coiled-coil domain
coiled-coil domain
2.5kb
0 Identity (%) 100
K1
K92
K2
K4
K110 unknown
K115 unknown
K5
K53
K93
K146 unknown
K52
K116 unknown
K7
K106 unknown
K12
K14
K157 unknown
K18_K22
K100
K51
K108 unknown
K6 unknown
K126 unknown
K118 unknown
K119 unknown
K125 unknown
K105 unknown
K123 unknown
K136 unknown
K111 unknown
K113 unknown
K148 unknown
K154 unknown
K124 unknown
K140 unknown
K132 unknown
K117 unknown
K122 unknown
K24
K13_K23
K20
K120 unknown
K104 unknown
K131 unknown
K16
K107 unknown
K74
K109 unknown
K95
K135 unknown
K97
K112 unknown
K127 unknown
K156 unknown
K155 unknown
K141 unknown
K19
K94 unknown
K128 unknown

244

245

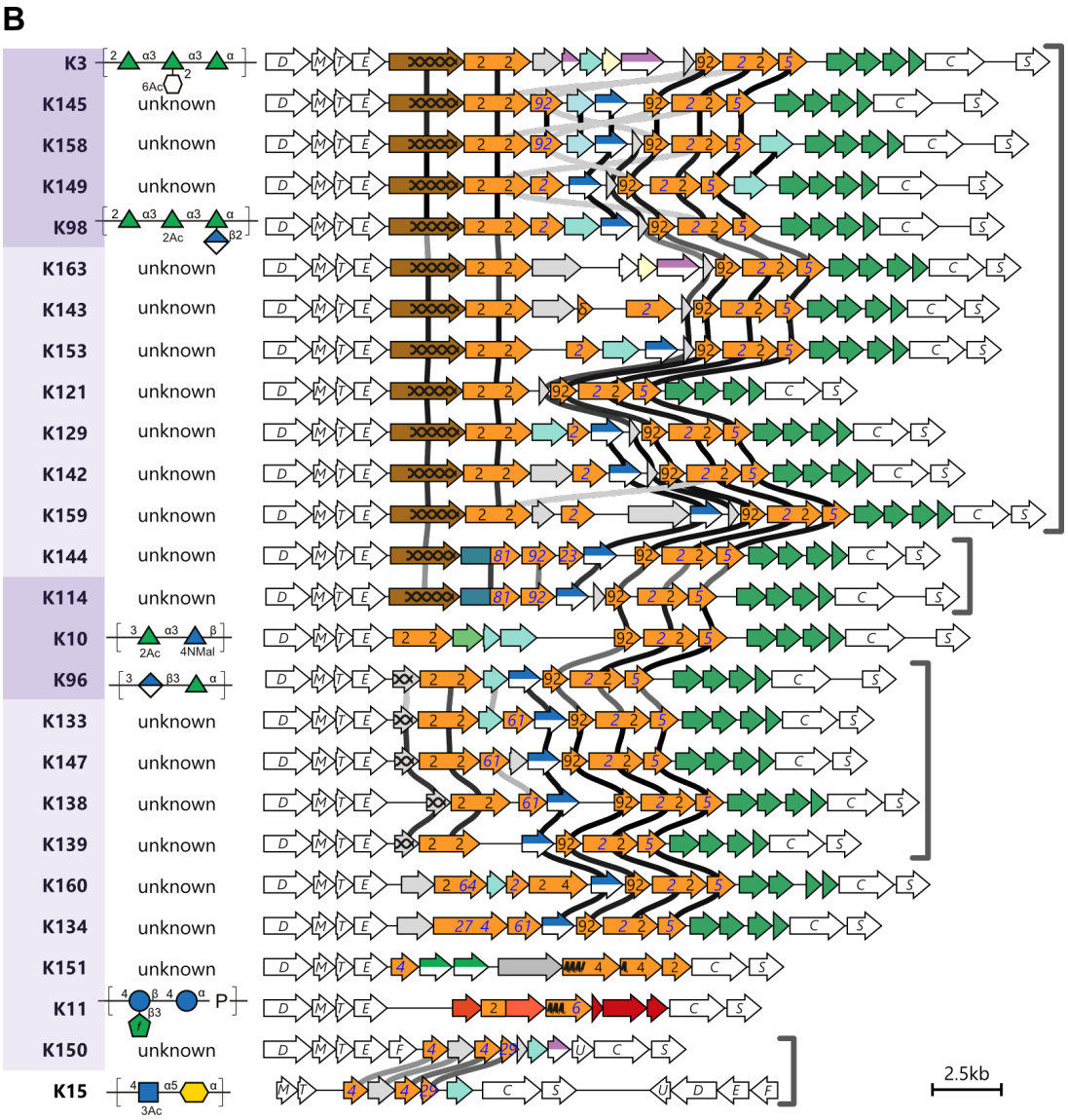



**Figure 2. Catalog of *E. coli* transporter-dependent capsule gene clusters.** The conserved biosynthetic and export genes are white. K-specific genes are colored according to known or predicted function. For characterized glycosyltransferases (GT), the transferred sugar is displayed above the gene. The CAZy family for GT domains are indicated in black numbers on the gene. For putative GT domains identified by predicted structural similarity, the CAZy family of the closest structural match is indicated in blue italics on the gene. Pairwise polypeptide identities >30% between vertically adjacent K-specific genes are indicated by grey-scale connecting lines. For clarity, sequence similarity between conserved genes is not shown. Gene clusters are organized to collect groups with functionally related gene content, indicated by grey brackets to the right of gene clusters. Polysaccharide repeat units, where known, are depicted according to SNFG conventions.[54]

(A) Gene clusters assigned to lineages 2A and 2B.

(B) Gene clusters assigned to lineages 3A, 3B and K15.

## Development of kTYPr as a genotype-to-serotype profiler in *E. coli*

261 We used the newly assembled *kps* catalog and HMMs to develop kTYPr, an *in silico* capsule
262 typing software applicable to both genomic and metagenomic data (Fig. 3, STAR Methods). The
263 tool screens genomes for *kpsC* gene presence, and processes *kps*-positive ones with the full set
264 of HMMs to identify conserved and K-type specific genes that meet HMM-specific bit score cutoffs.
265 K-types are defined by the presence of all essential genes (*kpsEDCSMT*) and K-type specific
266 gene sets. In cases where criteria for multiple K-types are met, the decision is based on the
267 highest accumulated bit scores. This allows the tool to distinguish between closely related gene
268 clusters (e.g. K51 and K108) and to avoid incorrectly assigning K-types that are defined by a
269 subset of genes that are also present in other K-types (e.g. K121 is a subset of K153 and K129,
270 see Fig. 2). Bit score values are reported for all K-types considered, together with the complete
271 set of hits for every gene against our HMM collection for further inspection or parametrization, a
272 genbank file of the extracted *kps*-related genes, and a visualization of its similarity to the predicted
273 or closest K-type cluster using clinker[55] (Fig. 3A).

274 We evaluated the performance of kTYPr on the source dataset (37,723 *E. coli* genomes from
275 RefSeq). Of these, 25.5% (n=9,607) were assigned a transporter-dependent K-type based on the
276 presence of the complete set of essential genes (*kpsEDCSMT*) and a complete set of K-type-
277 specific genes **(**Fig. S3A). Of the 28,082 genomes that were not assigned a transporter-
278 dependent K-type using default settings (STAR Methods), 92.2% (n=25,879) lacked a complete
279 set of essential and K-specific genes. The remaining 5.1% (n=1,436) had a complete set of either
280 *kpsEDCSMT* or K-specific genes, but not both. These cases may represent disrupted gene
281 clusters, genes missing in draft genomes, or uncataloged *kps* cluster diversity. While all 85 K-
282 types were identified in the source dataset, 11 were only identified once or twice (Fig. S3B),
283 indicating that our catalog has deeply sampled the diversity present in RefSeq, but that RefSeq
284 has not sampled the diversity of *E. coli kps* loci present in nature.

285

286 **Novel K-types are more represented in understudied hosts and environments**

287 To investigate the distribution of group 2 and 3 capsules across environments, hosts and body
288 sites in humans, we used kTYPr to profile a collection of 32,043 metadata-curated, globally
289 distributed genomes from NCBI (STAR Methods). To reduce genomic redundancy due to the
290 isolation of clonal strains from the same sample, or repeated isolation of strains from the same
291 individual over time, we dereplicated all genomes coupling a 99% ANI cutoff with metadata on
292 sample origin (host type, sample type, geography, health state) and bacterial characteristics
293 (sequence type, O-, H- and K-type), resulting in 23,206 representative genomes. We further
294 integrated this dataset with a collection of 825 *E. coli* genomes from urinary tract infection (UTI)
295 cases, 259 of which were associated with invasive disease (STAR Methods).[56]

296 The resulting collection (n=24,031) was biased towards human hosts (48.2%, Fig. 3B) and
297 westernized geographical areas (49.7% of genomes from Europe and North America, Fig. S4A).
298 33.5% of the genomes (n=8,062) contained a complete *kps* locus, with a higher proportion in
299 humans (47.7%), pets (37.6%) and human-associated environments (e.g. houses, public
300 transportation, sewage water and hospitals, 39.9%) (Fig. 3B). Most of the assigned transporter-
301 dependent capsules belonged to group 2 (86% of genomes, n=6,958), of which only 25 genomes
302 belonged to group 2B. 1,104 genomes (13.7%) harboured group 3 capsules, of which 609

genomes of group 3A and 495 of group 3B (Fig. 3C). Humans, human-associated environments and pets showed highly correlated K-type profiles (Fig. 3D, S4B-C), whereas livestock and wild animals were more similar, in agreement with previous reports of genetically distinct *E. coli* lineages between humans and livestock.[57] Animal hosts that are well-represented in the collection correlate according their K-type profiles, beyond geographical groups (Fig. S4C).

While 78.7% of the K-types in highly studied human hosts and associated environments were previously reported, the new K-types identified in this study were more frequent in less studied environments, such as non-domesticated animals (Fig. 3D-E). Novel K-types were also more diverse in terms of hosts and geography than known ones (Fig. 3F). The most frequent K-types overall were K1, K5, K2 and K14 (Fig. S4D). This ranking reflected the bias of the collection towards human samples and, except for the generally most abundant K1 type, was not always conserved in other environmental and geographical niches (Fig. S4E): for example K105 and K109 (newly defined types) were first and second most frequent in food (13.4%) and wild animals (7.1%), respectively (Fig. S4B). Overall, our approach unveiled a much higher diversity of transporter-dependent capsules than previously known from highly sampled, human-associated environments.

**Diversity of transporter-dependent capsules in human health and disease**

Consistent with previous reports,[58] 78% of *E. coli* genomes from phylogroups B2, D, and F harboured group 2 K-types (Fig. S5A). This was largely driven by human samples (Fig. 4A), where these phylogroups have been associated with extraintestinal pathogenic *E. coli* (ExPEC) (Fig. S5B).[59] Because most sequenced *E. coli* genomes are from clinical isolates, we have limited understanding of the prevalence and distribution of K-types in asymptomatic individuals. To address this gap, we additionally profiled a globally-distributed collection of *E. coli* metagenome-assembled genomes (MAGs) from the stool of healthy humans from 25 published studies (n=2,763 of which 921 *kps*-positive, Fig. S5C).[60]

Although group 2 capsules were also more frequent in MAGs belonging to phylogroups B2, D and F (Fig. S5D), their proportion was higher in extraintestinal disease (e.g. isolates from blood or cerebrospinal fluid) (Fig. 4B, S5D). Phylogroups B2, D and F had also the highest K-type diversity, higher in invasive disease than in asymptomatic carriage, except for phylogroup B2-2 (Fig. S5E-G). K-type profiles were more similar among MAGs and genomes from healthy individuals, patients with diseases not associated with *E. coli* and patients with intestinal pathogenic *E. coli* (InPEC), while genomes from urinary tract infections (UTIs), bloodstream infections (BSIs) and meningitis showed distinct separation (Fig. 4C). Transporter-dependent capsules were less common in InPEC (14%) compared to extraintestinal disease (67.3% in UTIs, 70.5% in BSI, 91.1% for meningitis), and to stool-derived MAGs (33.3%) or genomes (35.7%) from healthy individuals (Fig. S5H). Altogether, while consistent with previous reports of group 2 K-types in ExPEC,[61,62] our analysis revealed the extensive presence of typical ExPEC K-types (e.g. K1, K2, K5)[19,63,64] in the gut microbiome of asymptomatic individuals by overcoming previous limitations of isolate-based reference databases.

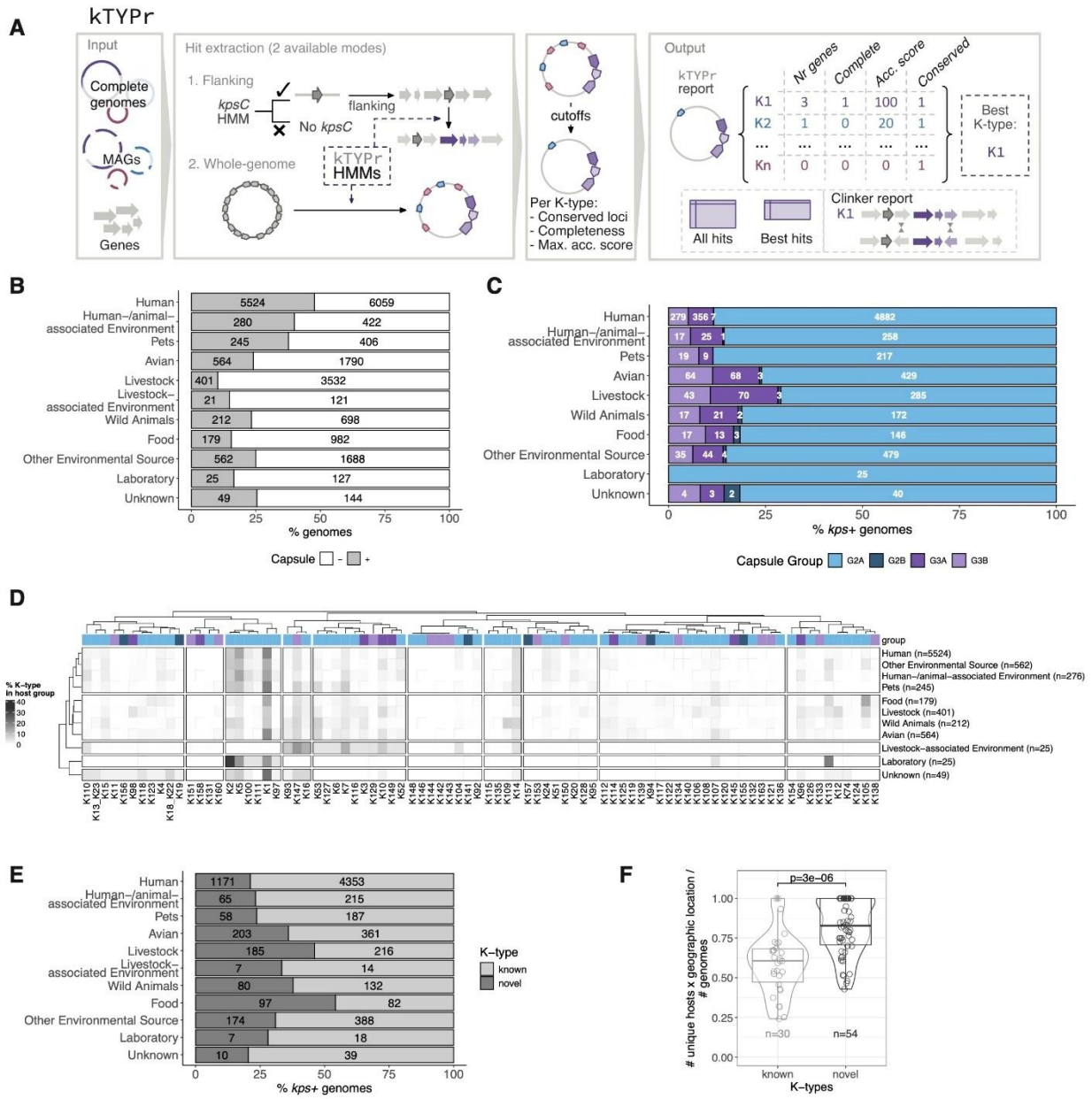



**Figure 3. kTYPr workflow and K-type diversity across hosts and environments**.

(A) Schematic representation of the kTYPr tool, highlighting allowed inputs (complete genomes, MAGs or genes) and two hit extraction modes (flanking and whole-genome). Once the custom cutoffs are applied, the hits are processed to produce a report output including for each input genome the accumulated bit-scores and completeness metrics for all K-types profiled, with the "best match" as the most complete K-type with highest accumulated bit-score. In addition, all gene hits against the HMM collection, a formatted genbank and a clinker visualization of the similarity between the genome and the best hit(s) are also reported (STAR Methods).

(B) Fraction of capsule-positive genomes (defined as containing a complete *kps* locus, STAR Methods) across hosts and environments. Absolute counts are indicated within the bars.

(C) Capsule group proportion across hosts and environments. Only capsule-positive genomes are considered. Absolute counts are indicated within the bars.

(D) K-type proportion in hosts and environments. Only capsule-positive genomes were considered. K-types and hosts are clustered according to Spearman correlation. K-type groups are color-coded as in Fig. 3C.

(E) Proportion of novel and known K-types across hosts and environments. Only capsule-positive genomes are considered.

(F) Number of distinct hosts and geographic locations (divided by number of genomes) for each K-type. The p-value between known and novel K-types is shown (two-sided Wilcoxon test); box limits correspond to first and third quartiles, with the median marked, and whiskers extending to the most extreme data points up to 1.5 times the interquartile range (IQR). Data point counts are indicated below the boxplots.

Furthermore, in agreement with the low frequency of transporter-dependent capsules in InPECs (Fig. S5H), we did not detect transporter-dependent K-types in genomes associated with the typical InPEC serotypes O157:H7 (enterohemorrhagic *E. coli*, EHEC) and O104:H4 (enteroaggregative *E. coli*, EAEC) (Fig. S6A). We note that these strains may harbour group 1 or 4 capsules that are not detected by kTYPr. On the other hand, typical ExPEC serotypes, such as O25:H4 and O1:H7, had the highest number of distinct transporter-dependent K-types and were among the O-types with highest proportion of capsule positivity (Fig. S6A). O- and K antigen co-occurrence may also reveal functional interdependencies between glycan biosynthetic gene clusters: for example, K12 polysaccharide contains rhamnose.[65] However, since we could not find rhamnose-biosynthesis genes in the *kps* cluster as for other rhamnose-containing K-types (Fig. 2), we hypothesized that these K-types might harness the rhamnose-biosynthesis machinery of the O antigen. Consistent with this hypothesis, K12 was exclusively found with rhamnose-containing O-types (O4, O139, O51, O1, O132, O13/O135, O18, O25) (Fig. S6B). Overall, this demonstrates how kTYPr can inform new hypotheses on functional relationships between different bacterial surface sugar antigens with potential relevance for human disease.

Finally, to quantify the association of K-types with invasiveness compared to asymptomatic carriage, we used a generalized linear model, with the invasive phenotype (defined as isolation from blood or cerebrospinal fluid in patients with bloodstream infection or meningitis) as binary outcome, K-type, O-type, H-type, sequence type (ST), phylogroup, host age, gender, and geographical location as fixed effects (STAR Methods). UTI cases were excluded to rule out possible contamination from non-pathogenic *E. coli* from the gut. The model confirmed previously proposed associations, including K1, K2, K4 and K5.[19,63,64] The K-types with the strongest association with invasiveness were K4, K12, K52 and K13_K23. Of the 20 K-types that were significantly and positively associated with invasiveness, 10 were previously reported (of which five only recently).[37] Of the 10 types that we reported for the first time with invasive potential, six were among those newly defined in our study: K104, K107, K111, K112, K116 and K147 (Fig. 4D), highlighting novel contributors to invasiveness that should be investigated in larger datasets and mechanistic studies.

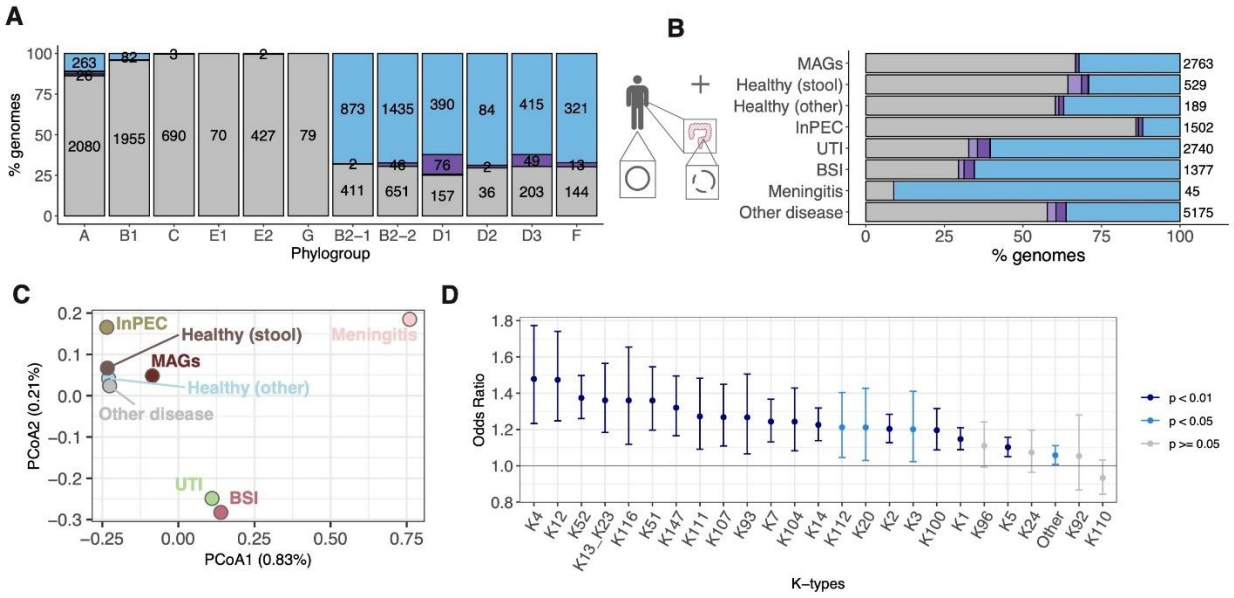



**Figure 4. K-type diversity in human health and disease**.

(A) Proportion of capsule groups across phylogroups in *E. coli* complete genomes from humans (n=11,537). Groups are color-coded as in Fig. 3C, with grey indicating *kps*-negative genomes.

(B) Proportion of capsule groups in *E. coli* genomes from humans across clinical categories and in asymptomatic carriers (MAGs from stool of healthy individuals). Groups are color-coded as in Fig. 3C. Non-*E. coli*-associated diseases are grouped as “other disease”. InPEC, Intestinal Pathogenic *E. coli*; UTI, Urinary Tract Infection; BSI, bloodstream infection.

(C) Principal Coordinates Analysis (PCoA) of K-type profiles from the clinical categories in 4B. The PCoA was based on a Bray-Curtis dissimilarity matrix calculated from the proportion of K-types in each group.

(D) Odds ratio (OR) and confidence interval (CI) from a generalized linear model evaluating the contribution of K-types to invasiveness. A dataset consisting of asymptomatic carriage (2,763 *E. coli* MAGs from healthy individuals, and 773 complete genomes from stool of healthy individuals) and invasive *E. coli*-associated disease (1377 genomes of which 1,181 genomes of isolates from blood or cerebrospinal fluid from NCBI and 259 from blood from a published study on urosepsis)[56] was used (STAR Methods). OR=1 is indicated (dashed line). Data points are color-coded according to statistical significance (STAR Methods).

## Discussion

Capsular polysaccharides have historically been classified through serology and bacteriophage sensitivity, but these approaches have inherent limitations in resolution, scalability and reproducibility.[8,11,66,67] Our work advances the field by integrating genomic classification, linking defined sets of genes to both known and putative novel serotypes. Based on a ground-truth database including all established *E. coli* K antigens, and an extended catalog of *kps* clusters assembled from >37’000 *E. coli* genomes, we provide a high-throughput, scalable and robust genome-driven approach to studying *E. coli* transporter-dependent capsules.

417 Our study significantly expands the known diversity of *E. coli* transporter-dependent capsules.
418 Our catalog of 85 K-types, including 55 putative novel types (Fig. 2) confirms the idea that *E. coli*
419 capsule diversity extends far beyond the set collected by serologists in the 20th century.[7,19,21,36,37]
420 Phylogenetic analysis of the conserved capsule biosynthesis machinery clarifies the evolutionary
421 relationships underlying the group 2 and 3 classifications (Fig. 1C-E). Our work also revealed a
422 previously unrecognized lineage, group 2B, characterized by the absence of *kpsU*. We confirm
423 and expand the group 3B lineage, first identified by Hong et al.[7,36] as plasmid-borne *kps* clusters.
424 While we have not investigated the presence of *kps* clusters on mobile genetic elements, our work
425 provides a method to explore this in future studies.

426 A key contribution of this study is the comprehensive characterization of the *kps* cluster catalog,
427 assigning putative functions to more than 90% of K-specific ORFs (Fig. 2). An earlier study of
428 TagF-like domains in capsule polymerases showed that sequences with as little as 15% identity
429 shared fold and function (sugar-phosphate and polyol-phosphate transfer).[68] Building on this, we
430 integrated structure prediction and similarity searches into our annotation pipeline, almost
431 doubling the number of glycosyltransferase domains identified and associated with a GT family in
432 the CAZy database. As our MALDI-MS results for K6 and K16 demonstrate (Fig. S2), this
433 functional and structural map can be used to inform mechanistic hypotheses on capsule
434 biosynthesis, to highlight gaps in our understanding of polysaccharide structure and biochemistry,
435 and to infer structures for novel K-types.

436 To open this resource for the research and public health communities, we developed kTYPr, an
437 HMM-based tool for robust and scalable determination of K-types from genome sequences.
438 Unlike earlier BLAST-based approaches,[21,37] kTYPr uses gene content-based K-type definitions
439 and sensitive HMMs for more accurate detection. These advances allow kTYPr to recognize gene
440 clusters that are highly divergent, reorganized, distributed across different genomic loci, or
441 fragmented in MAGs. The critical resource that sets kTYPr apart from other available tools is our
442 ground-truth dataset generated by sequencing the available reference strains. This makes kTYPr
443 the first tool that accounts for all 35 known transporter-dependent K-types (corresponding to 30
444 distinct gene clusters) and differentiates 55 putative novel types.

445 A limitation is that kTYPr's accuracy is inherently constrained by our understanding of the genetic
446 and biochemical determinants of capsule structure. For example, K18 and K22 share identical
447 capsule gene clusters, yet differ in acetylation of the polysaccharide (Fig. S1C). As the
448 responsible acetyltransferase gene is not located within the capsule locus, and remains
449 unidentified, we currently lack the knowledge to distinguish these serotypes using genomic data
450 alone. Similar ambiguities arise in other cases, such as K2a/K2ab and K13/K23. These examples
451 highlight the need to advance our understanding of capsule biosynthesis, and to develop methods
452 that increase the speed and throughput of structural analysis of bacterial polysaccharides.
453 Integration of these tools with genomic data will be essential to future advances in genome-driven
454 prediction of bacterial surface structures.

455 This work illustrates the power of kTYPr, applied to a unique collection for global distribution,
456 metadata resolution and genome dereplication, revealing type-resolved capsule ecology. We
457 found transporter-dependent capsules across diverse hosts and environments (Fig. 3B and S4C),
458 but the distribution of capsule lineages and types differs substantially between them (Fig. 3C-D).
459 The 30 known types were on average less ecologically diverse than 55 novel ones, which were

more often found in undersampled hosts, such as wild animals (Fig. 3E-F). For example, half of the K-types identified in livestock and food have not been previously described and the newly described K105 and K109 were the most frequently observed types after K1 in livestock and wild animals, respectively (Fig. 3D). Our initial screen on available strains with associated metadata hints at associations between transporter-dependent capsules and hosts and potentially geographical location (Fig. S4C, E), which should be further investigated with larger datasets representing undersampled hosts such as wild animals.

The inclusion of *E. coli* MAGs from healthy individuals as carriage control revealed previously unexplored associations of *E. coli* capsules with human health and disease. An important finding is the broadly similar distribution of K-types among *E. coli* isolates from the healthy gut and from clinically infected human patients (Fig. 4C, S5F). These results confirm, at unprecedented scale, the observations of serologists that ExPEC-associated serotypes are good colonizers of the intestinal tract.[69] These observations also support recent analyses suggesting that opportunistic infection of vulnerable hosts might explain a large portion of ExPEC-associated K-type diversity.[37]

Overall, we presented here a scalable, fast and flexible tool to type transporter-dependent capsules, kTYPr, and introduced three new resources: a well-characterized catalog of known and novel *kps* loci based on a ground-truth genotype-serotype map for all 35 reference transporter-dependent capsule types, and two curated, globally distributed, and dereplicated collections of *E. coli* genomes and MAGs, which can be used to revisit our current knowledge of *E. coli* ecology and genetic diversity at unprecedented scale. Our results enable structural inference for novel capsule types, highlight key gaps in our knowledge of polysaccharide structure and biochemistry, and outline the diversity of capsules across ecological niches, including novel K-types that are associated with understudied environments and invasive disease. This study provides a foundation for future work on *E. coli* glycobiology, ecology, evolution, epidemiology and for the prevention of ExPEC infections with capsule-based vaccines,[58] and treatment of infections with capsule-dependent monoclonal antibodies[70] or phages.[71]

## Acknowledgments

This work was supported by funding from the Basel Research Centre for Child Health (BRCCH) to E.S. and S.S., the Swiss National Science Foundation through the project grants 40B2-0_180953 and 310030_185128 to E.S. and the NCCR Microbiomes (51NF40_180575 and 51NF40_225148) to E.S. and S.S, and an Innosuisse grant (117.143 IP-LS) to T.G.K and E.S., and core funding from ETH Zürich. E.S. acknowledges support of an European Research Council Consolidator Grant (865730). E.S. and S.S. are supported by the LOOP Zurich mTORUS project. S.M.V. acknowledges funding from the Human Frontier Science Program (HFSP) through a postdoctoral fellowship [LT0050/2023-L]. E.C. acknowledges support from SNSF Spark (CRSK-3_228959) and Novartis Freenovation (FN24-0000000703). AlphaFold runs were performed on the ETH Euler cluster. We acknowledge the support of the IT Service and HPC facilities of ETH Zürich.

Conceptualization (E.S., S.S., T.G.K.), Data curation (C.R.M., E.C.), Formal analysis (C.R.M., S.M.V., E.C., T.G.K.), Funding acquisition (E.S., S.S., T.G.K.), Investigation (C.R.M., S.M.V., E.C., C.R., K.V., C-W.L., E.C.B, E.M., R.S., T.F., M.S., T.G.K.), Methodology (C.R.M., S.M.V., E.C., D.B., T.G.K.), Software (D.J.B., S.M.V.), Resources (H.J.R., A.C., A.E.), Supervision (E.S., S.S., T.G.K.), Validation (S.M.V., E.C., T.G.K.), Visualization (C.R.M., S.M.V., E.C., T.G.K.), Writing – original draft (C.R.M., S.M.V., E.C., T.G.K.), Writing – review & editing (C.R.M., S.M.V., E.C., E.S., S.S., T.G.K.). A.E. was supported by an unrestricted research grant by the University of Zurich.

## Declaration of interests

C.R., E.S., and T.G.K. are listed as inventors on a patent describing methods and reagents for preparation of capsule targeting vaccines. C.R. and T.G.K. are founders and shareholders of Baxiva AG.

## Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used ChatGPT to improve readability of the manuscript. The authors reviewed and edited all content and take full responsibility for the content of the published manuscript.

## Supplementary Figures

**Figure S1. Closely related capsule gene clusters identified in different K serotype reference strains do not support robust genetic differentiation of the respective serotypes, related to Figure 2.** Capsule gene clusters were compared and rendered using clinker. The proportion of identical amino acids is displayed numerically and in grey-scale between opposing ORFs. The reported base identities are from global Needle-Wunsch alignments and the protein alignments were made with the Smith-Waterman algorithm. Gene assignments and functional predictions were made as described in STAR Methods. Polysaccharides are depicted according to the conventions of the Symbol Nomenclature for Graphical Representation of Glycans and references for the displayed polysaccharide structures are provided in. (A) K2a and K2ab differ by acetylation of galactose in the repeat unit, consistent with a frameshift mutation in the putative acetyltransferase ORF in K2a. (B) The K13 and K23 polysaccharides differ by acetylation of the Kdo residue in the repeat unit and the candidate acetyltransferases differ by two missense mutations. Although the observed mutations may explain the K2a/K2ab and K13/K23 serotypes, they do not provide a robust basis for genetically distinguishing these serotypes. Therefore, these serotypes are grouped under the names K2 and K13_K23, respectively, in our catalog. (C) The K18 polysaccharide is an acetylated version of the K22 backbone, but the capsule gene clusters are identical (100% identity), and there is no candidate acetyltransferase gene within the clusters. The serological distinction between K18a and K18ab is unclear because chemical analysis of the polysaccharides revealed identical repeat unit structures. As there is currently no genetic basis for distinguishing these serotypes, they are grouped under the assignment K18_K22 in our catalog. (D) Our analysis of the K54 and K96 gene clusters is consistent with earlier reports suggesting that the gene(s) responsible for threonine modification of the polysaccharide backbone are located outside of the kps locus. Pending identification of these gene(s), the K54 and K96 clusters are grouped under the assignment K96 in the current version of the catalog.

542 **Figure S2. Mass spectrometry clarifies K antigen structures, related to Figure 2.** (A) Capsule gene
543 clusters from K serotype reference strains were compared and rendered with clinker, then colored
544 according to gene assignments and predicted domain functions. Amino acid identity > 30% is displayed in
545 greyscale between opposing genes. (B) Published polysaccharide repeat unit structures for each K
546 serotype. (C) Deconvoluted MALDI-TOF-TOF spectra of partially hydrolysed and permethylated capsular
547 polysaccharide (STAR Methods). Peak labels indicate the assigned structure and corresponding expected
548 m/z species calculated in GlycoWorkbench 2. We note that O-acetyl groups are expected to be lost during
549 the permethylation procedure and were not considered when assigning peaks.

550 **Figure S3. Evaluation of kTYPr on the RefSeq source data set.** (A) Upset plot of kTYPr results on
551 37,723 *E. coli* genomes from RefSeq, showing the distribution of *E. coli* genomes with complete and
552 incomplete capsule biosynthesis loci. (B) Absolute counts of K-types identified in the RefSeq data set.

553 **Figure S4. Global diversity of K-types, related to Figure 3.** (A) Fraction of capsule-positive genomes
554 (defined as containing a complete *kps* locus, STAR Methods) across geography. (B) K-type proportion in
555 hosts and environments. K-types with frequency < 5% are annotated as "other". (C) K-type proportion in
556 individual hosts and environments. K-types and sample origin are clustered according to Pearson's
557 correlation. Host-environment combinations with < 15 occurrences were not considered. K-type groups are
558 color-coded as in Fig. 3C. Hosts are color-coded according to their lower resolution groups shown in Fig.
559 3B. (D) Proportion and absolute counts of K-types, color-coded according to groups as in Fig. 3C. K-types
560 with proportion < 0.5% were annotated as "other". Only *kps*-positive genomes were considered. (E) K-type
561 proportion across geographic regions, grouped by hosts and environments. K-types with frequency < 8%
562 are annotated as "other". Only *kps*-positive genomes were considered.

563 **Figure S5. K-type associations with phylogroups and human health and disease, related to Figure**
564 **4.** (A) Proportion of capsule groups across phylogroups in *E. coli* genomes from all hosts and environments.
565 Groups are color-coded as in Fig. 3C. (B) *E. coli* human isolate source in the collection, grouped by
566 phylogroup (n=11,537). 20 genomes to which no phylogroup could be assigned were removed. (C) Fraction
567 of capsule-positive genomes (defined as containing a complete *kps* locus, STAR Methods) in *E. coli* human
568 gut metagenomes. (D) Capsule group proportion in each phylogroup in asymptomatic carriage (A,
569 corresponding to 2,762 *E. coli* MAGs from healthy individuals) and invasive *E. coli*-associated disease (I,
570 1,178 genomes of isolates from blood or cerebrospinal fluid from NCBI and 259 from blood from a published
571 study on urosepsis), STAR Methods. ns $p > 0.05$; * $p \leq 0.05$; ** $p \leq 0.01$; *** $p \leq 0.001$ (Fisher's exact test).
572 (E-G) K-type diversity (expressed as Simpson index, E), richness (F) and evenness (G) of phylogroups.
573 The same data as in D) were considered, filtering out phylogroups with ≤ 25 genomes in each group
574 (asymptomatic, A or invasive, I). The number of capsule-positive genomes in the two groups is indicated
575 above each boxplot. Box limits correspond to first and third quartiles, with the median marked, and whiskers
576 extending to the most extreme data points up to 1.5 times the interquartile range (IQR). (H) K-type
577 proportion across health groups defined as in Fig. 4B, considering all (left) or only capsule-positive genomes
578 (right). K-types with proportion < 1% (left) or < 3% (right) are grouped as "other".

579 **Figure S6. K-type associations with O and H antigens.** (A) Proportion of *kps*-positive genomes (top) and
580 the number of unique K-types (bottom) per O-type. H-type composition is indicated for each O-type
581 (bottom). Only O-types with > 25 genomes are shown. H-types with proportion < 30% are grouped as
582 "Other". (B) O-type proportion for each K-type. K- and O-negative cases are not shown. Rare O-types (total
583 proportion across all K-types ≤ 0.032, as first quartile of distribution) are not shown. K-type groups are color-
584 coded as in Fig. 3C. O-type sugar composition was obtained from ECODAB.[2,72] Only unique sugars are
585 listed. K12 co-occurrence with rhamnose-containing O-types is highlighted.

586 **Figure S7. Distribution of ANI scores across genome comparisons.** For 778,526,070 all-vs-all ANI
587 comparisons, we show the frequency distribution (normalized per million pairs) using 1,500 bins for all the

588 ANI values (X-axis). The 98% threshold was selected based on the separation of the two main modes of
589 the distribution.

590

591 ## STAR Methods

592 ### Sequencing and genome assembly for K antigen reference strains

593 Strain acquisition, growth, DNA extraction, sequencing, and assembly of genomes is described
594 (Roese Mores et al., Genome Announcements, 2025).

595

596 ### Developing a catalog of capsule gene clusters

597 To survey the diversity of ABC transporter-dependent capsules in *E. coli*, we cataloged the *kps*
598 gene clusters in a collection of 37,732 *E. coli* genomes available from RefSeq (downloaded 19th
599 May 2024). We used *kpsC* as a marker gene due to its large size and central role in capsule
600 biosynthesis.[7,14,17] *kpsC* positive genomes were identified with BLASTN (version 2.16.0)[73,74] with
601 the parameters `-max-target-seqs 100000 -max_hsps 1`, using *kpsC* genes from K1, K3, K11,
602 K15, and K19 reference clusters as queries. We extracted the *kpsC* flanking region +/-30 kb from
603 11,834 genomes. 130 sequences where *kpsC* occurred on a contig of <8 kb were excluded. The
604 remaining sequences were dereplicated at 90% identity and coverage using `mmseqs cluster`
605 algorithm (MMseqs2 Version 15-6f452)[75] with the parameters `–min-seq-id 0.98 -cov 0.9`
606 `cov-mode 0 -s 7.0`. The unique representatives (n=2,822) were annotated using DFAST Web
607 API[76] with the parameter `dataset=ecoli` because of its sensitivity for detection of
608 *kpsFEDUCSMT* genes. Filtering for the presence of all essential genes (*kpsEDCSMT*) yielded
609 1911 unique sequences with a potential *kps* locus. To account for K-specific genes often, but not
610 always, being flanked by the common *kps* gene, we trimmed each sequence to an approximate
611 cluster; including a minimum of 2 ORFs up- and downstream to the *kps* start and end coordinates,
612 or 10 ORFs up- and downstream in cases where no candidate K-specific ORFs were identified
613 within the bounds of *kpsFEDUCSMT*. Dereplication of the trimmed sequences using `mmseqs`
614 `cluster` with `-min-seq-id 0.98` yielded 292 representative sequences. Each sequence was
615 visually inspected to define the locus start and end coordinates, considering the organization of
616 *kpsFEDUCSMT*, functional annotations (DFAST), and identification of common flanking genes.
617 Dereplication of the manually trimmed sequences at 98% identity yielded a set of 190 unique *kps*
618 locus sequences.

619 We further refined the capsule locus catalog based on gene content. To do this, we first generated
620 a global protein catalog by clustering all serotype-specific ORFs (excluding transposase-related
621 sequences) at 90% amino acid identity and coverage. Each resulting cluster was aligned using
622 MUSCLE (v3.8.1511)[77] and used to build a protein HMM (HMMER 3.4)[78]. In an iterative process,
623 these HMMs were applied to automate identification of locus boundaries within each *kpsC*
624 flanking region +/-30 kb. This approach ensured the inclusion of functionally related genes that
625 may have been missed during initial manual locus curation, and it allowed us to merge loci with
626 the same gene content. In later iterations, we transitioned to K-type specific protein HMMs, built
627 exclusively with sequences assigned to that K-type by analysis with kTYPr. Exceptions to this rule
628 are proteins with a common function across multiple K-types, i.e. RfbBDAC, NeuDBACE, and

629 KpsFEDUCSMT. Because closely related gene clusters can produce serologically and chemically
630 distinct polysaccharides – such as K1 and K92, which differ by a single ORF with 83% protein
631 identity – we used a threshold of 90% identity across all ORFs for defining distinct K-types. This
632 led us to a final set of 85 K-types defined by a set of 456 protein HMMs.

633

634 **Phylogenetic analysis of conserved *kps* genes**

635 Phylogenetic relationships among the conserved *kpsEDCSTM* genes were inferred to support the
636 classification of *kps* loci into distinct groups. Protein sequences corresponding to each conserved
637 gene (*kpsD, kpsE, kpsC, kpsS, kpsT*, and *kpsM*) were individually aligned using MAFFT
638 (v7.305)[79] with default parameters. Phylogenetic trees were constructed using FastTree
639 (v2.1.10)[80] under the JTT+CAT model for maximum-likelihood inference. The resulting
640 phylogenies were visualized using iTOL v7[81].

641

642 **Definition and development of kTYPr**

643 The kTYPr tool developed and utilized in this study is fully written in Python (v≥3.9) and can be
644 downloaded and installed following the details in https://github.com/SushiLab/kTYPr. This
645 bioinformatic tool accepts two types of input in FASTA format: genomes or user-provided
646 translated gene sequences. If a genome is provided, whether a complete genome or a
647 metagenome-assembled genome, the analysis begins with gene prediction using *Pyrodigal*
648 (v3.4.1) and its find_genes function[82,83]. Once a set of predicted protein sequences is obtained,
649 kTYPr employs *PyHMMER* (v0.10.11) [84,85] to scan for the presence of *kpsC* within the sequences.
650 If a *kpsC* hit is detected, the tool extracts a ±30 kb region flanking the start and end coordinates
651 of the hit. At this stage, no thresholds or best-hit filtering are applied, meaning that if multiple *kpsC*
652 genes are present, regardless of score, all identified *kpsC* candidates (if multiple) and flanking
653 genes are retained for further analysis.

654 The extracted subset of genes is profiled against our custom catalog of 456 HMMs, each linked
655 to a known K-type, generating an initial kTYPr output table. This table, called 'all hits', includes all
656 identified genes, their corresponding HMM hits, and key metrics such as bit scores. HMM hits are
657 then filtered according to custom bit score thresholds. By default, thresholds are set to 60% of the
658 maximum bit score observed for each HMM in the RefSeq collection. In a handful of cases the
659 thresholds were decreased to accommodate highly polymorphic genes where the observed
660 variation is known (e.g. the K5 putative sulfotransferase, *kfiB*[86]), or not expected to be critical for
661 K antigen production (e.g. variation in length of the coiled-coil regions in the putative
662 methyltransferase genes present in many group 3 clusters). On the other hand, the thresholds for
663 dTDP-rhamnose biosynthesis genes (*rfbABCD*) were increased to limit false positive hits to
664 homologs present in many *E. coli* O antigen biosynthesis gene clusters. After applying these
665 thresholds, the filtered set of hits is analyzed to determine the presence or absence of the
666 conserved region, based on identification of *kpsEDCSMT* and *kpsFU* genes (assigned a value of
667 1 if all genes are present, 0 otherwise). The same strategy is applied to each K-specific gene set
668 associated with different K-types. The final output table reports the presence of the conserved
669 region and, for each K-type, whether the cluster is complete, along with its accumulated bit score.

670 Additionally, a separate set of columns designates the best K-type, prioritizing first the
671 completeness of the cluster and then the highest accumulated bit score. This approach allows
672 users to explore multiple or nearly complete clusters while still identifying the most probable gene-
673 based K-type. The outputs are a detailed kTYPr report, genbank file for the extracted cluster, and
674 an HTML file with comparison of the query cluster to the predicted or closest match K-type cluster.

675 Alternatively, to enable kTYPr to run on incomplete genomes, we have implemented a 'whole-
676 genome' mode as an alternative to the 'flanking mode' described above. In this mode, the initial
677 step of selecting neighboring genes around *kpsC* is skipped, and instead, every gene in the
678 genome is evaluated throughout the process. Furthermore, the described bioinformatics process
679 operates at the single-genome level; however, kTYPr also supports a batch mode (see
680 https://github.com/SushiLab/kTYPr). This mode allows users to provide a list of genome file paths,
681 enabling the same analytical steps to be applied across an entire genome collection. Additionally,
682 multiprocessing capabilities are available to enhance performance, ensuring efficient processing
683 of large datasets. In this mode, the final output consists of a comprehensive table that
684 consolidates predictions for the entire collection, streamlining downstream comparative analyses.

685 All analyses presented in this study were performed using kTYPr in 'flanking mode', considering
686 the custom definition of curated hit cutoffs, and in batch mode.

687

688 **Definition of reference gene clusters**

689 For each of the 85 defined K-types, we assigned a reference gene cluster. These were assigned
690 first based on kTYPr results, using the cluster with the highest accumulated bit score for each K-
691 type. The assigned reference was controlled by clinker comparison with all clusters (or a
692 representative set of all clusters identified using MMSeqs2 with a minimum sequence identity
693 threshold of 90%) to ensure that the gene organization was typical of the majority of clusters
694 assigned to the given K-type. We note that the reference gene cluster is never taken from the
695 corresponding K antigen reference strain, though in some cases they may be identical.

696 The reference gene clusters were compared in clinker, and K-specific ORFs extracted from the
697 reference gene clusters were used for functional annotation of the gene clusters (**Fig. 2**).

698

699 **Functional annotation of ORFs**

700 Functional annotation of serotype-specific ORFs was performed using InterProScan (v5.62-
701 94.0)[87] and dbCAN3[42] with default parameters. InterProScan result columns 'orf_id', 'length',
702 'domain_start', 'domain_end', 'DB', 'DB_hit', 'DB_hit_description', 'e_value', 'gene_ontology' were
703 taken to group results by ORF identifier while preserving the information by the database were
704 the hit was retrieved. This tab-delimited output and the standard output from EggNOG were then
705 combined by ORF identifier for further comparative and analysis.

706 Protein structure predictions were performed using ColabFold (v1.3.0).[88] The AlphaFold2 version
707 used within ColabFold was v2.1.14.[46] Multiple sequence alignments (MSAs) were generated
708 using MMseqs2.[89] To maximize sequence coverage, searches were conducted against the

709 UniRef30[90] and ColabFold environmental[91] databases. MSAs were obtained using the
710 colabfold_search command with default parameters, specifying both databases for alignment.

711 Protein structure predictions were then performed using the precomputed MSAs with the
712 colabfold_batch function. Five AlphaFold2 models were generated per query, with a maximum of
713 three recycles. Model confidence was ranked based on predicted local distance difference test
714 (pLDDT) scores, and structures were retained according to this ranking. Template-based
715 modeling and Amber refinement were not used. The models were generated using an ensemble
716 size of one, and the pairing mode was set to "unpaired+paired" to optimize MSA depth. The
717 MMseqs2 search mode included both UniRef30 and environmental databases to improve
718 structural predictions.

719 Serotype-specific ORFs were clustered according to sequence and structural similarity using
720 MMseqs2[75] and Foldseek[47], respectively. Sequence-based clustering was conducted using
721 MMseqs2 (v14) with a minimum sequence identity threshold of 30% and 90% coverage. Clusters
722 were generated using single-linkage clustering, ensuring that closely related sequences were
723 grouped while maintaining sensitivity to divergent homologs. Structural homology clustering was
724 performed using Foldseek (v8), using a TM-score threshold of 0.6 and a minimum coverage of
725 90%. Additionally, ORFs were searched against the Protein Data Bank (PDB) [92] to identify
726 potential structural homologs using Foldseek (v8) and default parameters.

727 **Purification, hydrolysis, permethylation and MS analysis of capsular polysaccharides**

728 K antigen reference strains were cultivated to stationary phase in TB media[93] at 37°C shaking at
729 180 rpm in Erlenmeyer flasks. Total surface polysaccharides, including lipopolysaccharide (LPS)
730 and capsular polysaccharide (CPS), were extracted from approximately 25 g (wet weight) cell
731 pellets according to the hot phenol/water procedure (Westphal and Jann, 1965)[94] with slight
732 modifications as follows. Membrane lipids were extracted from the biomass once with 500 mL of
733 ethanol overnight with stirring, then twice with 500 mL of acetone for 2 h with stirring. The cell
734 mass was recovered after each extraction by centrifugation at 2,500 g, 20°C, 30 min, with slow
735 deceleration, supernatant was discarded. After extraction of membrane lipids, the cell mass was
736 air dried and weighed, typically yielding 3-5 g dry weight. Dry cell mass was crushed to a fine
737 powder with a spatula, then resuspended with 60 mL of 50 mM Tris-HCl pH 8.0 at 65°C under
738 vigorous stirring in teflon coated centrifuge buckets. 60 mL of preheated phenol at 65°C was
739 added to the bucket, vigorous stirring of the aqueous-phenol mix continued for 15 min at 65°C.
740 The aqueous phenol mix was then cooled in an ice bath for 10 minutes, then centrifuged for 45
741 min at 4°C with slow deceleration to reduce disruption of the separated phases. The upper
742 aqueous phase containing LPS and CPS was collected to a clean vessel. The remaining phenol
743 phase was extracted a second time with 60 mL of 50 mM Tris-HCl pH 8.0 using the same
744 procedure, and the upper aqueous phases were pooled. The pooled aqueous extract was
745 dialysed 5 times over 5 days with 40-fold volume of water using 14 kDa Spectra/Por 4 regenerated
746 cellulose membranes to remove phenol. The dialysed extract was lyophilized to dryness,
747 weighed, and resuspended at 50 mg/ml in 20 mM Tris-HCl pH 8.0 with 1 mM $MgCl_2$, and treated
748 with 500 Units of Benzonase for 1 h at 37°C. Lipid A and short LPS molecules were removed by
749 three successive extractions with Triton X-114 as follows. 2% by volume of triton X-114 was added
750 to ice cold samples then vortexed until completely dissolved. After 10 minutes on ice, samples
751 were warmed to 42°C in a water bath to trigger phase separation of triton X-114, then centrifuged

752 for 10 min at 20,000 g in a hot rotor. The aqueous supernatant was removed to a fresh tube,
753 taking care not to disturb the triton X-114 extracted material at the bottom of the tube. Capsular
754 polysaccharides were then selectively precipitated by adding 1 volume of 10%
755 cetryltrimethylammonium bromide (CTAB). The precipitated material was recovered by
756 centrifugation for 15 min at 4,500 g, and dissolving the pellet in 7 mL of 1 M NaCl. Polysaccharides
757 were then precipitated from 1 M NaCl by addition of 3 volumes (27 mL) of ice cold ethanol, and
758 centrifugation at 4°C and 8,000 g for 15 min. Supernatant was discarded and the pellet was
759 resuspended in 7 mL of water. Purified capsular polysaccharides were partially hydrolysed in 50
760 mM TFA using the following conditions:

| K antigens | Hydrolysis conditions |
|---|---|
| K6, K16 | 25°C overnight |
| K13, K20, K23, K97 | 40°C 1 h, then 25°C overnight |
| K19 | 60°C 1 h, then 25°C overnight |

765

766 Hydrolysates were neutralised by addition of 7 mL 50 mM NaOH, then adjusted to 50 mL total
767 volume with 20 mM Tris-HCl pH 7.5. Capsular polysaccharide fragments were then loaded on a
768 10 mL Source 15Q column, washed with 20 mM Tris pH 7.5, and eluted in a step gradient up to
769 1 M NaCl. The eluted polysaccharides were dialysed against water, then lyophilized to dryness
770 and resuspended in water.

771 Polysaccharide fragments were permethylated to enhance glycan ionization efficiency. For each
772 sample, 50 µL (50 ug of polysaccharide) was dried in a Teflon-lined screw-capped glass tube. A
773 slurry was prepared by grinding NaOH pellets in dry DMSO, and about 0.5 mL was then added
774 to the sample, followed by about 0.5 mL methyl iodide. The mixture was shaken vigorously for 20
775 minutes at room temperature. The reaction was quenched by the dropwise addition of water,
776 followed by extraction with 2 mL chloroform. The chloroform layer was washed repeatedly with
777 water until clear, then dried under a stream of nitrogen. Permethylated glycans were dissolved in
778 20 µL of pure acetonitrile, and 1 µL was mixed with 1 µL of 2,5-dihydroxybenzoic acid matrix (10
779 mg/mL) on the MALDI-target plate. Samples were analyzed using a Bruker Rapiflex MALDI-TOF-
780 TOF mass spectrometer in positive ion mode.

781

782 **Genome collection preprocessing and annotation**

783 For the NCBI collection, we collected 39,460 genomes reported as *E. coli* from NCBI (October
784 11th, 2024) and removed all deprecated ones and duplicated accession numbers, leading to
785 32,043 genomes. O- and H-types were annotated using the tool ECTyper (v1.0) [1] and sequence
786 types were determined using the MLST tool (v2.23) available at
787 https://github.com/tseemann/mlst. The genomes were classified into phylogroups using a
788 published Mash distance-based approach [59]. For phylogroup C, an alternative reference genome
789 was selected (GCF_001515725.1) from the Microreact database, based on its classification under
790 multiple criteria and highest sequence scores [56]. Each queried genome was assigned to the
791 closest reference phylogroup using a Mash v2.3 distance cutoff of 0.04 [95]. Metadata available

from BV-BRC[96] was further manually curated. Since NCBI genomes are biased towards hosts and environments (e.g. clinical isolates) and include time-series from the same source, we dereplicated all genomes according to a 98% ANI cutoff using the tool skani (v0.2.2) [97]. The command `skani triangle -l genome_list.txt --fast` was used to generate an ANI percentage matrix by performing all-against-all genome comparisons within the collection (with genome file paths specified in the input text file). The fast mode was selected to enhance scalability, given that results for genomes with high N50 and >95% expected ANI, as in the case of this study, are comparable to those obtained using the standard mode [97]. This matrix was then processed to define clusters of genomes with at least 98% ANI. This threshold was determined based on the distribution of ANI values across the matrix, which exhibited a bimodal pattern, with the 98th percentile marking the boundary between the two modes (Fig. S7). Coupling this information with metadata on sample origin, we considered as distinct genomes with distinct 98% ANI clustering,geographic origin, host type, gender, age and health state when available and unique in sequence type, O-, H- and K-type. This resulted in 22846 distinct entries which we used for all downstream, considering as K-positive genomes fulfilling the following criteria: (i) a complete set of conserved and variable genes in the *kps* locus; (ii) *kps* clusters present on the same contig, which led to removing K137, K152, K161, K162.

**Statistical inference on strain invasiveness**

To assess K-type contribution to the invasive phenotype, the NCBI collection was filtered to include only human hosts and samples annotated as “Bloodstream Infection,” “Meningitis,” “Neonatal Bacteremia” and isolated from blood and cerebrospinal fluid (n=1118), or “Healthy” (n=773). This dataset was further integrated with 259 samples from bloodstream infections (Cuénod et al., 2023) and with publicly available, metadata-curated MAGs from the stool of healthy humans (n=2,763).[60] A generalized linear model was fitted using the R package stats v4.5.0, with invasiveness encoded as binary outcome, and K-type, O-type, phylogroup, ST, host gender, age and geography as fixed effects. To prevent sparsity and improve model stability, categorical variables with fewer than 15 (for O-types) or 10 occurrences (for ST and H-types) were grouped as “Other”. To prevent complete separation, K-types with ≤ 5 occurrences in both groups (invasive and not) were grouped as “Other”. Fixed effect estimates were extracted and odds ratios (ORs) with 95% confidence intervals (CIs) were computed by exponentiating the model estimates. All analyses were conducted in R (v4.5.0).

## References


1. Bessonov, K., Laing, C., Robertson, J., Yong, I., Ziebell, K., Gannon, V.P.J., Nichani, A., Arya, G., Nash, J.H.E., and Christianson, S. (2021). ECTyper: in silico Escherichia coli serotype and species prediction from raw and assembled whole-genome sequence data. Microb. Genomics *7*, 000728. https://doi.org/10.1099/mgen.0.000728.

2. Liu, B., Furevi, A., Perepelov, A.V., Guo, X., Cao, H., Wang, Q., Reeves, P.R., Knirel, Y.A., Wang, L., and Widmalm, G. (2020). Structure and genetics of *Escherichia coli* O antigens. FEMS Microbiol. Rev. *44*, 655–683. https://doi.org/10.1093/femsre/fuz028.

832 3. Whitfield, C., Wear, S.S., and Sande, C. (2020). Assembly of Bacterial Capsular
833 Polysaccharides and Exopolysaccharides. Annu. Rev. Microbiol. *74*, 521–543.
834 https://doi.org/10.1146/annurev-micro-011420-075607.

835 4. Gao, S., Jin, W., Quan, Y., Li, Y., Shen, Y., Yuan, S., Yi, L., Wang, Y., and Wang, Y. (2024).
836 Bacterial capsules: Occurrence, mechanism, and function. Npj Biofilms Microbiomes *10*, 1–
837 15. https://doi.org/10.1038/s41522-024-00497-6.

838 5. Mostowy, R.J., and Holt, K.E. (2018). Diversity-Generating Machines: Genetics of Bacterial
839 Sugar-Coating. Trends Microbiol. *26*, 1008–1021. https://doi.org/10.1016/j.tim.2018.06.006.

840 6. Chorro, L., Li, Z., Chu, L., Singh, S., Gu, J., Kim, J.H., Dutta, K., Pan, R., Kodali, S., Ndreu,
841 D., et al. (2022). Preclinical Immunogenicity and Efficacy of Optimized O25b O-Antigen
842 Glycoconjugates To Prevent MDR ST131 E. coli Infections. Infect. Immun. *90*.
843 https://doi.org/10.1128/iai.00022-22.

844 7. Hong, Y., Qin, J., Forga, X.B., and Totsika, M. (2023). Extensive Diversity in Escherichia coli
845 Group 3 Capsules Is Driven by Recombination and Plasmid Transfer from Multiple Species.
846 Microbiol. Spectr. *11*, e0143223. https://doi.org/10.1128/spectrum.01432-23.

847 8. Orskov, I., Orskov, F., Jann, B., and Jann, K. (1977). Serology, chemistry, and genetics of O
848 and K antigens of Escherichia coli. Bacteriol. Rev. *41*, 667–710.
849 https://doi.org/10.1128/br.41.3.667-710.1977.

850 9. Kauffmann, F. (1947). The Serology of the Coli Group. J. Immunol. *57*, 71–100.
851 https://doi.org/10.4049/jimmunol.57.1.71.

852 10. Snapper, C.M. (2016). Differential regulation of polysaccharide-specific antibody responses
853 to isolated polysaccharides, conjugate vaccines, and intact Gram-positive versus Gram-
854 negative extracellular bacteria. Vaccine *34*, 3542–3548.
855 https://doi.org/10.1016/j.vaccine.2015.12.077.

856 11. Semjen, G., Orskov, I., and Orskov, F. (1977). K antigen determination of Escherichia coli
857 by counter current immunoelectrophoresis (CIE). Acta Pathol. Microbiol. Scand. - Sect. B
858 Microbiol. Immunol. *85*, 103–107. https://doi.org/10.1111/j.1699-0463.1977.tb01680.x.

859 12. Guinée, P.A., Agterberg, C.M., Jansen, W.H., and Frik, J.F. (1977). Serological identification
860 of pig enterotoxigenic Escherichia coli strains not belonging to the classical serotypes.
861 Infect. Immun. *15*, 549–555. https://doi.org/10.1128/iai.15.2.549-555.1977.

862 13. Whitfield, C., and Roberts, I.S. (1999). Structure, assembly and regulation of expression of
863 capsules in Escherichia coli. Mol. Microbiol. *31*, 1307–1319. https://doi.org/10.1046/j.1365-
864 2958.1999.01276.x.

865 14. Doyle, L., Ovchinnikova, O.G., Myler, K., Mallette, E., Huang, B.-S., Lowary, T.L., Kimber,
866 M.S., and Whitfield, C. (2019). Biosynthesis of a conserved glycolipid anchor for Gram-
867 negative bacterial capsules. Nat. Chem. Biol. *15*, 632–640. https://doi.org/10.1038/s41589-
868 019-0276-8.

869 15. Ovchinnikova, O.G., Mallette, E., Koizumi, A., Lowary, T.L., Kimber, M.S., and Whitfield, C.
870 (2016). Bacterial β-Kdo glycosyltransferases represent a new glycosyltransferase family

(GT99). Proc. Natl. Acad. Sci. *113*, E3120–E3129. https://doi.org/10.1073/pnas.1603146113.

16. Willis, L.M., Stupak, J., Richards, M.R., Lowary, T.L., Li, J., and Whitfield, C. (2013). Conserved glycolipid termini in capsular polysaccharides synthesized by ATP-binding cassette transporter-dependent pathways in Gram-negative pathogens. Proc. Natl. Acad. Sci. *110*, 7868–7873. https://doi.org/10.1073/pnas.1222317110.

17. Willis, L.M., and Whitfield, C. (2013). KpsC and KpsS are retaining 3-deoxy-D-manno-oct-2-ulosonic acid (Kdo) transferases involved in synthesis of bacterial capsules. Proc. Natl. Acad. Sci. U. S. A. *110*, 20753–20758. https://doi.org/10.1073/pnas.1312637110.

18. Kuklewicz, J., and Zimmer, J. (2024). Molecular insights into capsular polysaccharide secretion. Nature *628*, 901–909. https://doi.org/10.1038/s41586-024-07248-9.

19. Arredondo-Alonso, S., Blundell-Hunter, G., Fu, Z., Gladstone, R.A., Fillol-Salom, A., Loraine, J., Cloutman-Green, E., Johnsen, P.J., Samuelsen, Ø., Pöntinen, A.K., et al. (2023). Evolutionary and functional history of the Escherichia coli K1 capsule. Nat. Commun. *14*, 3294. https://doi.org/10.1038/s41467-023-39052-w.

20. Gladstone, R.A., Pesonen, M., Pöntinen, A.K., Mäklin, T., MacAlasdair, N., Thorpe, H., Shao, Y., Mallawaarachchi, S., Arredondo-Alonso, S., Parcell, B.J., et al. (2024). Group 2 and 3 ABC-transporter dependant capsular K-loci contribute significantly to variation in the invasive potential of Escherichia coli. Preprint at medRxiv, https://doi.org/10.1101/2024.11.22.24317484 https://doi.org/10.1101/2024.11.22.24317484.

21. Holt, K.E., Lassalle, F., Wyres, K.L., Wick, R., and Mostowy, R.J. (2020). Diversity and evolution of surface polysaccharide synthesis loci in Enterobacteriales. ISME J. *14*, 1713–1730. https://doi.org/10.1038/s41396-020-0628-0.

22. Wyres, K.L., Wick, R.R., Gorrie, C., Jenney, A., Follador, R., Thomson, N.R., and Holt, K.E. (2016). Identification of Klebsiella capsule synthesis loci from whole genome data. Microb. Genomics *2*, e000102. https://doi.org/10.1099/mgen.0.000102.

23. Rendueles, O., Garcia-Garcerà, M., Néron, B., Touchon, M., and Rocha, E.P.C. (2017). Abundance and co-occurrence of extracellular capsules increase environmental breadth: Implications for the emergence of pathogens. PLOS Pathog. *13*, e1006525. https://doi.org/10.1371/journal.ppat.1006525.

24. Echarti, C., Hirschel, B., Boulnois, G.J., Varley, J.M., Waldvogel, F., and Timmis, K.N. (1983). Cloning and analysis of the K1 capsule biosynthesis genes of Escherichia coli: lack of homology with Neisseria meningitidis group B DNA sequences. Infect. Immun. *41*, 54–60. https://doi.org/10.1128/iai.41.1.54-60.1983.

25. Ninomiya, T., Sugiura, N., Tawada, A., Sugimoto, K., Watanabe, H., and Kimata, K. (2002). Molecular cloning and characterization of chondroitin polymerase from Escherichia coli strain K4. J. Biol. Chem. *277*, 21567–21575. https://doi.org/10.1074/jbc.M201719200.

26. Roberts, I., Mountford, R., High, N., Bitter-Suermann, D., Jann, K., Timmis, K., and Boulnois, G. (1986). Molecular cloning and analysis of genes for production of K5, K7, K12,

910 and K92 capsular polysaccharides in Escherichia coli. J. Bacteriol. *168*, 1228–1233.
911 https://doi.org/10.1128/jb.168.3.1228-1233.1986.

912 27. Yang, S., Xi, D., Jing, F., Kong, D., Wu, J., Feng, L., Cao, B., and Wang, L. (2018). Genetic
913 diversity of K-antigen gene clusters of Escherichia coli and their molecular typing using a
914 suspension array. Can. J. Microbiol. *64*, 231–241. https://doi.org/10.1139/cjm-2017-0620.

915 28. Kunduru, B.R., Nair, S.A., and Rathinavelan, T. (2016). EK3D: an E. coli K antigen 3-
916 dimensional structure database. Nucleic Acids Res. *44*, D675-681.
917 https://doi.org/10.1093/nar/gkv1313.

918 29. Hofmann, P., Jann, B., and Jann, K. (1985). Structure of the amino acid-containing capsular
919 polysaccharide (K54 antigen) from escherichia coli O6:K54:H10. Carbohydr. Res. *139*, 261–
920 271. https://doi.org/10.1016/0008-6215(85)90025-4.

921 30. Jann, B., Kochanowski, H., and Jann, K. (1994). Structure of the capsular K96
922 polysaccharide (K96 antigen) from Escherichia coli O77:K96:H- and comparison with the
923 capsular K54 polysaccharide (K54 antigen) from Escherichia coli O6:K54:H10. Carbohydr.
924 Res. *253*, 323–327.

925 31. Jann, K., Jann, B., Schmidt, M.A., and Vann, W.F. (1980). Structure of the Escherichia coli
926 K2 capsular antigen, a teichoic acid-like polymer. J. Bacteriol. *143*, 1108–1115.
927 https://doi.org/10.1128/jb.143.3.1108-1115.1980.

928 32. JANN, K., and Schmidt, M.A. (1980). COMPARATIVE CHEMICAL ANALYSIS OF TWO
929 VARIANTS OF THE *ESCHERICHIA COLI* K2 ANTIGEN. FEMS Microbiol. Lett. *7*, 79–81.
930 https://doi.org/10.1111/j.1574-6941.1980.tb01580.x.

931 33. Rodriguez, M.-L., Jann, B., and Jann, K. (1988). Comparative structural elucidation of the
932 K18, K22, and K100 antigens of Escherichia coli as related ribosyl-ribitol phosphates.
933 Carbohydr. Res. *173*, 243–253. https://doi.org/10.1016/S0008-6215(00)90820-6.

934 34. Vann, W.F., Soderstrom, T., Egan, W., Tsui, F.P., Schneerson, R., Orskov, I., and Orskov,
935 F. (1983). Serological, chemical, and structural analyses of the Escherichia coli cross-
936 reactive capsular polysaccharides K13, K20, and K23. Infect. Immun. *39*, 623–629.
937 https://doi.org/10.1128/iai.39.2.623-629.1983.

938 35. Vann, W.F., Soderstrom, T., Egan, W., Tsui, F.P., Schneerson, R., ØRskov, I., and ØRskov,
939 F. (1983). Serological, chemical, and structural analyses of the Escherichia coli cross-
940 reactive capsular polysaccharides K13, K20, and K23. Infect. Immun. *39*, 623–629.
941 https://doi.org/10.1128/iai.39.2.623-629.1983.

942 36. Hong, Y., Cunneen, M.M., and Reeves, P.R. (2019). Two extremely divergent sequence
943 forms of the genes that define Escherichia coli group 3 capsules suggest a very long history
944 since their common ancestor. FEMS Microbiol. Lett. *366*.
945 https://doi.org/10.1093/femsle/fnz091.

946 37. Gladstone, R.A., Pesonen, M., Pöntinen, A.K., Mäklin, T., MacAlasdair, N., Thorpe, H.,
947 Shao, Y., Mallawaarachchi, S., Arredondo-Alonso, S., Parcell, B.J., et al. (2024). Group 2
948 and 3 ABC-transporter dependant capsular K-loci contribute significantly to variation in the

949 invasive potential of Escherichia coli. medRxiv.
950 https://doi.org/10.1101/2024.11.22.24317484.

951 38. Sande, C., and Whitfield, C. (2021). Capsules and Extracellular Polysaccharides in
952 Escherichia coli and Salmonella. EcoSal Plus *9*. https://doi.org/10.1128/ecosalplus.esp-
953 0033-2020.

954 39. Azurmendi, H.F., Veeramachineni, V., Freese, S., Lichaa, F., Freedberg, D.I., and Vann,
955 W.F. (2020). Chemical structure and genetic organization of the E. coli O6:K15 capsular
956 polysaccharide. Sci. Rep. *10*, 12608. https://doi.org/10.1038/s41598-020-69476-z.

957 40. Schneider, G., Dobrindt, U., Brüggemann, H., Nagy, G., Janke, B., Blum-Oehler, G.,
958 Buchrieser, C., Gottschalk, G., Emödy, L., and Hacker, J. (2004). The Pathogenicity Island-
959 Associated K15 Capsule Determinant Exhibits a Novel Genetic Structure and Correlates
960 with Virulence in Uropathogenic *Escherichia coli* Strain 536. Infect. Immun. *72*, 5993–6001.
961 https://doi.org/10.1128/IAI.72.10.5993-6001.2004.

962 41. Zdobnov, E.M., and Apweiler, R. (2001). InterProScan – an integration platform for the
963 signature-recognition methods in InterPro. Bioinformatics *17*, 847–848.
964 https://doi.org/10.1093/bioinformatics/17.9.847.

965 42. Zheng, J., Ge, Q., Yan, Y., Zhang, X., Huang, L., and Yin, Y. (2023). dbCAN3: automated
966 carbohydrate-active enzyme and substrate annotation. Nucleic Acids Res. *51*, W115–W121.
967 https://doi.org/10.1093/nar/gkad328.

968 43. Coutinho, P.M., Deleury, E., Davies, G.J., and Henrissat, B. (2003). An evolving hierarchical
969 family classification for glycosyltransferases. J. Mol. Biol. *328*, 307–317.
970 https://doi.org/10.1016/s0022-2836(03)00307-3.

971 44. Campbell, J.A., Davies, G.J., Bulone, V., and Henrissat, B. (1997). A classification of
972 nucleotide-diphospho-sugar glycosyltransferases based on amino acid sequence
973 similarities. Biochem. J. *326 ( Pt 3)*, 929–939. https://doi.org/10.1042/bj3260929u.

974 45. Lombard, V., Ramulu, H.G., Drula, E., Coutinho, P.M., and Henrissat, B. (2014). The
975 carbohydrate-active enzymes database (CAZy) in 2013. Nucleic Acids Res. *42*, D490–
976 D495. https://doi.org/10.1093/nar/gkt1178.

977 46. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O.,
978 Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate
979 protein structure prediction with AlphaFold. Nature *596*, 583–589.
980 https://doi.org/10.1038/s41586-021-03819-2.

981 47. van Kempen, M., Kim, S.S., Tumescheit, C., Mirdita, M., Lee, J., Gilchrist, C.L.M., Söding,
982 J., and Steinegger, M. (2024). Fast and accurate protein structure search with Foldseek.
983 Nat. Biotechnol. *42*, 243–246. https://doi.org/10.1038/s41587-023-01773-0.

984 48. Cifuente, J.O., Schulze, J., Bethe, A., Di Domenico, V., Litschko, C., Budde, I., Eidenberger,
985 L., Thiesler, H., Ramón Roth, I., Berger, M., et al. (2023). A multi-enzyme machine
986 polymerizes the Haemophilus influenzae type b capsule. Nat. Chem. Biol.
987 https://doi.org/10.1038/s41589-023-01324-3.

988 49. Kelly, S.D., Williams, D.M., Nothof, J.T., Kim, T., Lowary, T.L., Kimber, M.S., and Whitfield,
989 C. (2022). The biosynthetic origin of ribofuranose in bacterial polysaccharides. Nat. Chem.
990 Biol. *18*, 530–537. https://doi.org/10.1038/s41589-022-01006-6.

991 50. J. Jennings, H., Rosell, K.-G., and G. Johnson, K. (1982). Structure of the 3-deoxy-d-
992 manno-octulosonic acid-containing polysaccharide (K6 antigen) from escherichia coli LP
993 1092. Carbohydr. Res. *105*, 45–56. https://doi.org/10.1016/S0008-6215(00)81853-4.

994 51. Lenter, M., Jann, B., and Jann, K. (1990). Structure of the K16 antigen from escherichia coli
995 O7:K16:H-, A Kdo-containing capsular polysaccharide. Carbohydr. Res. *197*, 197–204.
996 https://doi.org/10.1016/0008-6215(90)84142-H.

997 52. Litschko, C., Romano, M.R., Pinto, V., Claus, H., Vogel, U., Berti, F., Gerardy-Schahn, R.,
998 and Fiebig, T. (2015). The capsule polymerase CslB of Neisseria meningitidis serogroup L
999 catalyzes the synthesis of a complex trimeric repeating unit comprising glycosidic and
1000 phosphodiester linkages. J. Biol. Chem. *290*, 24355–24366.
1001 https://doi.org/10.1074/jbc.M115.678094.

1002 53. Keys, T.G., Fuchs, H.L.S., Galuska, S.P., Gerardy-Schahn, R., and Freiberger, F. (2013). A
1003 single amino acid toggles Escherichia coli polysialyltransferases between mono- and
1004 bifunctionality. Glycobiology *23*, 613–618. https://doi.org/10.1093/glycob/cwt003.

1005 54. Varki, A., Cummings, R.D., Aebi, M., Packer, N.H., Seeberger, P.H., Esko, J.D., Stanley, P.,
1006 Hart, G., Darvill, A., Kinoshita, T., et al. (2015). Symbol Nomenclature for Graphical
1007 Representations of Glycans. Glycobiology *25*, 1323–1324.
1008 https://doi.org/10.1093/glycob/cwv091.

1009 55. Gilchrist, C.L.M., and Chooi, Y.H. (2021). Clinker & clustermap.js: Automatic generation of
1010 gene cluster comparison figures. Bioinformatics *37*, 2473–2475.
1011 https://doi.org/10.1093/bioinformatics/btab007.

1012 56. Cuénod, A., Agnetti, J., Seth-Smith, H.M.B., Roloff, T., Wälchli, D., Shcherbakov, D.,
1013 Akbergenov, R., Tschudin-Sutter, S., Bassetti, S., Siegemund, M., et al. (2023). Bacterial
1014 genome-wide association study substantiates papGII of Escherichia coli as a major risk
1015 factor for urosepsis. Genome Med. *15*, 89. https://doi.org/10.1186/s13073-023-01243-x.

1016 57. Ludden, C., Raven, K.E., Jamrozy, D., Gouliouris, T., Blane, B., Coll, F., De Goffau, M.,
1017 Naydenova, P., Horner, C., Hernandez-Garcia, J., et al. (2019). One Health Genomic
1018 Surveillance of Escherichia coli Demonstrates Distinct Lineages and Mobile Genetic
1019 Elements in Isolates from Humans versus Livestock. mBio *10*, e02693-18.
1020 https://doi.org/10.1128/mBio.02693-18.

1021 58. Clark, J.R., and Maresso, A.M. (2021). Comparative Pathogenomics of *Escherichia coli* :
1022 Polyvalent Vaccine Target Identification through Virulome Analysis. Infect. Immun. *89*,
1023 e00115-21. https://doi.org/10.1128/IAI.00115-21.

1024 59. Abram, K., Udaondo, Z., Bleker, C., Wanchai, V., Wassenaar, T.M., Robeson, M.S., and
1025 Ussery, D.W. (2021). Mash-based analyses of Escherichia coli genomes reveal 14 distinct
1026 phylogroups. Commun. Biol. *4*, 117. https://doi.org/10.1038/s42003-020-01626-5.

1027 60. Dmitrijeva, M., Ruscheweyh, H.-J., Feer, L., Li, K., Miravet-Verde, S., Sintsova, A., Mende,
1028 D.R., Zeller, G., and Sunagawa, S. (2025). The mOTUs online database provides web-
1029 accessible genomic context to taxonomic profiling of microbial communities. Nucleic Acids
1030 Res. *53*, D797–D805. https://doi.org/10.1093/nar/gkae1004.

1031 61. Köhler, C.-D., and Dobrindt, U. (2011). What defines extraintestinal pathogenic Escherichia
1032 coli? Int. J. Med. Microbiol. *301*, 642–647. https://doi.org/10.1016/j.ijmm.2011.09.006.

1033 62. Hu, D., Fuller, N.R., Caterson, I.D., Holmes, A.J., and Reeves, P.R. (2022). Single-gene
1034 long-read sequencing illuminates Escherichia coli strain dynamics in the human intestinal
1035 microbiome. Cell Rep. *38*, 110239. https://doi.org/10.1016/j.celrep.2021.110239.

1036 63. Whitfield, C. (2006). Biosynthesis and assembly of capsular polysaccharides in Escherichia
1037 coli. Annu. Rev. Biochem. *75*, 39–68.

1038 64. Buckles, E.L., Wang, X., Lane, M.C., Lockatell, C.V., Johnson, D.E., Rasko, D.A., Mobley,
1039 H.L.T., and Donnenberg, M.S. (2009). Role of the K2 capsule in Escherichia coli urinary
1040 tract infection and serum resistance. J. Infect. Dis. *199*, 1689–1697.
1041 https://doi.org/10.1086/598524.

1042 65. Schmidt, M.A., and Jann, K. (1983). Structure of the 2-Keto-3-deoxy-d-manno-octonic-acid-
1043 Containing Capsular Polysaccharide (K12 Antigen) of the Urinary-Tract-Infective
1044 Escherichia coli O4:K12:H-. Eur. J. Biochem. *131*, 509–517. https://doi.org/10.1111/j.1432-
1045 1033.1983.tb07291.x.

1046 66. Gross, R.J., Cheasty, T., and Rowe, B. (1977). Isolation of bacteriophages specific for the
1047 K1 polysaccharide antigen of Escherichia coli. J. Clin. Microbiol. *6*, 548–550.
1048 https://doi.org/10.1128/jcm.6.6.548-550.1977.

1049 67. Gupta, D.S., Jann, B., Schmidt, G., Golecki, J.R., Ørskov, I., Ørskov, F., and Jann, K.
1050 (1982). Coliphage K5, specific for E. coli exhibiting the capsular K5 antigen. FEMS
1051 Microbiol. Lett. *14*, 75–78. https://doi.org/10.1111/j.1574-6968.1982.tb08638.x.

1052 68. Litschko, C., Oldrini, D., Budde, I., Berger, M., Meens, J., Gerardy-Schahn, R., Berti, F.,
1053 Schubert, M., and Fiebig, T. (2018). A new family of capsule polymerases generates
1054 teichoic acid-like capsule polymers in gram-negative pathogens. mBio *9*.
1055 https://doi.org/10.1128/mBio.00641-18.

1056 69. Orskov, F., and Orskov, I. (1992). Escherichia coli serotyping and disease in man and
1057 animals. Can. J. Microbiol. *38*, 699–704.

1058 70. Raff, H.V., Devereux, D., Shuford, W., Abbott-Brown, D., and Maloney, G. (1988). Human
1059 Monoclonal Antibody with Protective Activity for Escherichia coli K1 and Neisseria
1060 meningitidis Group B Infections. J. Infect. Dis. *157*, 118–126.
1061 https://doi.org/10.1093/infdis/157.1.118.

1062 71. Gaborieau, B., Vaysset, H., Tesson, F., Charachon, I., Dib, N., Bernier, J., Dequidt, T.,
1063 Georjon, H., Clermont, O., Hersen, P., et al. (2024). Prediction of strain level phage–host
1064 interactions across the Escherichia genus using only genomic information. Nat. Microbiol. *9*,
1065 2847–2861. https://doi.org/10.1038/s41564-024-01832-5.

1066 72. Rojas-Macias, M.A., Stahle, J., Lutteke, T., and Widmalm, G. (2015). Development of the
1067 ECODAB into a relational database for Escherichia coli O-antigens and other bacterial
1068 polysaccharides. Glycobiology *25*, 341–347. https://doi.org/10.1093/glycob/cwu116.

1069 73. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local
1070 alignment search tool. J. Mol. Biol. *215*, 403–410. https://doi.org/10.1016/S0022-
1071 2836(05)80360-2.

1072 74. Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and
1073 Madden, T.L. (2009). BLAST+: architecture and applications. BMC Bioinformatics *10*, 421.
1074 https://doi.org/10.1186/1471-2105-10-421.

1075 75. Steinegger, M., and Söding, J. (2017). MMseqs2 enables sensitive protein sequence
1076 searching for the analysis of massive data sets. Nat. Biotechnol. *35*, 1026–1028.
1077 https://doi.org/10.1038/nbt.3988.

1078 76. Tanizawa, Y., Fujisawa, T., and Nakamura, Y. (2018). DFAST: a flexible prokaryotic
1079 genome annotation pipeline for faster genome publication. Bioinformatics *34*, 1037–1039.
1080 https://doi.org/10.1093/bioinformatics/btx713.

1081 77. Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high
1082 throughput. Nucleic Acids Res. *32*, 1792–1797. https://doi.org/10.1093/nar/gkh340.

1083 78. Eddy, S.R. (2011). Accelerated Profile HMM Searches. PLoS Comput. Biol. *7*, e1002195.
1084 https://doi.org/10.1371/journal.pcbi.1002195.

1085 79. Katoh, K., and Standley, D.M. (2013). MAFFT multiple sequence alignment software version
1086 7: improvements in performance and usability. Mol. Biol. Evol. *30*, 772–780.
1087 https://doi.org/10.1093/molbev/mst010.

1088 80. Price, M.N., Dehal, P.S., and Arkin, A.P. (2010). FastTree 2 – Approximately Maximum-
1089 Likelihood Trees for Large Alignments. PLOS ONE *5*, e9490.
1090 https://doi.org/10.1371/journal.pone.0009490.

1091 81. Letunic, I., and Bork, P. (2021). Interactive Tree Of Life (iTOL) v5: an online tool for
1092 phylogenetic tree display and annotation. Nucleic Acids Res. *49*, W293–W296.
1093 https://doi.org/10.1093/nar/gkab301.

1094 82. Hyatt, D., Chen, G.-L., LoCascio, P.F., Land, M.L., Larimer, F.W., and Hauser, L.J. (2010).
1095 Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC
1096 Bioinformatics *11*, 119. https://doi.org/10.1186/1471-2105-11-119.

1097 83. Larralde, M. (2022). Pyrodigal: Python bindings and interface to Prodigal, an efficient
1098 method for gene prediction in prokaryotes. J. Open Source Softw. *7*, 4296.
1099 https://doi.org/10.21105/joss.04296.

1100 84. Eddy, S.R. (2011). Accelerated Profile HMM Searches. PLOS Comput. Biol. *7*, e1002195.
1101 https://doi.org/10.1371/journal.pcbi.1002195.

85. Larralde, M., and Zeller, G. (2023). PyHMMER: a Python library binding to HMMER for efficient sequence analysis. Bioinformatics *39*, btad214. https://doi.org/10.1093/bioinformatics/btad214.

86. Rutschmann, C., Vershynina, K., Mathew, C., Piëch, L., Sulewska, M., Schilling, P.E., Wenner, N., Larsson, L., Bertola, A., Barazzone, E.C., et al. (2025). DMB labelling for detection and analysis of capsular polysaccharides. Preprint at Biochemistry, https://doi.org/10.1101/2025.01.29.635457 https://doi.org/10.1101/2025.01.29.635457.

87. Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., et al. (2014). InterProScan 5: genome-scale protein function classification. Bioinformatics *30*, 1236–1240. https://doi.org/10.1093/bioinformatics/btu031.

88. Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., and Steinegger, M. (2022). ColabFold: making protein folding accessible to all. Nat. Methods *19*, 679–682. https://doi.org/10.1038/s41592-022-01488-1.

89. Mirdita, M., Steinegger, M., and Söding, J. (2019). MMseqs2 desktop and local web server app for fast, interactive sequence searches. Bioinformatics *35*, 2856–2858. https://doi.org/10.1093/bioinformatics/bty1057.

90. Mirdita, M., von den Driesch, L., Galiez, C., Martin, M.J., Söding, J., and Steinegger, M. (2017). Uniclust databases of clustered and deeply annotated protein sequences and alignments. Nucleic Acids Res. *45*, D170–D176. https://doi.org/10.1093/nar/gkw1081.

91. Mitchell, A.L., Almeida, A., Beracochea, M., Boland, M., Burgin, J., Cochrane, G., Crusoe, M.R., Kale, V., Potter, S.C., Richardson, L.J., et al. (2020). MGnify: the microbiome analysis resource in 2020. Nucleic Acids Res. *48*, D570–D578. https://doi.org/10.1093/nar/gkz1035.

92. Burley, S.K., Bhikadiya, C., Bi, C., Bittrich, S., Chen, L., Crichlow, G.V., Christie, C.H., Dalenberg, K., Di Costanzo, L., Duarte, J.M., et al. (2021). RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. Nucleic Acids Res. *49*, D437–D451. https://doi.org/10.1093/nar/gkaa1038.

93. Terrific Broth (TB) Medium (2015). Cold Spring Harb. Protoc. *2015*, pdb.rec085894. https://doi.org/10.1101/pdb.rec085894.

94. De Castro, C., Parrilli, M., Holst, O., and Molinaro, A. (2010). Microbe-Associated Molecular Patterns in Innate Immunity. In Methods in Enzymology (Elsevier), pp. 89–115. https://doi.org/10.1016/S0076-6879(10)80005-9.

95. Ondov, B.D., Treangen, T.J., Melsted, P., Mallonee, A.B., Bergman, N.H., Koren, S., and Phillippy, A.M. (2016). Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol. *17*, 132. https://doi.org/10.1186/s13059-016-0997-x.

96. Olson, R.D., Assaf, R., Brettin, T., Conrad, N., Cucinell, C., Davis, J.J., Dempsey, D.M., Dickerman, A., Dietrich, E.M., Kenyon, R.W., et al. (2023). Introducing the Bacterial and Viral Bioinformatics Resource Center (BV-BRC): a resource combining PATRIC, IRD and ViPR. Nucleic Acids Res. *51*, D678–D689. https://doi.org/10.1093/nar/gkac1003.

97. Shaw, J., and Yu, Y.W. (2023). Fast and robust metagenomic sequence comparison through sparse chaining with skani. Nat. Methods *20*, 1661–1665. https://doi.org/10.1038/s41592-023-02018-3.

# Supplemental figures

## The natural diversity of *E. coli* transporter-dependent capsules

Carine Roese Mores[1, 14], Samuel Miravet-Verde[1, 14, *], Elisabetta Cacace[1, 14, *], Christoph Rutschmann[2, 12], Chia-Wei Lin[6], Hans J. Ruscheweyh[1], Aline Cuenod[8], Elisa Cappio Barazzone[2], Enora Marrec[1], Kateryna Vershynina[2], Raffael Schumann[2, 13], Dan J. Bower[5, 11], Mario Schubert[7], Adrian Egli[8], Timm Fiebig[9, 10], Emma Slack[2, 3, 4, *], Shinichi Sunagawa[1, *], Timothy G. Keys[2, 12, 15, *]

**Affiliations**

[1] Department of Biology, Institute of Microbiology and Swiss Institute of Bioinformatics, ETH Zurich, Zurich, Switzerland
[2] Department of Health Sciences and Technology, Institute of Food, Nutrition and Health, ETH Zurich, Zurich, Switzerland
[3] Basel Research Center for Child Health, Basel, Switzerland
[4] Sir William Dunn School of Pathology, University of Oxford, United Kingdom
[5] QuadraticMed LLC, Baar, Switzerland
[6] Functional Genomics Centre Zurich, Zurich, Switzerland
[7] Department of Biology, Chemistry and Pharmacy, Free University of Berlin, Berlin, Germany
[8] Institute for Medical Microbiology, University of Zurich, Zurich, Switzerland
[9] German Center for Infection Research (DZIF), Partner Site Hannover-Braunschweig, Hannover, Germany
[10] Institute of Clinical Biochemistry, Hannover Medical School, Hannover, Germany
[11] Department of Earth and Planetary Sciences, ETH Zurich, Switzerland.
[12] Baxiva AG, Schlieren, Switzerland.
[13] Present address: MRC Laboratory of Molecular Biology and University of Cambridge, United Kingdom
[14] These authors contributed equally.
[15] Lead contact.
* Correspondence: smiravet@biol.ethz.ch (S.M.V.), ecacace@biol.ethz.ch (E.C.), emma.slack@immune.engineering (E.S.), ssunagawa@ethz.ch (S.S.), keyst@ethz.ch (T.G.K).

## Contents

Supplemental figures 1-7

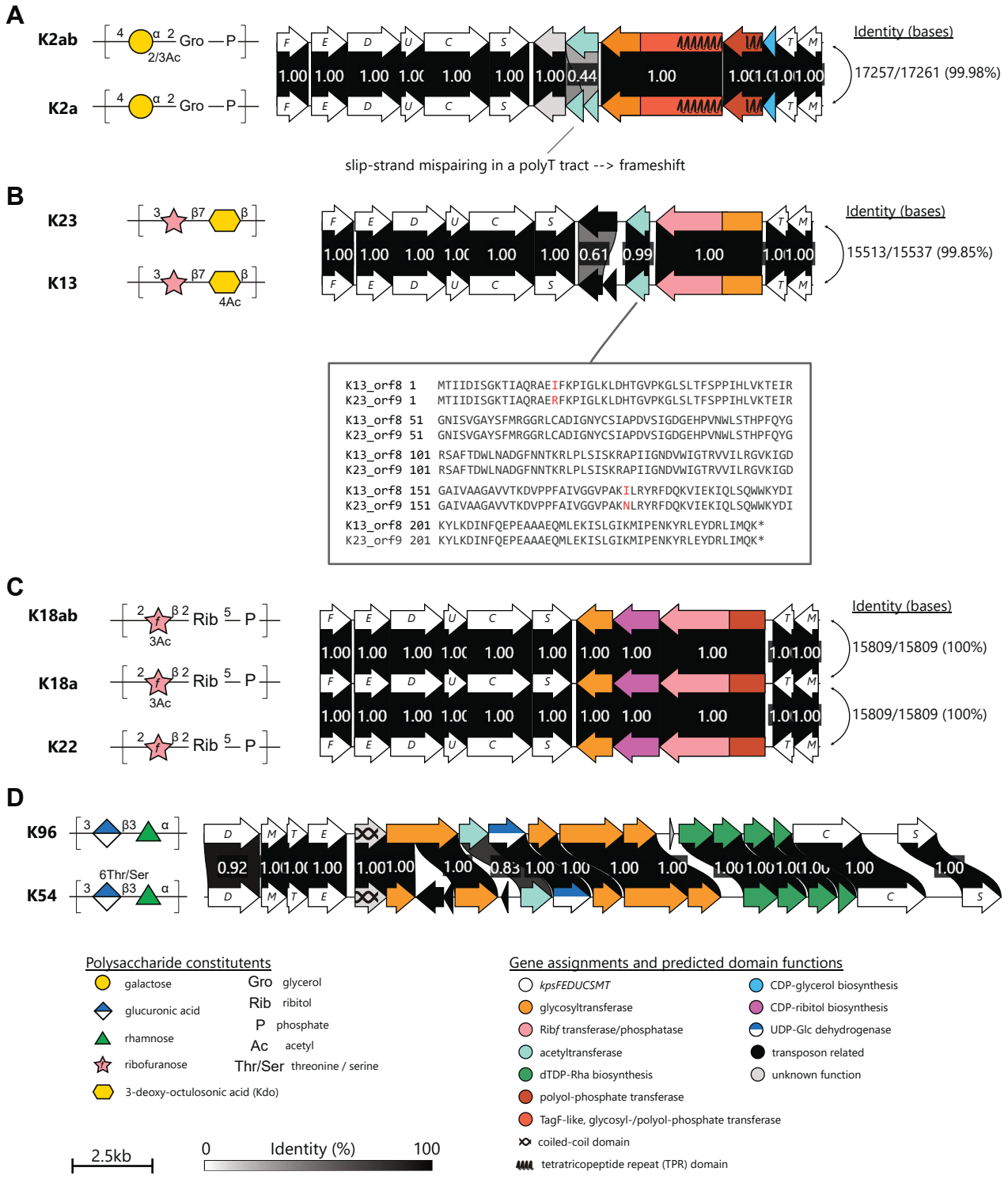



**Figure S1. Closely related capsule gene clusters identified in different K serotype reference strains do not support robust genetic differentiation of the respective serotypes, related to Figure 2.** Capsule gene clusters were compared and rendered using clinker. The proportion of identical amino acids is displayed numerically and in grey-scale between opposing ORFs. The reported base identities are from global Needle-Wunsch alignments and the protein alignments were made with the Smith-Waterman algorithm. Gene assignments and functional predictions were made as described in STAR Methods. Polysaccharides are depicted according to the conventions of the Symbol Nomenclature for Graphical Representation of Glycans and references for the displayed polysaccharide structures are provided in Table S1. (A) K2a and K2ab differ by acetylation of galactose in the repeat unit, consistent with a frameshift mutation in the putative acetyltransferase ORF in K2a. (B) The K13 and K23 polysaccharides differ by acetylation of the Kdo residue in the repeat unit and the candidate acetyltransferases differ by two missense mutations. Although the observed mutations may explain the K2a/K2ab and K13/K23 serotypes, they do not provide a robust basis for genetically distinguishing these serotypes. Therefore, these serotypes are grouped under the names K2 and K13_K23, respectively, in our catalog (Table S2).

**Figure S1 (**continued). (C) The K18 polysaccharide is an acetylated version of the K22 backbone, but the capsule gene clusters are identical (100% identity), and there is no candidate acetyltransferase gene within the clusters. The serological distinction between K18a and K18ab is unclear because chemical analysis of the polysaccharides revealed identical repeat unit structures. As there is currently no genetic basis for distinguishing these serotypes, they are grouped under the assignment K18_K22 in our catalog (Table S2). (D) Our analysis of the K54 and K96 gene clusters is consistent with earlier reports suggesting that the gene(s) responsible for threonine modification of the polysaccharide backbone are located outside of the kps locus. Pending identification of these gene(s), the K54 and K96 clusters are grouped under the assignment K96 in the current version of the catalog (Table S2).

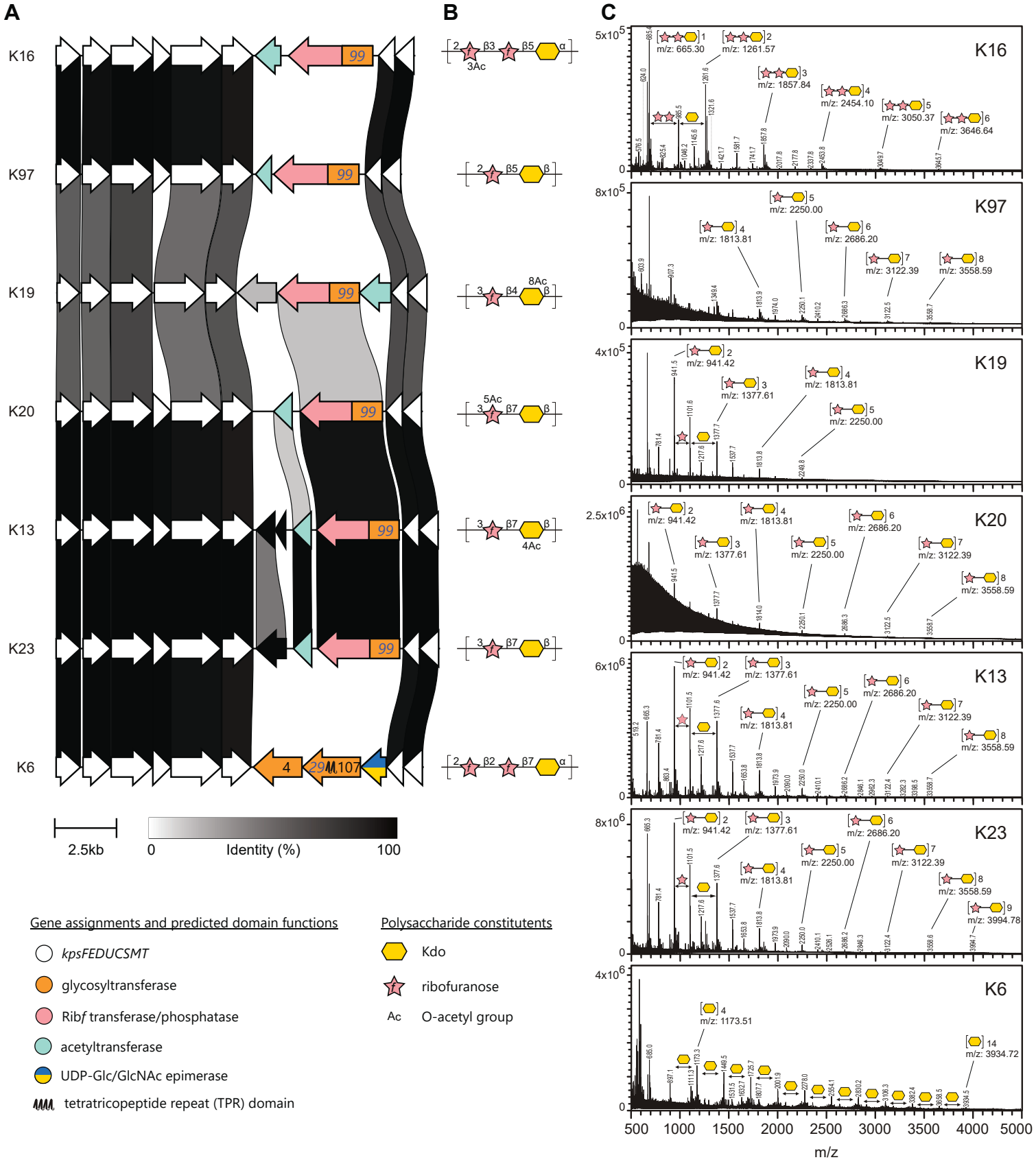
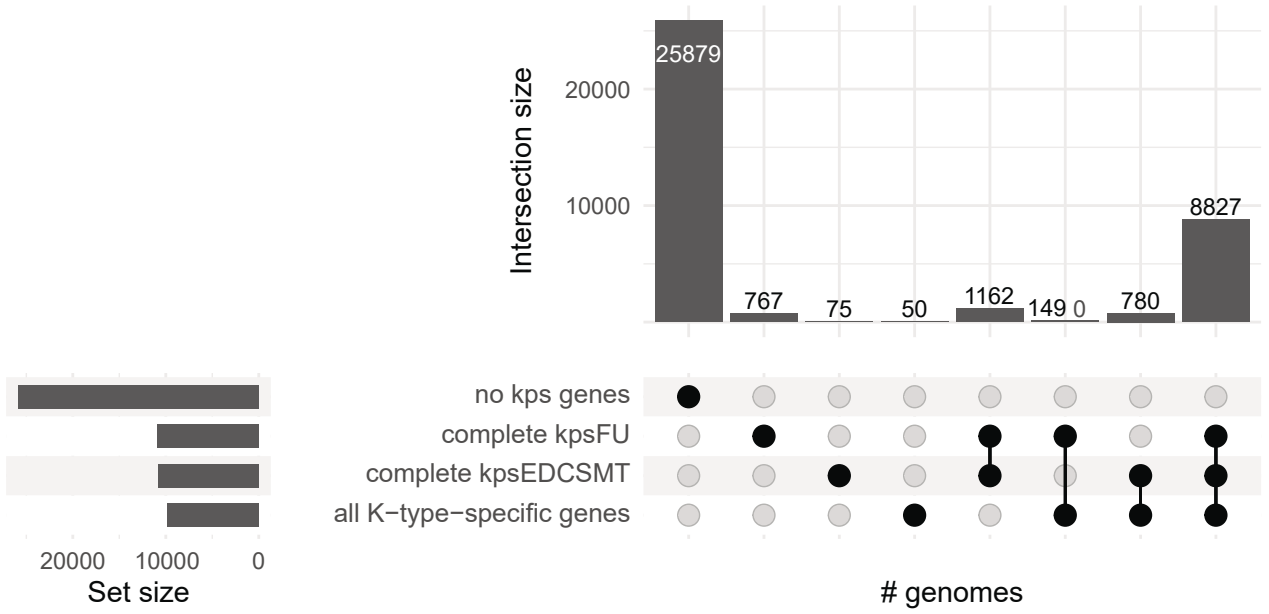



**Figure S2. Mass spectrometry clarifies K antigen structures, related to Figure 2.** (A) Capsule gene clusters from K serotype reference strains were compared and rendered with clinker, then colored according to gene assignments and predicted domain functions. Amino acid identity > 30% is displayed in greyscale between opposing genes. (B) Published polysaccharide repeat unit structures for each K serotype (see references in Table S1). (C) Deconvoluted MALDI-TOF-TOF spectra of partially hydrolysed and permethylated capsular polysaccharide (**STAR Methods**). Peak labels indicate the assigned structure and corresponding expected m/z species calculated in GlycoWorkbench. We note that O-acetyl groups are expected to be lost during the permethylation procedure and were not considered when assigning peaks.

A




B

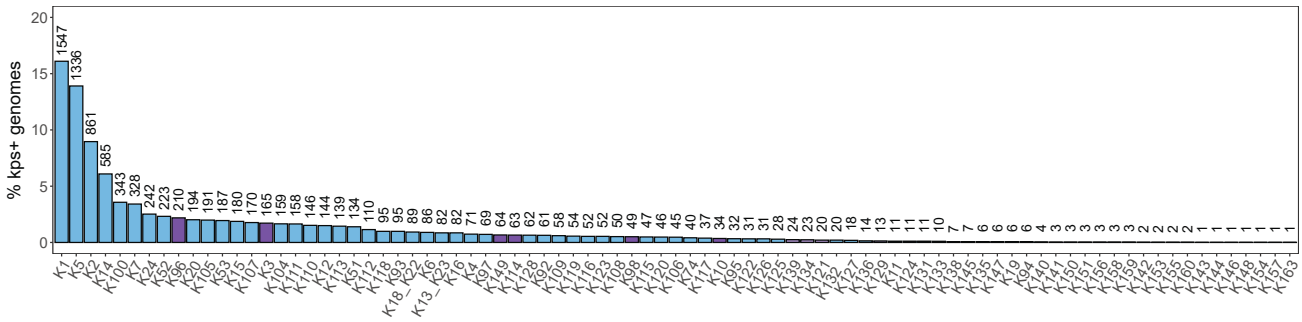



**Figure S3. Evaluation of kTYPr on the RefSeq source data set.** (A) Upset plot of kTYPr results on 37,723 *E. coli* genomes from RefSeq, showing the distribution of *E. coli* genomes with complete and incomplete capsule biosynthesis loci. (B) Absolute counts of K-types identified in the RefSeq data set.

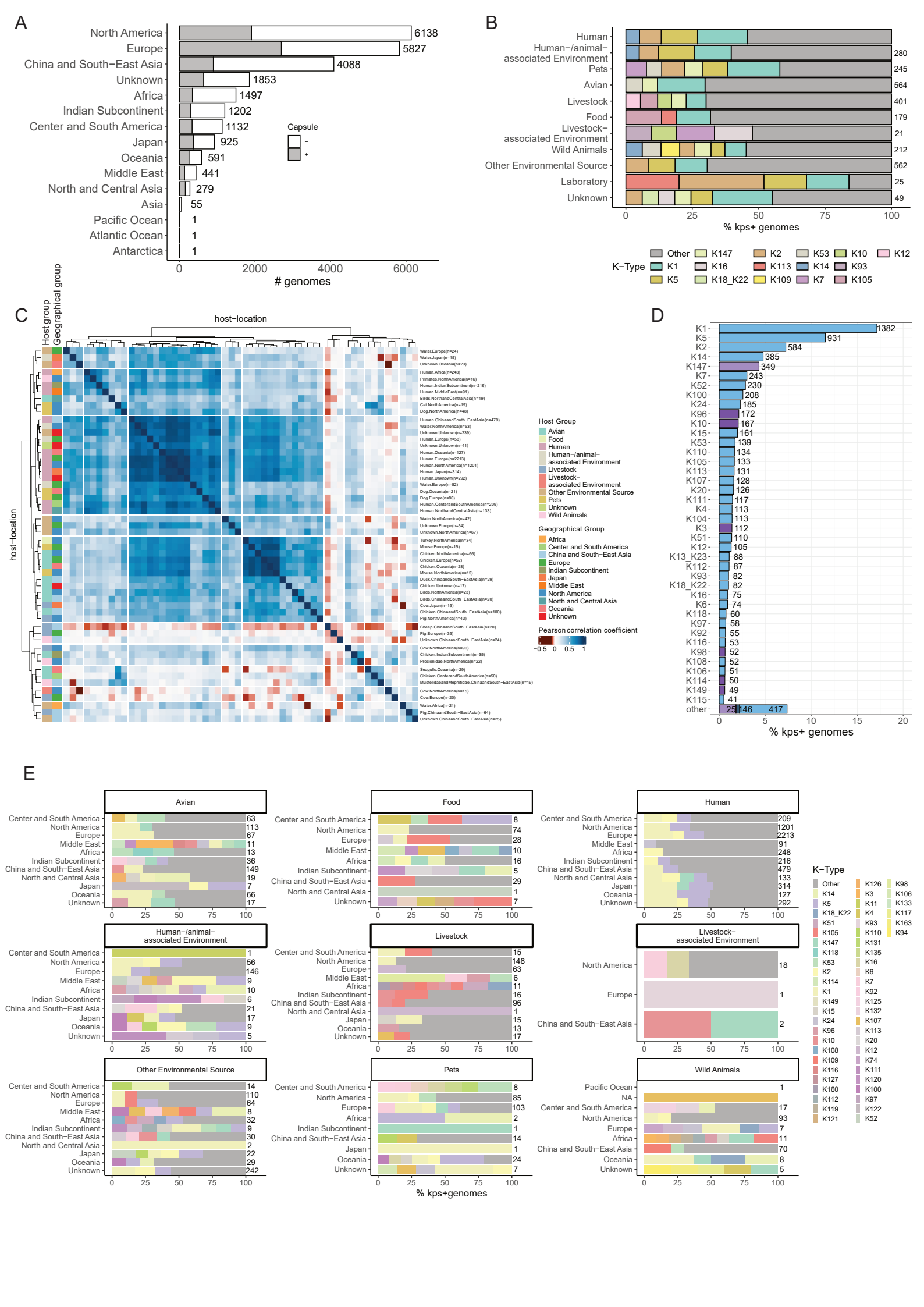

A
North America 6138
Europe 5827
China and South−East Asia 4088
Unknown 1853
Africa 1497
Indian Subcontinent 1202
Center and South America 1132
Japan 925
Oceania 591
Middle East 441
North and Central Asia 279
Asia 55
Pacific Ocean 1
Atlantic Ocean 1
Antarctica 1
Capsule
−
+
0 2000 4000 6000
# genomes
B
Human
Human−/animal−associated Environment 280
Pets 245
Avian 564
Livestock 401
Food 179
Livestock−associated Environment 21
Wild Animals 212
Other Environmental Source 562
Laboratory 25
Unknown 49
0 25 50 75 100
% kps+ genomes
K−Type
Other K147 K2 K53 K10 K12
K1 K16 K113 K14 K93
K5 K18_K22 K109 K7 K105
C
host−location
Host group
Geographical group
Host Group
Avian
Food
Human
Human−/animal−associated Environment
Livestock
Livestock−associated Environment
Other Environmental Source
Pets
Unknown
Wild Animals
Geographical Group
Africa
Center and South America
China and South−East Asia
Europe
Indian Subcontinent
Japan
Middle East
North America
North and Central Asia
Oceania
Unknown
Pearson correlation coefficient
−0.5 0 0.5 1
D
K1 1382
K5 931
K2 584
K14 385
K147 349
K7 243
K52 230
K100 208
K24 185
K96 172
K10 167
K15 161
K53 139
K110 134
K105 133
K113 131
K107 128
K20 126
K111 117
K4 113
K104 113
K3 112
K51 110
K12 105
K13_K23 88
K112 87
K93 82
K18_K22 82
K16 75
K6 74
K118 60
K97 58
K92 55
K116 53
K98 52
K108 52
K106 51
K114 50
K149 49
K115 41
other 25 146 417
0 5 10 15 20
% kps+ genomes
E
Avian
Center and South America 63
North America 113
Europe 67
Middle East 11
Africa 13
Indian Subcontinent 36
China and South−East Asia 149
North and Central Asia 19
Japan 7
Oceania 66
Unknown 17
Food
Center and South America 8
North America 74
Europe 28
Middle East 10
Africa 16
Indian Subcontinent 5
China and South−East Asia 29
North and Central Asia 1
Unknown 7
Human
Center and South America 209
North America 1201
Europe 2213
Middle East 91
Africa 248
Indian Subcontinent 216
China and South−East Asia 479
North and Central Asia 133
Japan 314
Oceania 127
Unknown 292
Human−/animal−associated Environment
Center and South America 1
North America 56
Europe 146
Middle East 9
Africa 10
Indian Subcontinent 6
China and South−East Asia 21
Japan 17
Oceania 9
Unknown 5
Livestock
Center and South America 15
North America 148
Europe 63
Middle East 6
Africa 11
Indian Subcontinent 16
China and South−East Asia 96
North and Central Asia 1
Japan 15
Oceania 13
Unknown 17
Livestock−associated Environment
North America 18
Europe 1
China and South−East Asia 2
Other Environmental Source
Center and South America 14
North America 110
Europe 64
Middle East 8
Africa 32
Indian Subcontinent 9
China and South−East Asia 30
North and Central Asia 2
Japan 22
Oceania 29
Unknown 242
Pets
Center and South America 8
North America 85
Europe 103
Africa 2
Indian Subcontinent 1
China and South−East Asia 14
Japan 1
Oceania 24
Unknown 7
Wild Animals
Pacific Ocean 1
NA
Center and South America 17
North America 93
Europe 7
Africa 11
China and South−East Asia 70
Oceania 8
Unknown 5
0 25 50 75 100
% kps+genomes
K−Type
Other K126 K98
K14 K3 K106
K5 K11 K133
K18_K22 K4 K117
K51 K93 K163
K105 K110 K94
K147 K131
K118 K135
K53 K16
K2 K6
K114 K7
K1 K92
K149 K125
K15 K132
K24 K107
K96 K113
K10 K20
K108 K12
K109 K74
K116 K111
K127 K120
K160 K100
K112 K97
K119 K122
K121 K52

**Figure S4. Global diversity of K-types, related to Figure 3.** (A) Fraction of *kps*-positive genomes (defined as containing a complete *kps* locus, STAR Methods) across geography. (B) K-type proportion in hosts and environments. K-types with frequency < 5% are annotated as “other”. (C) K-type proportion in individual hosts and environments. K-types and sample origin are clustered according to Pearson’s correlation. Host-environment combinations with < 15 occurrences were not considered. K-type groups are color-coded as in Fig. 3C. Hosts are color-coded according to their lower resolution groups shown in Fig. 3B. (D) Proportion and absolute counts of K-types, color-coded according to groups as in Fig. 3C. K-types with proportion < 0.5% were annotated as “other”. Only *kps*-positive genomes were considered. (E) K-type proportion across geographic regions, grouped by hosts and environments. K-types with frequency < 8% are annotated as “other”. Only *kps*-positive genomes were considered.

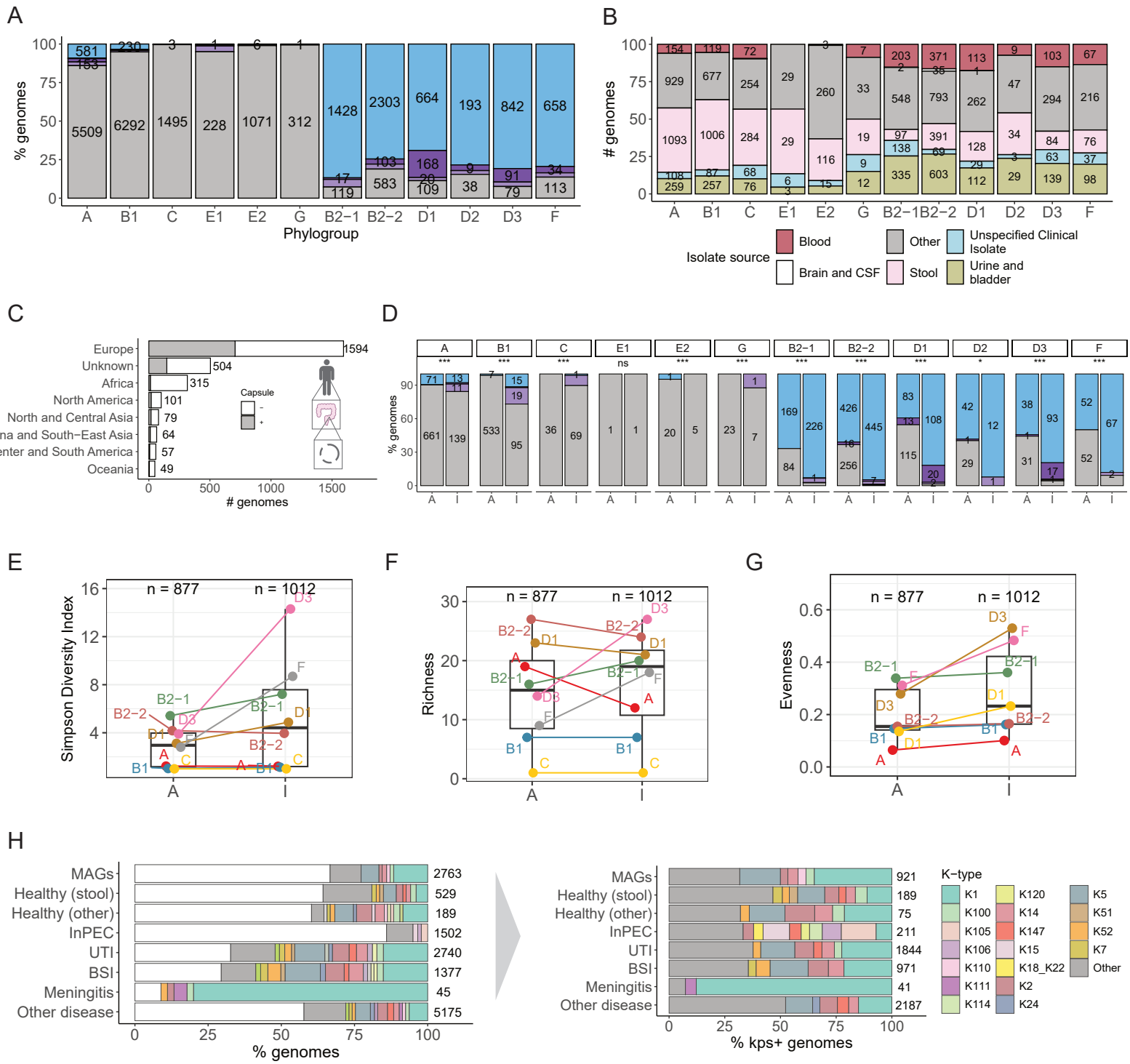
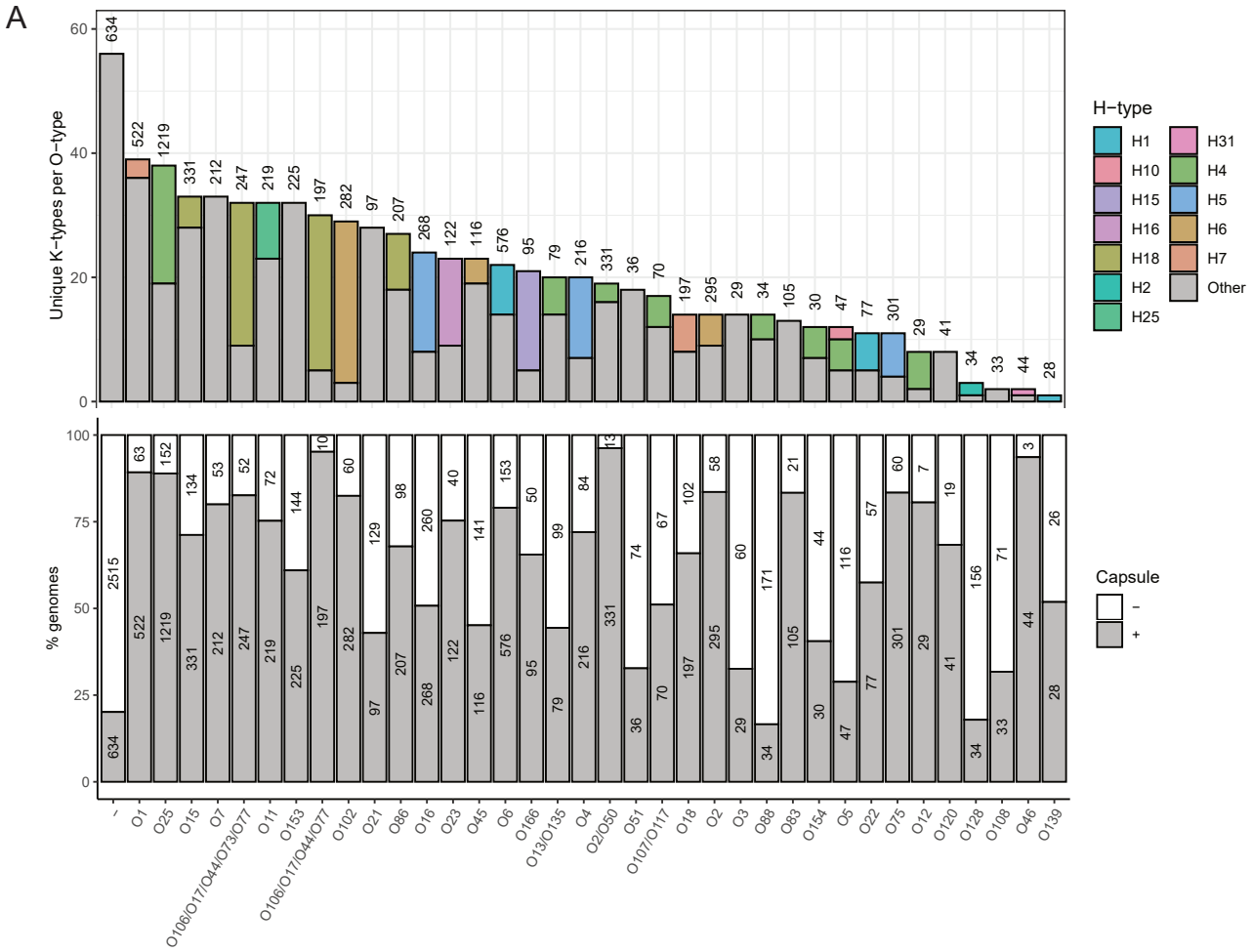



**Figure S5. K-type associations with phylogroups and human health and disease, related to Figure 4.** (A) Proportion of capsule groups across phylogroups in *E. coli* genomes from all hosts and environments. Groups are color-coded as in Fig. 3C. (B) *E. coli* human isolate source in the collection, grouped by phylogroup (n=11,537). 20 genomes to which no phylogroup could be assigned were removed. (C) Fraction of capsule-positive genomes (defined as containing a complete *kps* locus, STAR Methods) in *E. coli* human gut metagenomes (Table S4). (D) Capsule group proportion in each phylogroup in asymptomatic carriage (A, corresponding to 2,762 *E. coli* MAGs from healthy individuals, Table S4) and invasive *E. coli*-associated disease (I, 1,178 genomes of isolates from blood or CSF from NCBI and 259 from blood from a published study on urosepsis), STAR Methods. ns $p > 0.05$; * $p \leq 0.05$; ** $p \leq 0.01$; *** $p \leq 0.001$ (Fisher's exact test). (E-G) K-type diversity (expressed as Simpson index, E), richness (F) and evenness (G) of phylogroups. The same data as in D) were considered, filtering out phylogroups with ≤ 25 genomes in each group (asymptomatic, A or invasive, I). The number of capsule-positive genomes in the two groups is indicated above each boxplot. Box limits correspond to first and third quartiles, with the median marked, and whiskers extending to the most extreme data points up to 1.5 times the interquartile range (IQR). (H) K-type proportion across health groups defined as in Fig. 4B, considering all (left) or only capsule-positive genomes (right). K-types with proportion < 1% (left) or < 3% (right) are grouped as "other".


A
Unique K−types per O−type
% genomes
H−type
H1
H10
H15
H16
H18
H2
H25
H31
H4
H5
H6
H7
Other
Capsule
−
+
−
O1
O25
O15
O7
O106/O17/O44/O73/O77
O11
O153
O106/O17/O44/O77
O102
O21
O86
O16
O23
O45
O6
O166
O13/O135
O4
O2/O50
O51
O107/O117
O18
O2
O3
O88
O83
O154
O5
O22
O75
O12
O120
O128
O108
O46
O139

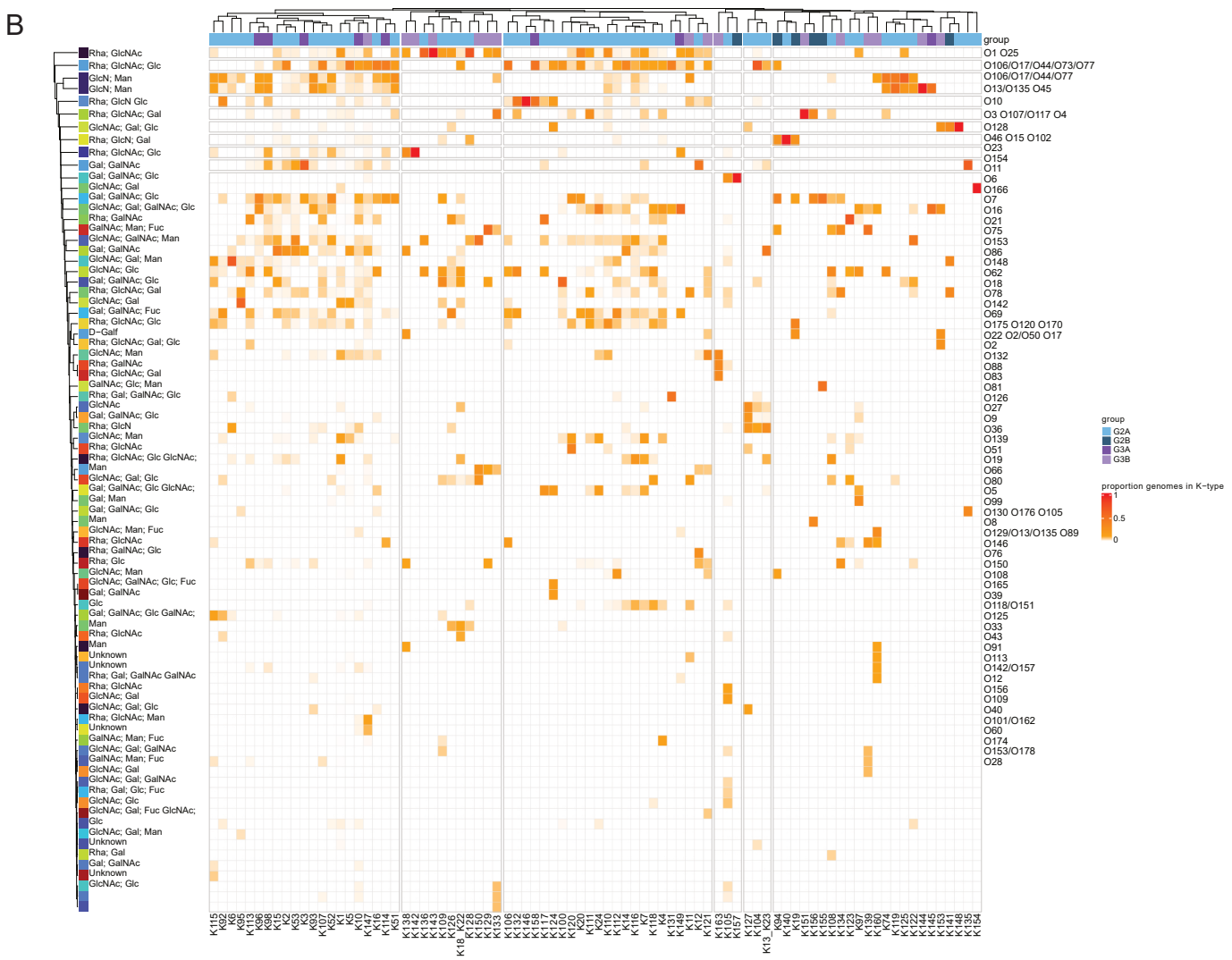

B
group
O1 O25
O106/O17/O44/O73/O77
O106/O17/O44/O77
O13/O135 O45
O10
O3 O107/O117 O4
O128
O46 O15 O102
O23
O154
O11
O6
O166
O7
O16
O21
O75
O153
O86
O148
O62
O18
O78
O142
O69
O175 O120 O170
O22 O2/O50 O17
O2
O132
O88
O83
O81
O126
O27
O9
O36
O139
O51
O19
O66
O80
O5
O99
O130 O176 O105
O8
O129/O13/O135 O89
O146
O76
O150
O108
O165
O39
O118/O151
O125
O33
O43
O91
O113
O142/O157
O12
O156
O109
O40
O101/O162
O60
O174
O153/O178
O28
group
G2A
G2B
G3A
G3B
proportion genomes in K−type
1
0.5
0

**Figure S6. K-type associations with O- and H-antigens.** (A) Proportion of *kps*-positive genomes (top) and the number of unique K-types (bottom) per O-type. H-type composition is indicated for each O-type (bottom). Only O-types with > 25 genomes are shown. H-types with proportion < 30% are grouped as “Other”. (B) O-type proportion for each K-type. K- and O-negative cases are not shown. Rare O-types (total proportion across all K-types ≤ 0.032, as first quartile of distribution) are not shown. K-type groups are color-coded as in Fig. 3C. O-type sugar composition was obtained from ECODAB. Only unique sugars are listed. K12 co-occurrence with rhamnose-containing O-types is highlighted.

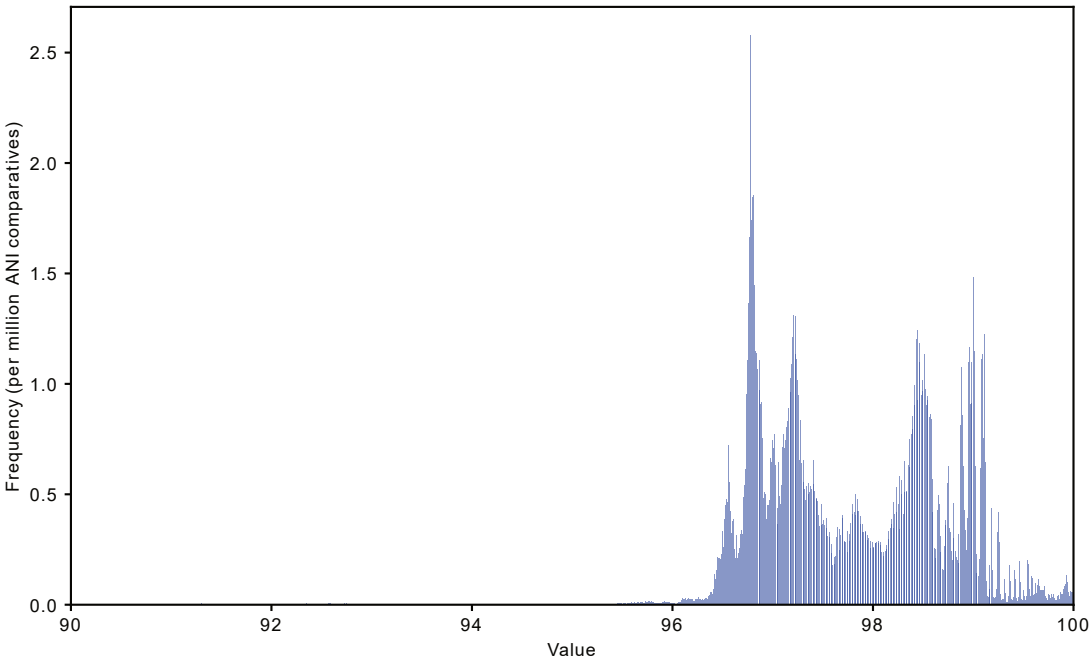



**Figure S7. Distribution of ANI scores across genome comparisons.** For 778,526,070 all-vs-all ANI comparisons, we show the frequency distribution (normalized per million pairs) using 1,500 bins for all the ANI values (X-axis). The 98% threshold was selected based on the separation of the two main modes of the distribution.