Supplemental Information for:

Persisting uropathogenic Escherichia coli lineages show signatures of niche-specific within-host adaptation mediated by mobile genetic elements

Robert Thänert, JooHee Choi, Kimberly A. Reske, Tiffany Hink, Anna Thänert, Meghan A. Wallace, Bin Wang, Sondra Seiler, Candice Cass, Margaret H. Bost, Emily L. Struttmann, Zainab Hassan Iqbal, Steven R. Sax, Victoria J. Fraser, Arthur W. Baker, Katherine R. Foy, Brett Williams, Ben Xu, Pam Capocci-Tolomeo, Ebbing Lautenbach, Carey-Ann D. Burnham, Erik R. Dubberke, Jennie H. Kwon, Gautam Dantas for the CDC Prevention Epicenter Program

Table S1 | UPEC sequence type (ST) distribution. Related to Figure 1.

Phylogroup (Prevalence %)	Clonal groups (Prevalence %)	*fimH* type	Dual colonizer (*n*=32)	Gut colonizer (*n*=51)	Urinary colonizer (*n*=4)
A (2.3%)	**410** (1.1%)	**24**	**0**	**1** (1.9%)	**0**
	744 (1.1%)	**54**	**1** (3.1%)	**0**	**0**
B2 (75.9%)	**73** (1.1%)	**103**	**0**	**0**	**1** (25%)
	95 (1.1%)	**27**	**1** (3.1%)	**0**	**0**
	131 (47.1%)	**30**	**14** (43.75%)	**23** (45.1%)	**0**
		41	**3** (9.4%)	**0**	**0**
		undefined	**0**	**1** (1.9%)	**0**
	636 (1.1%)	**undefined**	**0**	**1** (1.9%)	**0**
	1193 (25.3%)	**64**	**8** (25%)	**12** (23.5%)	**2** (50%)
C (1.1%)	**10** (1.1%)	**171**	**0**	**1** (1.9%)	**0**
D (13.8%)	**38** (2.3%)	**5**	**1** (3.1%)	**0**	**0**
		65	**0**	**1** (1.9%)	**0**
	69 (3.4%)	**27**	**1** (3.1%)	**2** (3.9%)	**0**
	70 (1.1%)	**65**	**0**	**0**	**1** (25%)
	405 (3.4%)	**27**	**0**	**3** (5.9%)	**0**
	501 (1.1%)	**undefined**	**1** (3.1%)	**0**	**0**
	1177 (1.1%)	**65**	**0**	**1** (1.9%)	**0**
	2003 (1.1%)	**65**	**1** (3.1%)	**0**	**0**
F (5.7%)	**354** (2.3%)	**58**	**0**	**2** (3.9%)	**0**
	648 (2.3%)	**29**	**1** (3.1%)	**0**	**0**
		undefined	**0**	**1** (1.9%)	**0**
	6870 (1.1%)	**undefined**	**0**	**1** (1.9%)	**0**
Unknown (1.1%)	**2006** (1.1%)	**61**	**0**	**1** (1.9%)	**0**

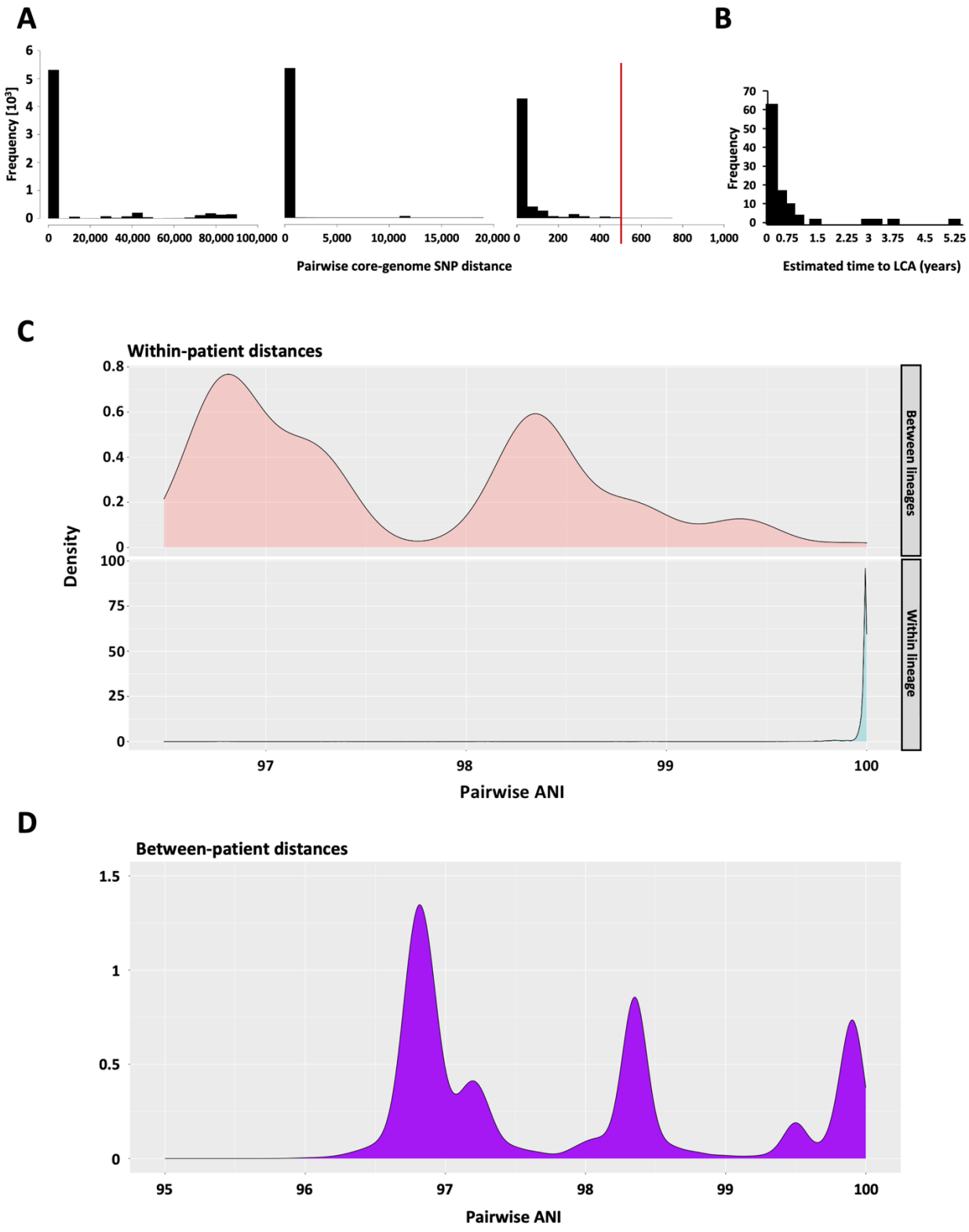

Figure S1 | Lineage definition. Related to STAR Methods and Figure 1

A) Histogram of *E. coli* pairwise within-patient core-genome SNP distances. Panels from left to right depict the same data using sequentially shorter x-axis ranges. Red line indicates cutoff used to define lineages.

B) Histogram of time to last common ancestor for UPEC lineages applying a 500 core-genome SNP cutoff to define lineages. **C)** Pairwise ANI values between same-patient isolates of different (top) and the same (bottom) *E. coli* lineage applying a 500 core-genome SNP cutoff to define lineages. D) Pairwise ANI values between different-patient isolates.

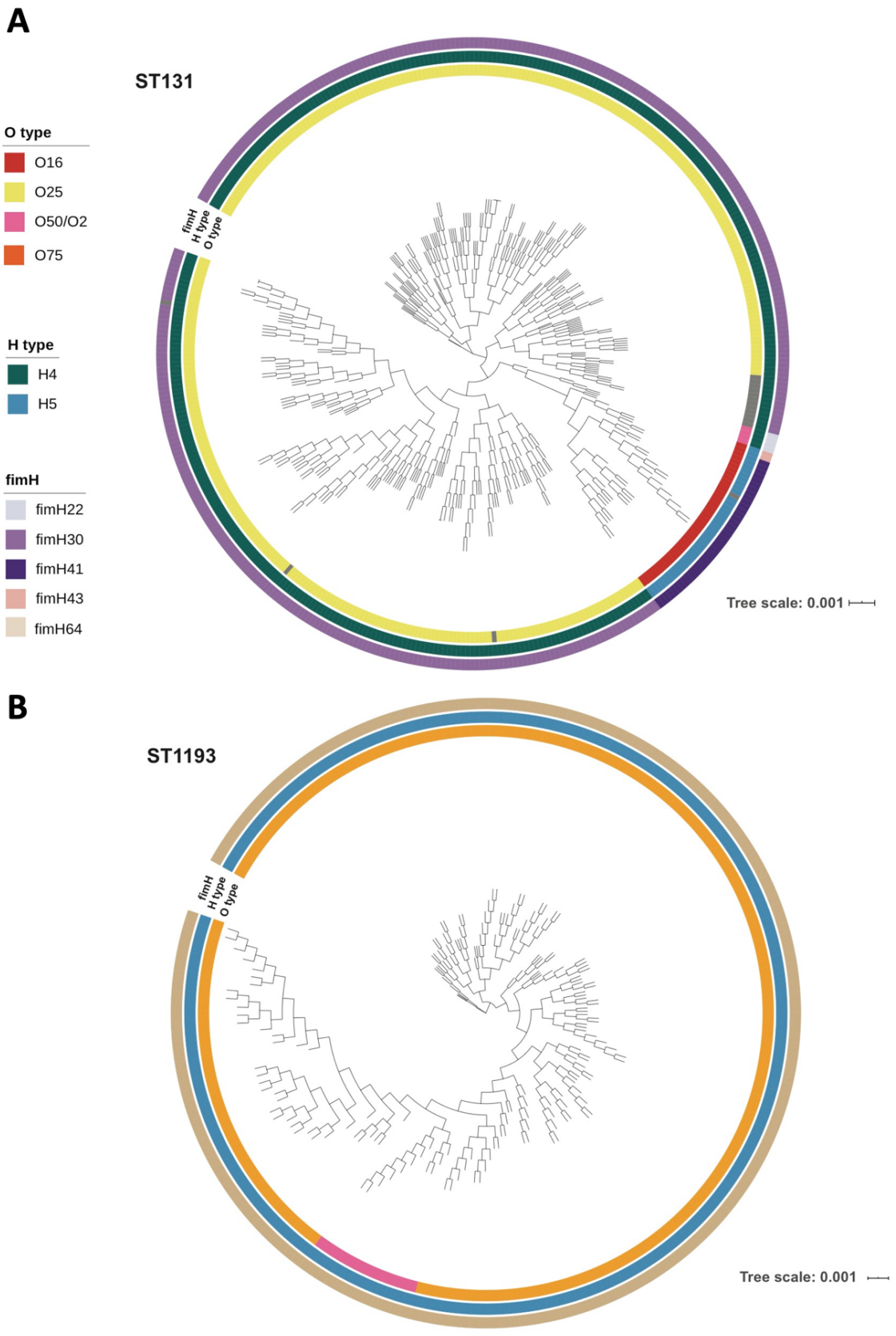

Figure S2 | Phylogenetic analysis of ST131 and ST1193. Related to STAR Methods and Figure 1

A) Unrooted core genome phylogeny of E. coli A) ST131 and B) ST1193. The outer rings annotate the

O-type, H-type, and *fimH*-type of each isolate.

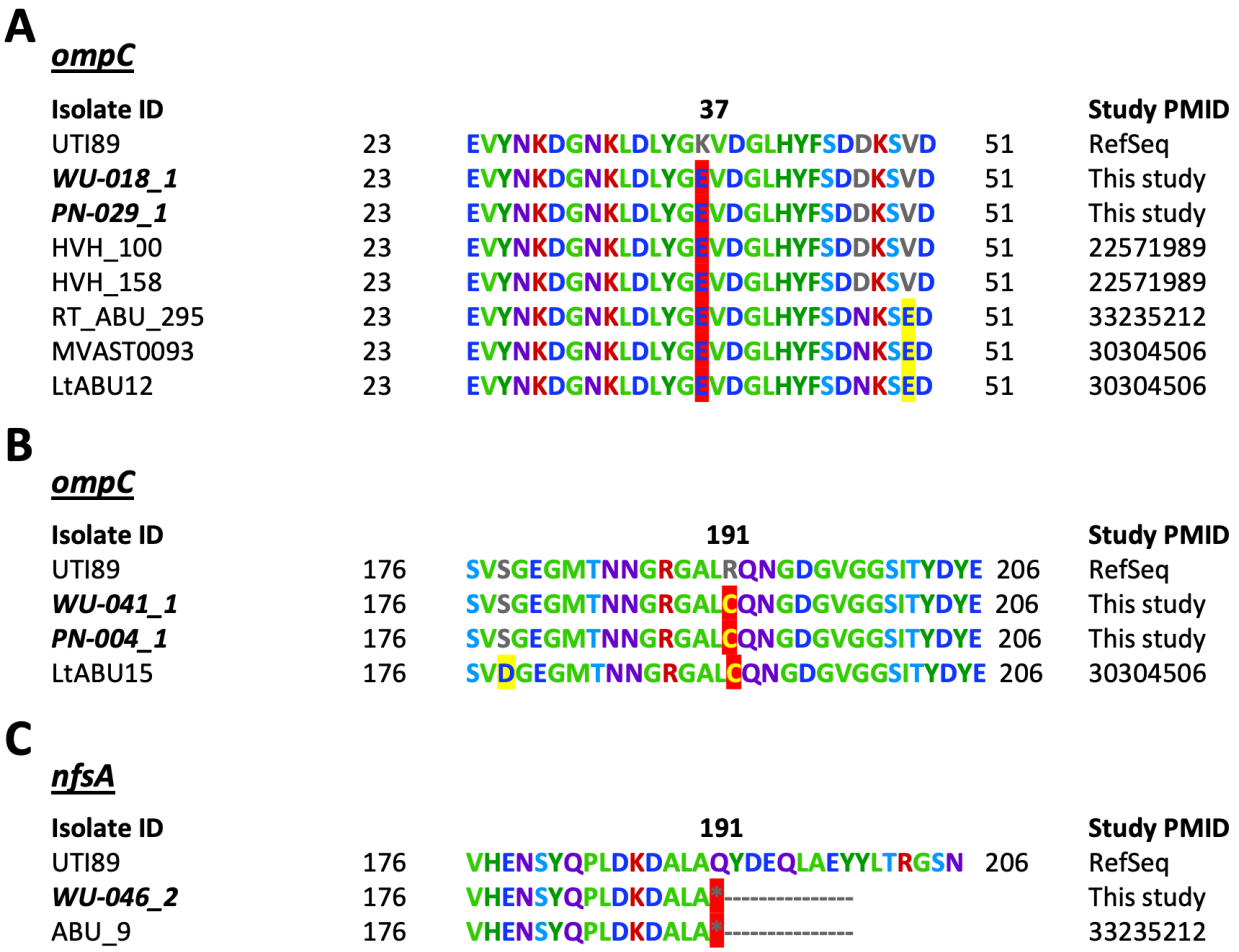

Figure S3. Related to Figure 3 | Multiple sequence alignment of variable regions in *ompC* and *nfsA*. A) Multiple sequence alignment of region of *ompC* region 23-51 between lineages with the 37 K->E found in this study and previously published genomes. **B)** Multiple sequence alignment of region of *ompC* egion 176-206 between lineages with the 191 R->S found in this study and previously published genomes. **C)** Multiple sequence alignment of region of *nfsA* region 176-206 between lineages with the 191 Q->* found in this study and previously published genomes. UTI89 sequence is added as a reference in all panels. Study PMID for published genomes are provided.

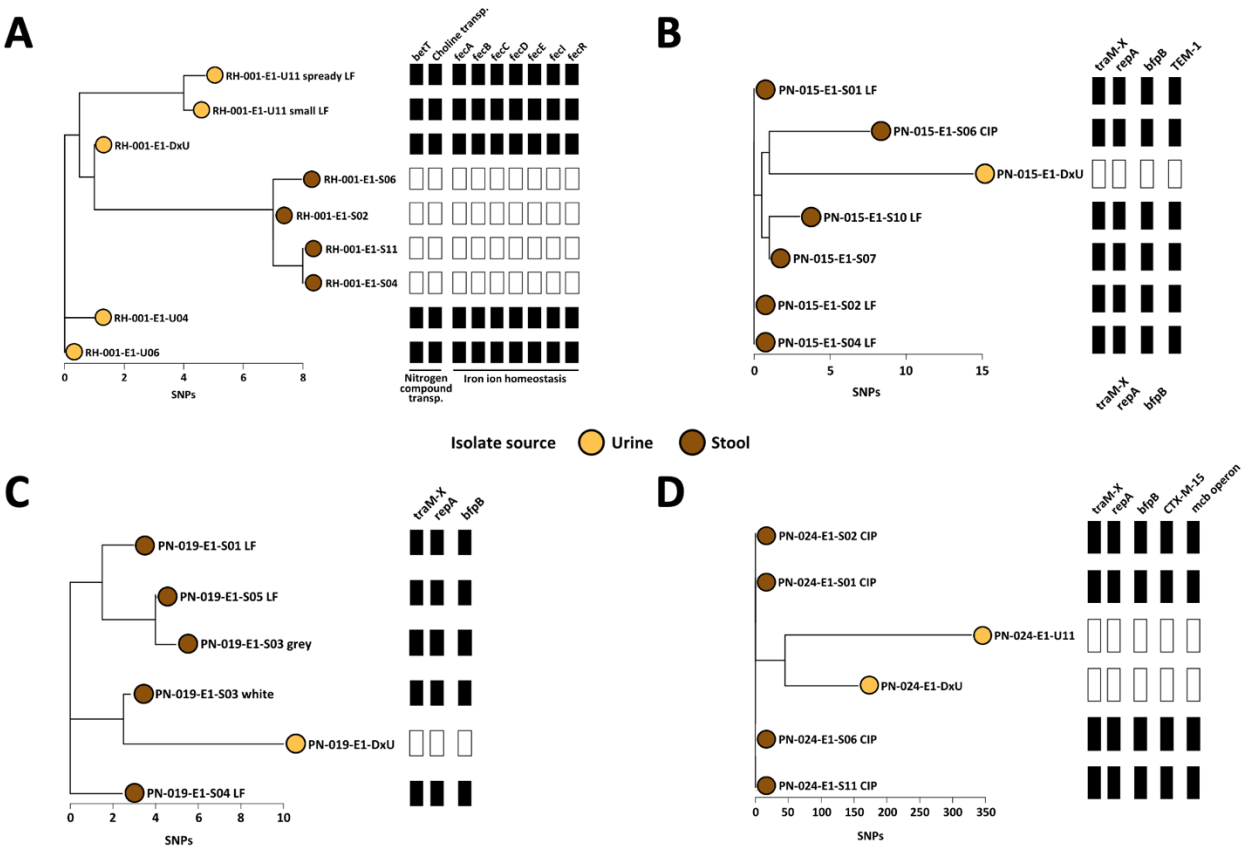

Figure S4 | A set of virulence and resistance genes is habitat-specific in persisting UPEC lineages. Related to Figure 4. A) Unrooted phylogeny of lineage RH-001_1 based on SNP distances annotated with selected habitat specific genes. **B)** Unrooted phylogeny of lineage PN-015_1 based on SNP distances annotated with selected habitat specific genes. **C)** Unrooted phylogeny of lineage PN-19_2 based on SNP distances annotated with selected habitat specific genes. **D)** Unrooted phylogeny of lineage PN-024_1 based on SNP distances annotated with selected habitat specific genes.

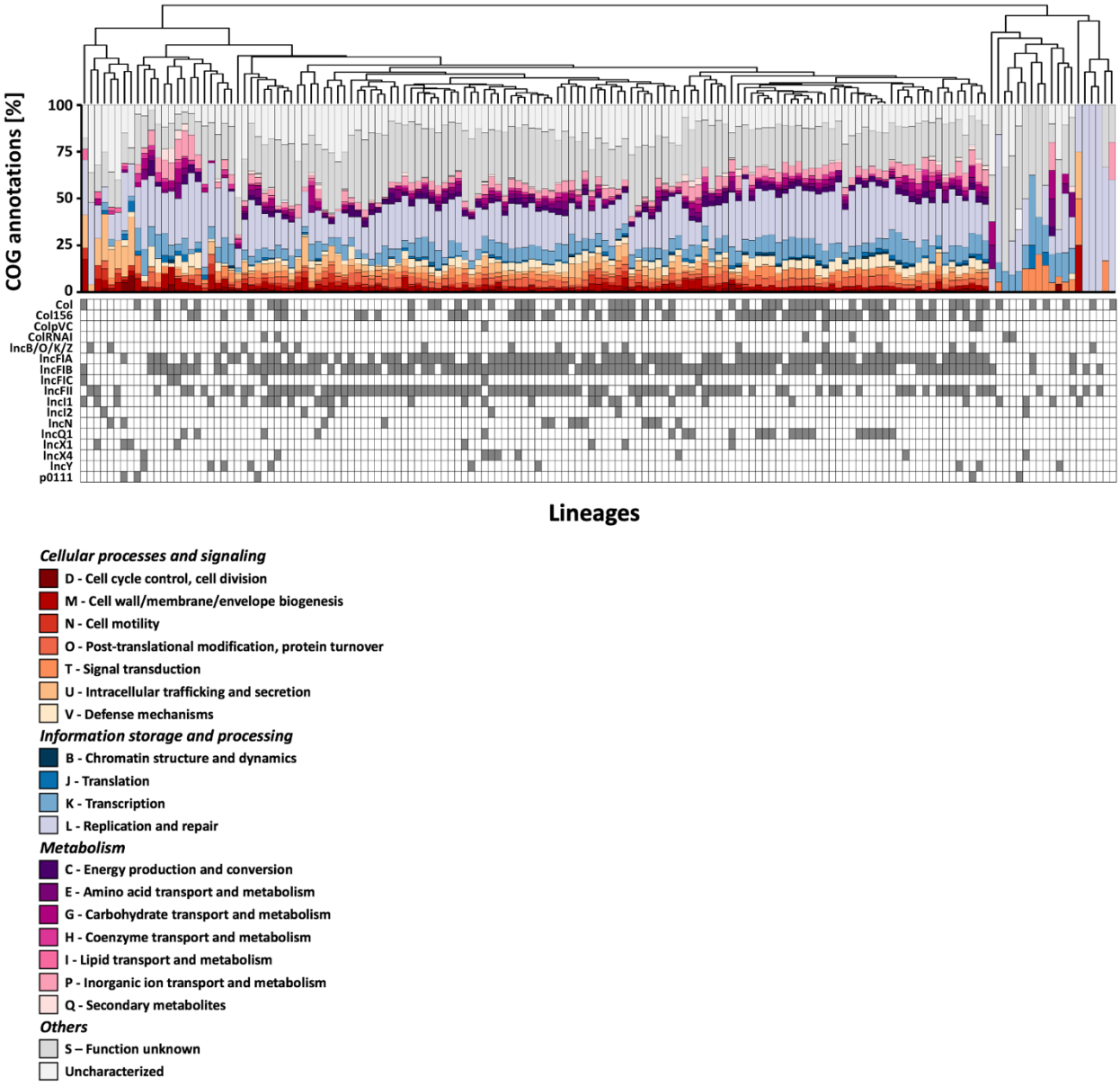

Figure S5 | The predicted lineage-specific plasmid repertoire of AR *E. coli* differs. Related to Figure 5 and STAR Methods. (Top) Lineage-specific GO-term annotation of coding sequences on contigs identified *in silico* to be of putative plasmidic origin. Only lineages with predicted plasmidic contigs are shown. (Bottom) Corresponding lineage-specific replicon-repertoire as determined using plasmidFinder.

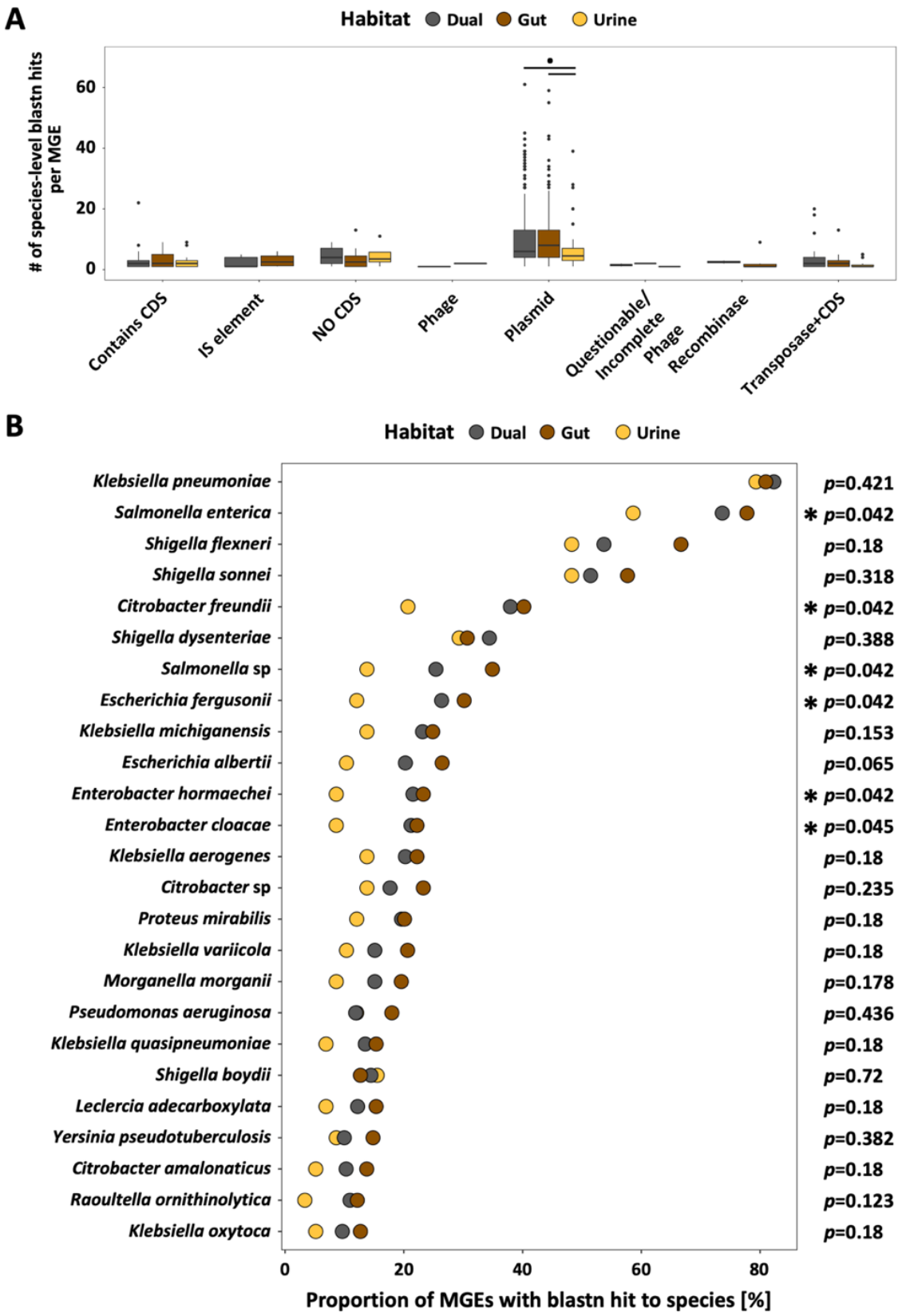

Figure S6 | Predicted host-range of putative MGEs. Related to Figure 5. A) UPEC putative plasmidic MGEs are commonly found in other species. Blastn results of putative MGEs classified as plasmidic against the NCBI nucleotide database (>95% identity, >95% query coverage.)

Urinary plasmidic MGEs were found in significantly less species compared to contigs present in stool or across habitats (Two-way ANOVA $P \leq 1.57e^{-05}$, Tukey post-hoc $P<0.001$ and $P=0.014$, respectively) **B)** Percentage of plasmidic MGE sharing between UPEC and the 25 species found to share the most plasmidic contigs with UPEC. *P*-values indicate significance values for the underrepresentation of species in the pool of urinary MGEs compared to the combined stool/dual plasmidic MGE pool as determined using Fisher's exact test. *P*-values are FDR corrected.

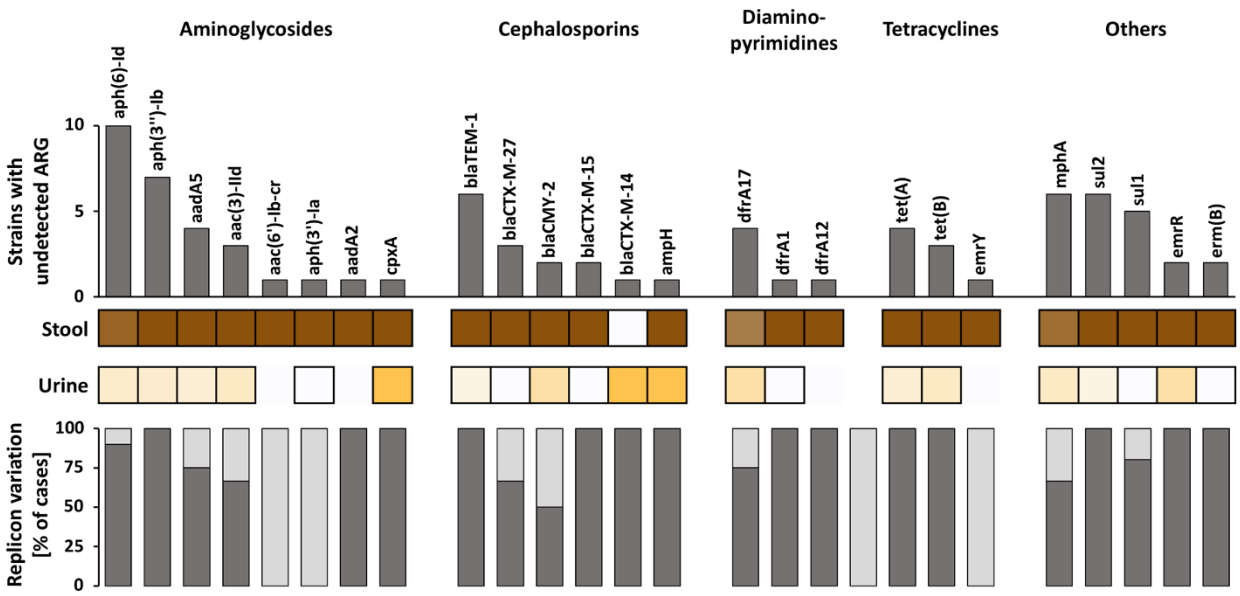

Figure S7 | Intestinally persistent UPEC are a reservoir for ARGs. Related to Figure 5. (Top) Number of lineages with 'hidden' ARGs grouped by resistance class (see Results). (Middle) Heatmap indicating the percentage of 'hidden' ARG cases where the ARG is found in an asymptomatic isolate recovered from urine (yellow) or stool (brown). (Bottom) Percentage of cases where 'hidden' ARGs are accompanied by variation in the replicon repertoire of the isolate carrying the compared to the DxU isolate.

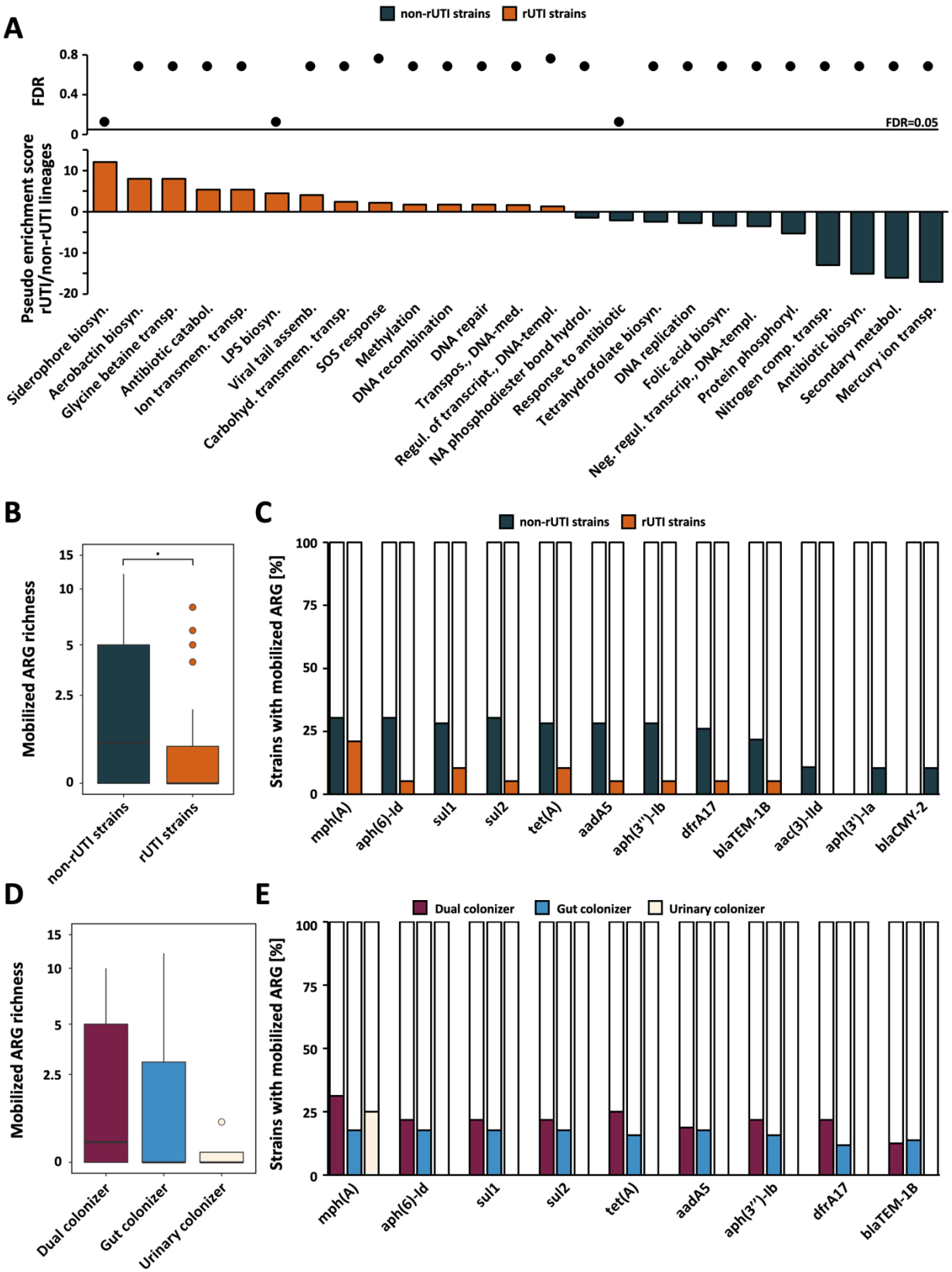

Figure S8 | Enrichment of MGE GO terms and mobilized ARGs by lineage recurrence status

and persistence type. Related to Figure 6. A) Despite variability no GO terms are over- or underrepresented in the mobilized gene pool of rUTI (orange) and non-rUTI (green) UPEC lineages (*n*=69 lineages, Fisher's exact test, all FDR corrected *P*-values >0.05). GO term overrepresentation was assessed using Fisher's exact test. *P*-values were FDR corrected. Pseudo enrichment scores were calculated comparing observed GO term abundances between compared groups adding the minimal value in the array as a pseudo-count. **B)** Mobilized ARG richness between rUTI (orange) and non-rUTI (green) lineages (*n*=69 lineages, Wilcoxon rank-sum test *P*=0.055). **C)** Prevalence of specific mobilized ARGs did not vary significantly between rUTI (orange) and non rUTI lineages (green, *n*=69 lineages, Fisher's exact test, all FDR corrected *P*-values >0.05). **D)** Mobilized ARG richness did not differe significantly between dual colonizers (maroon), gut colonizers (blue) and urinary colonizing lineages (light yellow, *n*=87 lineages, Kruskal-Wallis *P*=0.231). **E)** Prevalence of specific mobilized ARGs did not vary significantly between dual colonizers (maroon), gut colonizers (blue) and urinary colonizing lineages (light yellow, Fisher's exact test, all FDR corrected *P*-values >0.05).