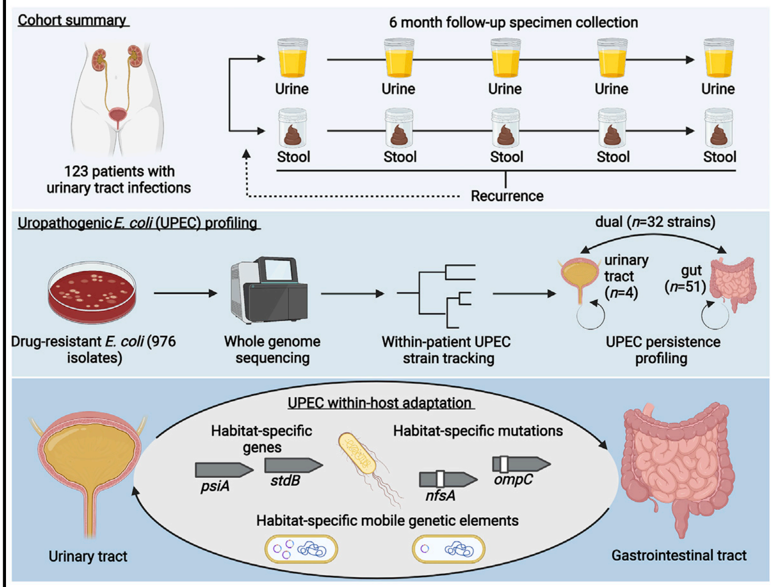
Resource

Cell Host & Microbe

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

Graphical abstract



Cohort summary
6 month follow-up specimen collection
Urine
Urine
Urine
Urine
Urine
Stool
Stool
Stool
Stool
Stool
123 patients with urinary tract infections
Recurrence
Uropathogenic *E. coli* (UPEC) profiling
dual (*n*=32 strains)
urinary tract (*n*=4)
gut (*n*=51)
Drug-resistant *E. coli* (976 isolates)
Whole genome sequencing
Within-patient UPEC strain tracking
UPEC persistence profiling
UPEC within-host adaptation
Habitat-specific genes
Habitat-specific mutations
psiA
stdB
nfsA
ompC
Habitat-specific mobile genetic elements
Urinary tract
Gastrointestinal tract

Highlights

- UPEC lineages persist within the gastrointestinal and urinary tracts of urinary tract infection (UTI) patients
- Habitat-specific selection impacts UPEC within-host adaptation
- Genomic plasticity facilitates UPEC niche adaptation
- Within-lineage genomic plasticity is facilitated by mobile genetic elements

Authors

Robert Thänert, JooHee Choi, Kimberly A. Reske, ..., Jennie H. Kwon, Gautam Dantas, CDC Prevention Epicenters Program

Correspondence

edubberk@wustl.edu (E.R.D.),
j.kwon@wustl.edu (J.H.K.),
dantas@wustl.edu (G.D.)

In brief

Thänert and Choi et al. show that lineages of uropathogenic *E. coli* (UPEC) persisting after resolution of symptomatic urinary tract infections adapt to the gastrointestinal and urinary environments. During this, mobile genetic elements facilitate the establishment of habitat-specific gene pools, providing UPEC with a mechanism to adapt to distinct physiological conditions.

Thänert et al., 2022, Cell Host & Microbe *30*, 1–14
July 13, 2022 © 2022 Elsevier Inc.
https://doi.org/10.1016/j.chom.2022.04.008

Cell Host & Microbe

CellPress

Resource

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

Robert Thänert,[1,2,13] JooHee Choi,[1,13] Kimberly A. Reske,[3] Tiffany Hink,[3] Anna Thänert,[1] Meghan A. Wallace,[3] Bin Wang,[1,2] Sondra Seiler,[3] Candice Cass,[3] Margaret H. Bost,[3] Emily L. Struttmann,[3] Zainab Hassan Iqbal,[3] Steven R. Sax,[3] Victoria J. Fraser,[3] Arthur W. Baker,[4,5] Katherine R. Foy,[4,5] Brett Williams,[6] Ben Xu,[6] Pam Capocci-Tolomeo,[7,8] Ebbing Lautenbach,[7,8,9] Carey-Ann D. Burnham,[2,3,10,11] Erik R. Dubberke,[3,*] Jennie H. Kwon,[3,*] Gautam Dantas,[1,2,11,12,14,*] and CDC Prevention Epicenters Program

[1]The Edison Family Center for Genome Sciences and Systems Biology, Washington University School of Medicine, St. Louis, MO, USA
[2]Department of Pathology and Immunology, Washington University School of Medicine, St. Louis, MO, USA
[3]Division of Infectious Diseases, Washington University School of Medicine, St. Louis, MO, USA
[4]Division of Infectious Diseases, Duke University School of Medicine, Durham, NC, USA
[5]Duke Center for Antimicrobial Stewardship and Infection Prevention, Durham, NC, USA
[6]Division of Infectious Diseases, Department of Internal Medicine, Rush Medical College, Chicago, IL, USA
[7]Department of Biostatistics, Epidemiology, and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
[8]Center for Clinical Epidemiology and Biostatistics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
[9]Division of Infectious Diseases, Department of Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
[10]Department of Pediatrics, Washington University School of Medicine, St. Louis, MO, USA
[11]Department of Molecular Microbiology, Washington University School of Medicine, St. Louis, MO, USA
[12]Department of Biomedical Engineering, Washington University in St. Louis, St. Louis, MO, USA
[13]These authors contributed equally
[14]Lead contact
*Correspondence: edubberk@wustl.edu (E.R.D.), j.kwon@wustl.edu (J.H.K.), dantas@wustl.edu (G.D.)

https://doi.org/10.1016/j.chom.2022.04.008

SUMMARY

Large-scale genomic studies have identified within-host adaptation as a hallmark of bacterial infections. However, the impact of physiological, metabolic, and immunological differences between distinct niches on the pathoadaptation of opportunistic pathogens remains elusive. Here, we profile the within-host adaptation and evolutionary trajectories of 976 isolates representing 119 lineages of uropathogenic *Escherichia coli* (UPEC) sampled longitudinally from both the gastrointestinal and urinary tracts of 123 patients with urinary tract infections. We show that lineages persisting in both niches within a patient exhibit increased allelic diversity. Habitat-specific selection results in niche-specific adaptive mutations and genes, putatively mediating fitness in either environment. Within-lineage inter-habitat genomic plasticity mediated by mobile genetic elements (MGEs) provides the opportunistic pathogen with a mechanism to adapt to the physiological conditions of either habitat, and reduced MGE richness is associated with recurrence in gut-adapted UPEC lineages. Collectively, our results establish niche-specific adaptation as a driver of UPEC within-host evolution.

INTRODUCTION

During infection or colonization, bacterial pathogens adapt to their host by optimizing their ability to replicate, disseminate, and evade host immunity (Marvig et al., 2015; Sheppard et al., 2018). Under strong selection, mutations arise continuously within persisting strains but rarely sweep to fixation, resulting in lasting intraspecies allelic diversity that provides a record of the pressures encountered (Lieberman et al., 2014; Lourenço et al., 2016). Parallel signatures in unrelated hosts can identify pathoadaptive mutations in persisting pathogens, revealing common drivers of within-host adaptation (Lieberman et al., 2011). Although a wealth of microbial whole-genome sequencing (WGS) data has identified common patterns of pathogen adaptation (pathoadaptation) (Didelot et al., 2016; Gatt and Margalit, 2021; Rossi et al., 2021), studies of within-host evolution have, with few exceptions (Lees et al., 2017; Young et al., 2017), been limited to specific niches in the human body. This potentially overlooks the effects of population dynamics of opportunistic pathogens occupying multiple body habitats.

Accordingly, there is a limited understanding of how physiological barriers between habitats may impact pathoadaptation.

One in four women affected by a urinary tract infection (UTI) will experience a recurrence (rUTI) within 6 months of initial infection (Foxman, 2014). Uropathogenic *Escherichia coli* (UPEC) are the most common cause of UTIs, accounting for approximately 75% of uncomplicated cases (Flores-Mireles et al., 2015). The recovery of UPEC from the gastrointestinal tract at asymptomatic time points before rUTI supports a model in which UPEC lineages can persist intestinally and reseed the urinary tract (Chen et al., 2013; Nielsen et al., 2016; Thänert et al., 2019). Emergence of uroadaptive mutations of the type 1 fimbrial adhesin FimH in urinary isolates that are rarely present in intestinal isolates suggests rapid adaptation to habitat-specific conditions (Chattopadhyay et al., 2007; Schwartz et al., 2013; Sokurenko, 2004; Weissman et al., 2007). In some patients, however, the absence of UPEC in the intestine and the recovery of UPEC from urine at asymptomatic time points (asymptomatic bacteriuria) highlight patient-specific patterns of persistence, which may differentially shape UPEC pathoadaptation (Thänert et al., 2019). It is unclear how the distinct physiological, metabolic, immunologic, and microbial conditions of the gastrointestinal and urinary tract impact UPEC within-host adaptation. Evolutionary trade-offs between habitats pose the question as to which molecular mechanisms enable UPEC lineages to persist, adapt, and cause repeated episodes of UTI (Bricio-Moreno et al., 2018).

Here, we investigate the hypothesis that habitat-specific selection in the gastrointestinal and urinary tracts differentially shapes UPEC within-host evolution. To assess this hypothesis, we characterize colonization patterns of persisting UPEC lineages in a longitudinal, prospective cohort of UTI patients. We contrast the adaptation of lineages colonizing the gastrointestinal tract with those also recovered from the urinary tracts to identify habitat-specific adaptations of UPEC. By characterizing within-lineage mutational diversity, we identify distinct patterns of within-host adaptation between UPEC colonization types, indicating that niche-adaptation shapes UPEC within-host adaptation. Finally, we identify mobile genetic elements (MGEs) as a major facilitator of within-lineage genomic plasticity associated with a pool of habitat-specific genes, putatively mediating UPEC fitness in either habitat and impacting recurrence in gut-adapted UPEC lineages.

RESULTS

UPEC lineages persist in the gastrointestinal and urinary tracts

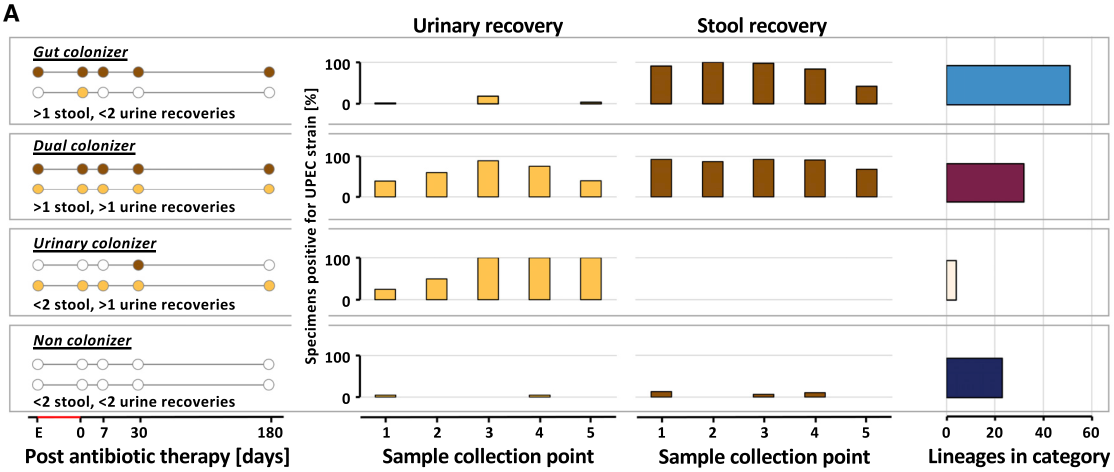
We collected 976 drug-resistant *Escherichia coli* isolates from a prospective, longitudinal cohort study of 123 patients presenting with symptomatic UTI caused by antibiotic resistant (AR) uropathogens. *E. coli* were cultured from 1,752 stool and urine specimens collected at study enrollment and subsequently at 10 asymptomatic time points over a 6-month follow-up period using a home shipment protocol (see STAR Methods). Patients who experienced a rUTI within the follow-up period were able to restart sample collection (42 patients, 34.15%).

To identify UPEC lineages persisting within patients, we characterized genomic relatedness of same-patient isolates using WGS of all 976 *E. coli* isolates (average of 8.2 isolates/patient; Data S1). Following methodologies implemented in similar studies (Bronson et al., 2021; Coll et al., 2017), we profiled single-nucleotide polymorphism (SNP) distances based on patient-specific core-genomes to differentiate isolates belonging to the same *E. coli* lineage as the causative agent of the index UTI from isolates representing distinct subspecies clusters. We observed that within-patient SNP distances followed a multimodal distribution (Figure S1A), with a notable paucity of within-patient pairwise isolate SNP distances between 500 and 10,000 SNPs. To assess plausibility of 500 SNPs as the upper limit of a UPEC lineage definition for this study, we estimated the average duration since last common ancestor (LCA) for each lineage. For each persistent lineage, we generated whole-genome SNP trees based on lineage-specific reference assemblies and calculated the median branch length. We then divided this value by a previously reported estimated rate of *E. coli* base substitution (8.9×10^{-11} bp/generation) (Wielgoss et al., 2011). Importantly, because our estimate is based on within-gut *E. coli* generation times, values for urinary persisters are likely less accurate. We estimated an average of $\sim$0.33 (0–5.39, Figure S1B) years since the LCA, consistent with the reported history of recurrent UTIs in our patient cohort. Whole-genome pairwise average nucleotide identity (ANI) values calculated between same-patient isolates further showed that isolates typed to the same lineage based on the 500 core genome SNP cutoff exhibited high pairwise ANI values (99.991% [0.0127]—median [IQR]), whereas isolates from the same patient typed into distinct lineages and from distinct patients displayed lower, variable ANI values (97.288% [1.531], 97.268% [1.588], Figures S1C and S1D).

We applied the 500 core genome SNP cutoff to all isolates cultured from the same patient and identified a total of 187 distinct subspecies clusters of *E. coli* (hereafter referred to as "lineages"—Figure S1; Data S1). In total, 702 isolates recovered at asymptomatic time points belonged to 119 lineages that were isolated as the causative agent of a UTI (diagnostic urinary isolate: DxU) and were defined as UPEC for the purpose of this study. The majority of these lineages belonged to the pandemic extraintestinal pathogenic *E. coli* (ExPEC) sequence type complexes (STcs) 131 (36.97%, Serotypes O25:H4 and O16:H5), predominantly ST131-fimH30, and STc14 (21.85%, Serotype O75:H5; Data S1), predominantly ST1193 (Table S1; Figure S2).

We characterized asymptomatic persistence of UPEC lineages based on longitudinal recovery of same-lineage *E. coli* from patient-matched urine and stool specimens, using standard-of-care clinical microbiology culturing methods (Figure 1A; STAR Methods). We classified three distinct patterns of UPEC lineage persistence (see STAR Methods): (1) gastrointestinal persistence ("gut colonizer," 51 lineages, 46.4%), (2) persistence in both habitats ("dual colonizer", 32 lineages, 29.1%), or (3) persistence in the urinary tract ("urinary colonizer," 4 lineages, 3.6%, Figure 1A). Isolates belonging to these categories were used in downstream analysis to investigate UPEC within-host evolution. In 23 patients (20.9%), we did not find evidence for UPEC persistence in either the urinary or the gastrointestinal tract. Although sequence type (ST) distribution did not differ between persistence types (Figure 1B), STs of nonpersisting lineages differed significantly from that of persisters (Figure 1C, Fisher's exact test $p < 0.001$), with ST131 and ST1193 underrepresented among nonpersisting lineages (Fisher's exact test

Please cite this article in press as: Thänert et al., Persisting uropathogenic *Escherichia coli* lineages show signatures of niche-specific within-host adaptation mediated by mobile genetic elements, Cell Host & Microbe (2022), https://doi.org/10.1016/j.chom.2022.04.008



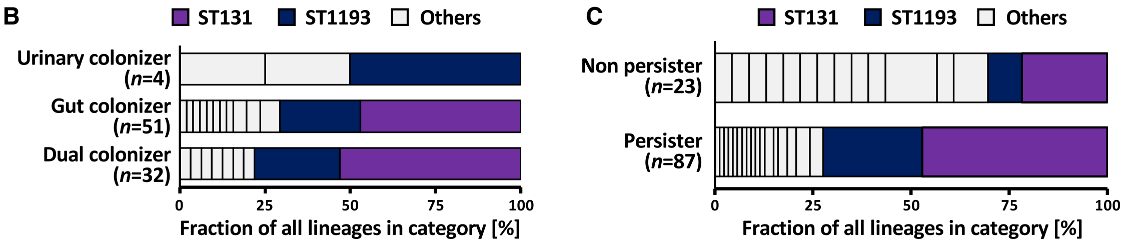


Figure 1. Persistent UPEC lineages group into distinct colonization patterns

(A) Schematic representation of UPEC colonization patterns (left) as determined by recovery from stool (brown circles) and urine (yellow circles) from UTI patients with available DxU isolates. The definition for each colonization type is given below the schematic. UPEC lineages (n = 119) are classified into four persistence types: gut colonizer, dual colonizer, urinary colonizer, and noncolonizer. (Middle) UPEC lineage presence at follow-up sample collection points as determined by whole-genome sequencing of isolates (key: 1, enrollment; 2, 0–3 days postantibiotic treatment [pAT]; 3, 7–14 days pAT; 4, 30–60 days pAT; 5, 150–180 days pAT). Bars indicate the fraction of patients' urine (yellow) and stool (brown) specimens positive for the disease-causing UPEC lineage at each sampling point. Patients are grouped by UPEC lineage persistence type. Only data from the first episode caused by a UPEC lineage is shown. (Right) Number of UPEC lineages falling into each colonization category (gut colonizer = 51, dual colonizer = 32, urinary colonizer = 4, and noncolonizer = 23). Boxes group together panels showing data of the same persistence type.

(B) Sequence types (STs) are evenly distributed between UPEC persistence types. Prevalence of the two dominant STs, ST131 and ST1193, is highlighted in color.

(C) ST composition varies significantly between persisting and nonpersisting lineages (n = 110 lineages, Fisher's exact test, p < 0.001). ST131 (light purple) and ST1193 (dark purple) are significantly underrepresented in the set of nonpersisting UPEC lineages (n = 110 lineages, Fisher's exact test, p < 0.001). Prevalence of the two dominant STs, ST131 and ST1193, is highlighted in color.

$p < 0.001$). Interestingly, dual colonizers were associated with the majority of rUTI events attributable to a specific lineage during the 6-month follow-up period (57.9% (11/19 lineages), 36.8% (7/19) gut colonizer, and 5.3% (1/19) urinary colonizer). Collectively, these observations suggest that colonization of the gut (gut colonizer) or both environments (dual colonizer) describe the majority of persistence of UPEC lineages.

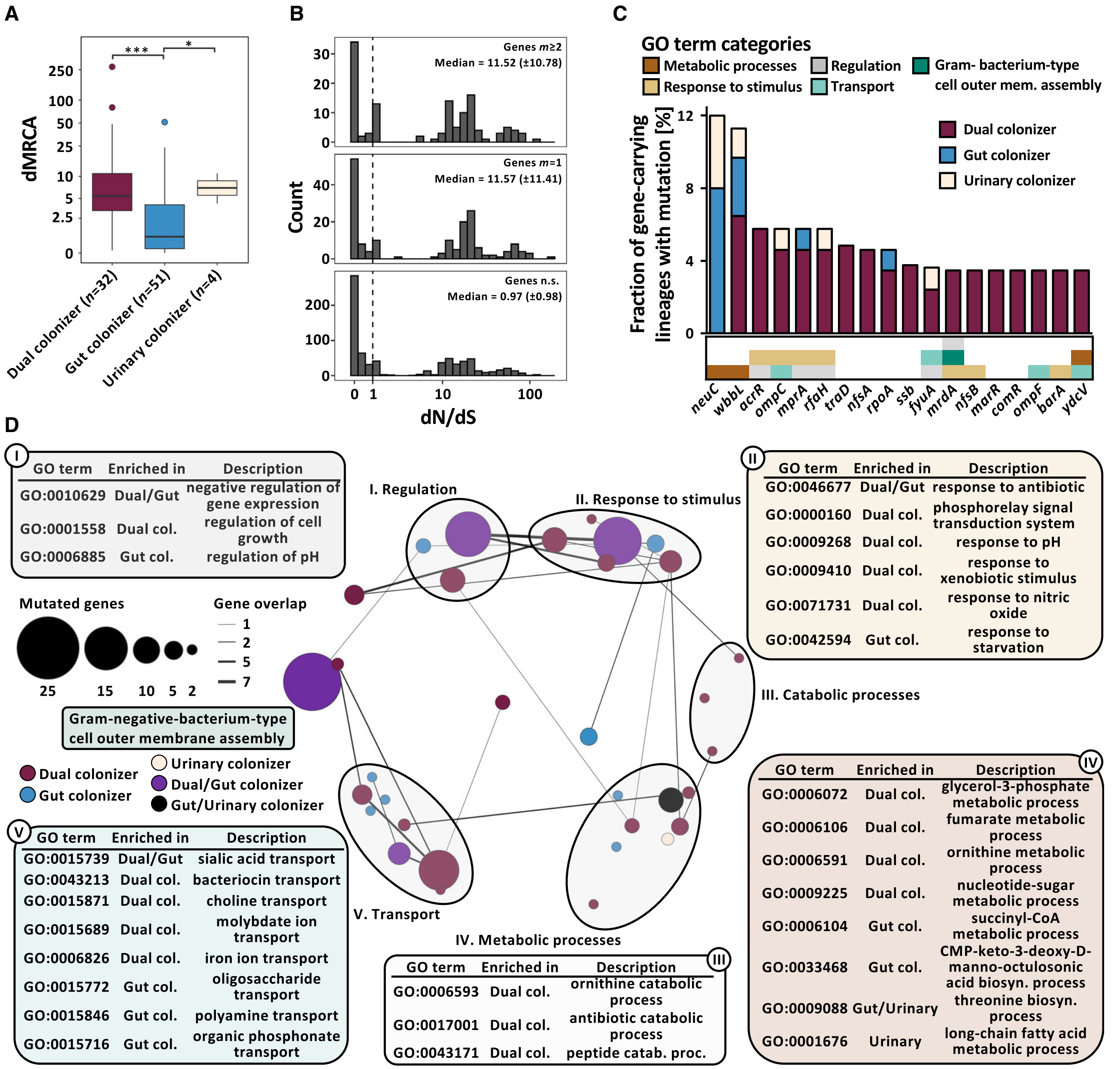
Urinary persistence is associated with increased allelic diversity of UPEC lineages

To assess the impact of environmental selection on UPEC within-host evolution, we profiled the within-host adaptation of UPEC lineages in their persistence habitats (i.e., gut colonizers in the gut, dual colonizers in gut and urinary tract, and urinary colonizers in the urinary tract). We identified all within-lineage SNPs by aligning reads against lineage-specific pseudo-assemblies, as previously described (Thänert et al., 2019; Zhao et al., 2019).

By inferring the ancestral sequence through maximum parsimony, we found that urinary persistence is associated with significantly increased distance to the most recent common ancestor (dMRCA) compared with gut-colonizing lineages (Figure 2A, n = 87 lineages, Kruskal-Wallis p = 1.38e−5, Dunn post hoc test, gut versus dual colonizer p = 2.39e−5, gut versus urinary colonizer p = 3.32e−2). These observations are consistent with two potential explanations: first, urinary persistence may enable UPEC lineages to persist within a host for longer durations. Alternatively, considering that *E. coli* are native to the gut, disparate selective pressure in the urinary tract could result in habitat-specific fitness maxima distinct from those of the gastrointestinal tract and extend the spectrum of positively selected mutations, diversifying the allelic repertoire of persisting UPEC lineages.

UPEC niche-specific adaptation shapes within-host adaptation

To test the hypothesis that urinary persistence results in trajectories of within-host adaptation distinct from those observed in the gut, we annotated within-lineage allelic diversity (SNPs,



I

GO term	Enriched in	Description
GO:0010629	Dual/Gut	negative regulation of gene expression
GO:0001558	Dual col.	regulation of cell growth
GO:0006885	Gut col.	regulation of pH

II

GO term	Enriched in	Description
GO:0046677	Dual/Gut	response to antibiotic
GO:0000160	Dual col.	phosphorelay signal transduction system
GO:0009268	Dual col.	response to pH
GO:0009410	Dual col.	response to xenobiotic stimulus
GO:0071731	Dual col.	response to nitric oxide
GO:0042594	Gut col.	response to starvation

IV

GO term	Enriched in	Description
GO:0006072	Dual col.	glycerol-3-phosphate metabolic process
GO:0006106	Dual col.	fumarate metabolic process
GO:0006591	Dual col.	ornithine metabolic process
GO:0009225	Dual col.	nucleotide-sugar metabolic process
GO:0006104	Gut col.	succinyl-CoA metabolic process
GO:0033468	Gut col.	CMP-keto-3-deoxy-D-manno-octulosonic acid biosyn. process
GO:0009088	Gut/Urinary	threonine biosyn. process
GO:0001676	Urinary	long-chain fatty acid metabolic process

V

GO term	Enriched in	Description
GO:0015739	Dual/Gut	sialic acid transport
GO:0043213	Dual col.	bacteriocin transport
GO:0015871	Dual col.	choline transport
GO:0015689	Dual col.	molybdate ion transport
GO:0006826	Dual col.	iron ion transport
GO:0015772	Gut col.	oligosaccharide transport
GO:0015846	Gut col.	polyamine transport
GO:0015716	Gut col.	organic phosphonate transport

III

GO term	Enriched in	Description
GO:0006593	Dual col.	ornithine catabolic process
GO:0017001	Dual col.	antibiotic catabolic process
GO:0043171	Dual col.	peptide catab. proc.

Figure 2. Niche-specific adaptation shapes UPEC within-host adaptation

(A) Boxplot of lineage dMRCA values (n = 87 lineages, Kruskal-Wallis p = 1.38e−5, Dunn post hoc test, gut versus dual colonizer p = 2.39e−5, gut versus urinary colonizer p = 3.32e−2). Outliers (outside 1.5× interquartile range) are depicted as points. Whiskers represent 1.5× interquartile range. Upper, middle, and lower box lines indicate 75th, 50th, and 25th percentiles, respectively.

(B) Histogram of dN/dS values of genes with signatures of nonrandom mutation (permutation test, $p < 0.05$) mutated in parallel across more than two lineages ($m \geq 2$, top) or in one lineage ($m = 1$, middle) and in genes nonsignificant in permutation test (bottom). Median and median absolute deviation (MAD) are given for both gene groups. Dashed vertical line indicates neutral selection at dN/dS = 1.

(C) Genes found to be mutated in parallel in ≥3 lineages, normalized by the total number of gene-carrying lineages. Hypothetical genes are not shown. Color of the bar corresponds to colonization type in which mutations were found (gut colonizer, blue; dual colonizer, maroon; and urinary colonizer, light yellow). Color bar below the histogram provides GO category (as shown in Figure 2D) for all genes with GO terms annotation found to be significantly enriched in a colonization type.

(D) Network visualization of GO terms significantly overrepresented in the pool of genes with nonrandom signature of selection within-lineages as defined by the permutation test. Bubble size represents the number of mutations in genes categorized into each GO term. Color of bubbles corresponds to colonization type GO terms were enriched in (gut colonizer: blue; dual colonizer: maroon; urinary colonizer: light yellow; gut/dual colonizer: purple; and gut/urinary colonizer: black). GO terms were clustered semantically into the 2D space using REVIGO. Circles group together semantically related GO terms.

insertions, deletions) at the gene level. We implemented permutation tests, randomly distributing the number of observed mutations over each lineage's pseudoassembly to generate a null distribution. We then compared observed against expected frequencies to identify genes with signatures of nonrandom evolution across lineages. Permutation tests were conducted independently for colonization types to characterize the effect of distinct persistence patterns.

Our analysis identified 253 genes with mutational signatures indicating nonrandom selection (n = 87 lineages, permutation test, confidence interval 95%). To validate that positive selection drives mutations in this gene set, we calculated per-gene dN/dS ratios, a canonical metric for selection. We found a robust enrichment of elevated dN/dS values in both genes mutated in a single lineage (Figure 2B, m = 1, median 11.57 ± 11.41 median absolute deviation [MAD]) and in parallel across multiple lineages ($m \geq 2$, 11.52 ± 10.78) compared with genes nonsignificant by permutation test (median 0.97 ± 0.98). Consistent with this observation, the group dN/dS value for all genes significant by permutation test and mutated in parallel across lineages, 1.34 (0.96–2.02, 95% confidence interval by binomial sampling) indicated that adaptation drives mutation in these genes. In contrast, genes carrying mutations but nonsignificant by permutation test were under purifying selection (group dN/dS 0.32, 0.30–0.35), consistent with previous literature (Zhao et al., 2019).

Mutation of a single gene (*wbbL*) was observed in all colonization types, whereas 12 genes were shared between at least two groups (Data S2). Virulence- and drug-associated genes were mutated in parallel frequently across colonization types (Figure 2C), including capsule-related genes *neuC* (dN/dS 7.3) and *mprA* (dN/dS 17.5), as well as *wbbL* (dN/dS 59.4), coding a rhamnosyl transferase critical for O-antigen synthesis. As both capsule and O-antigen directly affect UPEC fitness *in vivo* (Sarkar et al., 2014), these mutations may also affect UPEC persistence. Further, genes implicated in antibiotic resistance, including *ompC* (dN/dS 17.8), *acrR* (dN/dS 5.8), *nfsA* (dN/dS 17.8), and *nfsB* (dN/dS 10.9) (Choi and Lee, 2019; Osei Sekyere, 2018; Su et al., 2007), were found to be under positive selection across lineages. Interestingly, mutations of the biofilm suppressing antiterminator RfaH encoding gene (dN/dS 33.5) were exclusively found in lineages persisting within the urinary tract. Biofilms are critical UPEC colonization factors, enabling adhesion to abiotic (catheter) and biotic (urinary tract) surfaces (Beloin et al., 2006).

To assess functional adaptation of UPEC during persistence comprehensively, we performed gene ontology term overrepresentation analysis (GOOA) in the pool of all genes mutated within-lineages that exhibited a signature of nonrandom selection (Data S3). Strikingly, functional categories under selection differed between colonization types, with only a small set of core functions (sialic acid transport, membrane assembly, antibiotic resistance, and negative regulation of transcription) found to be under selection in multiple colonization types (Figure 2D). Distinct transport capabilities, response to environmental stressors, metabolic processes, and regulatory functions were selected in gut-restricted and dual colonizers (Figure 2D), indicating that distinct persistence patterns differentially shape within-host adaptation of persisting UPEC lineages. Functions found to be under selection in dual colonizers, including iron-ion transport, response to pH, response to nitric oxide, ornithine metabolism, or fumarate metabolism (Figure 2D), have been linked to urinary fitness of UPEC and likely are direct adaptations towards the habitat-specific conditions of the urinary tract (Hibbing et al., 2020; Mann et al., 2017). Collectively, these results support the idea that niche-specific selection shapes the evolutionary trajectories of persisting UPEC, altering the landscape of positively selected functionalities for multihabitat lineages.

Within-host adaptation of UPEC impacts resistance phenotypes

We observed that 79.4% of the within-lineage allelic diversity in genes mutated in parallel among dual colonizing lineages was structured by habitat, with mutations only occurring in a single habitat within a lineage (Figure 3A). Similarly, when including 71 additional urinary isolates from the 51 gut-colonizing lineages and implementing our permutation test to identify genes under positive selection (Data S2), we found that an even larger fraction of mutations in genes with parallel signature across lineages was only found in isolates cultured from one sample type (93.5%, Fisher's exact test, p = 0.001). As urinary colonizers had no representative gut isolates, they were not included in this analysis. We reasoned that this phenomenon could result from two potential processes: (1) a consequence of genetic bottlenecks upon habitat transition, or (2) habitat-specific selection resulting in divergent subpopulations within the same lineage in the gastrointestinal and urinary tracts.

To test whether niche-specific adaptation may in fact play a role in shaping allelic breakdown along habitat lines in persisting UPEC lineages, we focused on a subset of mutations with a tractable phenotypic impact. We had previously observed strong selection for mutations in antibiotic resistance-associated genes during persistence (Figure 2D) and reasoned that niche-specific adaptation would result in niche-dependent resistance phenotypes. Therefore, we identified mutations in antibiotic resistance genes (ARGs) and profiled isolate resistance phenotypes for both dual and gut-colonizing lineages. We found that the nonsynonymous *ompC* R191C mutation in dual colonizing lineage WU-041_1 was exclusively found in urinary isolates and coincided with the gain of ampicillin/sulbactam resistance (Figure 3B). Importantly, we found that nonsynonymous mutations of *ompC*, including another instance of R191C in lineage PN-004_1, were restricted to urinary isolates. Similarly, we found *nfsA* Q191* mutation in gut-colonizing lineage WU-046_2 exclusively in isolates cultured from urine specimens during symptomatic disease and immediately preceding recurrence (Figure 3C), associated with the gain of phenotypic nitrofurantoin resistance. Moreover, we identified that resistance-conferring mutations of *nfsA*, including another premature stop codon in lineage PN-004_1 (*nfsA* W237*), were restricted to urinary isolates. Together, these findings indicate niche-dependent fitness benefits of mutations in these two genes and a role of niche-specific adaptation in shaping within-host adaptation of persisting UPEC lineages.

We further reasoned that if these observed mutations provide UPEC with direct fitness benefits, they may also be found in UPEC genomes sequenced in different studies. To test this, we downloaded a set of 703 UPEC genomes previously curated from multiple studies (Biggel et al., 2020) and profiled allelic identify of *ompC* and *nfsA* at all positions observed to be variable in

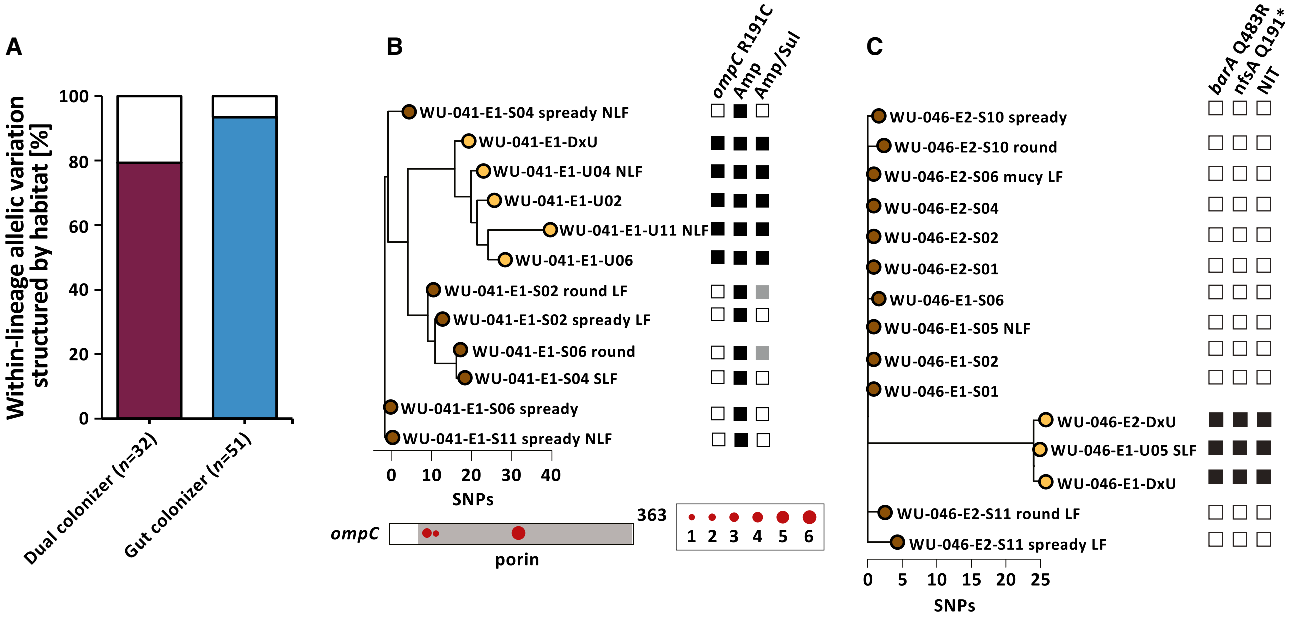


Figure 3. UPEC niche-specific adaptation impacts antibiotic resistance phenotypes

(A) The majority of allelic diversity in genes found to be mutated in parallel in gut and dual colonizers is structured by habitat (Fisher's exact test p = 0.001). Color of the bar corresponds to either dual colonizers (maroon) or gut colonizers (blue).

(B) (Top) Phylogeny of lineage WU-041_1 with annotated nonsynonymous *ompC* mutation and corresponding phenotypic resistance to ampicillin/sulbactam. Black squares denote gene presence or antibiotic resistance. White squares indicate gene absence or drug susceptibility. Gray squares indicate intermediate drug susceptibility. Phylogeny is unrooted based on SNP distances. (Bottom) SNP locations on the *ompC* gene. The porin domain is annotated in gray. Circle size corresponds to number of isolates carrying that mutation.

(C) Lineage WU-046_2 exhibited nonsynonymous *barA* and *nfsA* mutations in urinary isolates only, corresponding to phenotypic resistance to nitrofurantoin. Phylogeny is unrooted based on SNP distances. Labels as in (B).

this study. We found 2/4 *ompC* and 1/4 *nfsA* mutations identified in our study in the set of published UPEC genomes (Figure S3). This suggests that similar selective pressures to the ones characterized in this study are shaping adaptation of *ompC* and *nfsA* in the larger UPEC population.

Genomic plasticity facilitates UPEC niche adaptation

Differential abundance of genes within an otherwise clonal population, termed genomic plasticity, can facilitate rapid adaptation of bacterial pathogens to new environments (Darch et al., 2015; Gabrielaite et al., 2020; Hammond et al., 2020). The distinct physiological conditions of the gastrointestinal and urinary tracts are likely to require disparate metabolic and colonization factors. We therefore hypothesized that genomic plasticity may enable persisting UPEC lineages to maintain fitness in both the gastrointestinal and urinary environments.

Persisting gut populations of gut colonizers exhibited more homogeneous gene profiles than dual colonizers (Figure 4A, n = 87 lineages, Kruskal-Wallis test p = 0.009, Dunn post hoc test p = 0.012), indicating that habitat diversification is associated with a larger pool of flexible genes. We hypothesized that this difference may be caused by greater interhabitat heterogeneity in persisting dual colonizers not observed in gut populations. To test this hypothesis, we analyzed interhabitat similarity of same-lineage isolate gene profiles, including all 71 urinary isolates from the 51 gut-colonizing lineages. We found that isolates collected from the same habitat were significantly more likely to carry similar genes, although colonization types did not differ significantly (Figure 4B, n = 87 lineages, two-way ANOVA, habitat p = 5.94e−4, colonization type p > 0.05), suggesting that genomic plasticity contributes to niche adaptation of all persisting UPEC lineages.

1,553 genes were restricted to either urinary or stool isolates in the 83 UPEC gut and dual colonizing lineages and therefore may play a role in habitat adaptation (Figure 4C; Data S4). Interestingly, three plasmid-associated genes, *psiA*, *yggR*, and *stbB*, were found to be restricted to gut isolates in 5 independent lineages. To comprehensively profile functional selection on the variable genetic portion of each lineage in either habitat, we performed GOOA on the pool of habitat-specific genes. We identified nitrogen compound and iron uptake mechanisms as key factors for urinary adaptation in both dual and gut-colonizing lineages (Figures 4D and S4A, Fisher's exact test GO:0071705 p = 0.018 [dual] and p = 0.002 [gut], GO:0055072 p = 1.81e−4 and p = 2.51e−7, GO:0044718 p = 0.024 and p = 0.018, Data S3). Specifically, systems facilitating the uptake of ferric-citrate complexes that are abundant in urine were found to be habitat-associated in gut as well as dual colonizers (Figure S4; Robinson et al., 2018).

Few functionalities were overrepresented in stool isolates of dual colonizing lineages (Figure 4E). Conversely, the gut-specific gene pool of gut colonizers exhibited enrichment of multiple functionalities implicated in *E. coli* gut colonization and virulence, including antibiotic resistance, fumarate transport, type IV secretion, and pilus assembly (Elhenawy et al., 2021; Jones et al.,

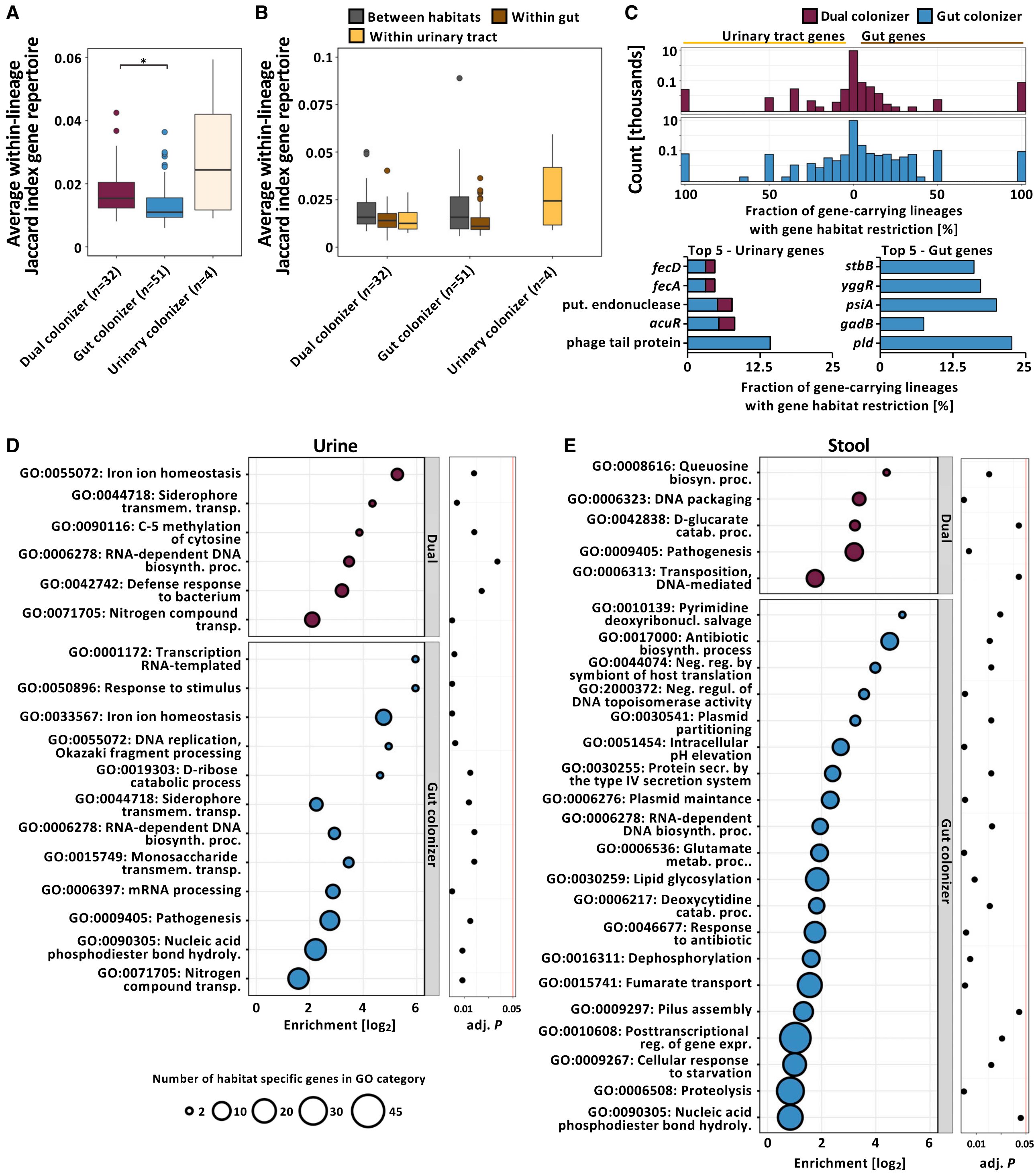


Figure 4. Persisting UPEC lineages exhibit niche-specific genomic plasticity

(A) Boxplot of average within-lineage Jaccard distances based on gene presence/absence data (n = 87 lineages, Kruskal-Wallis test p = 0.009, Dunn post hoc test gut versus dual colonizer p = 0.012). Outliers (outside 1.5× interquartile range) are depicted as points. Whiskers represent 1.5× interquartile range. Upper, middle, and lower box lines indicate 75th, 50th, and 25th percentiles, respectively.

(B) Average between- and within-habitat lineage Jaccard distances based on gene presence/absence data of same-lineage isolates by colonization type (n = 87 lineages, two-way ANOVA, habitat p = 5.94e−4, colonization type p > 0.05). Outliers (outside 1.5× interquartile range) are depicted as points. Whiskers represent 1.5× interquartile range. Upper, middle, and lower box lines indicate 75th, 50th, and 25th percentiles, respectively. Colors correspond to within-lineage comparison (between habitats: gray; within gut: brown; and within urinary tract: yellow).

(legend continued on next page)

2011; Spaulding et al., 2017). Notably, GO terms associated with plasmid maintenance genes were found to be enriched in intestinal isolates of gut-colonizing lineages, commonly coinciding with presence/absence of virulence and resistance genes (Figures S4A-S4D, Fisher's exact test GO:0030541 p = 0.044, GO:0006276 p = 1.77e−3, Data S3). We therefore hypothesized that MGEs may facilitate niche adaptation in persisting UPEC lineages.

Heterogeneous MGE carriage facilitates habitat-associated genomic plasticity

To evaluate the role of MGEs in the genomic plasticity of persisting UPEC lineages, we comprehensively identified regions of differential coverage in isolates of the same lineage as previously described (Zhao et al., 2019). These regions are candidate MGEs differentially abundant in isolates of the same lineage. We annotated the list of putative MGEs (Figure 5A; Data S5), combining *in silico* detection of plasmidic contigs and database-driven annotation of *de novo* identified MGEs as previously described (see STAR Methods; Figure S5; Durrant et al., 2020; Thänert et al., 2019). In total, 57.1% (887/1,553 genes) of the habitat-specific gene pool mapped back to putative MGEs. As expected, we found ARGs, proteolysis, and conjugation mechanisms associated with plasmidic MGEs (Figure 5-B). Pathofunctions that were implicated as habitatspecific in our previous analysis, including iron import systems, type II and type IV secretion systems, and cell adhesion genes, were found to be enriched within MGE subcategories.

To profile potential sharing of UPEC MGEs with other species, we mapped all MGE contigs to the NCBI nucleotide database. We found that plasmidic MGEs had the broadest putative host range (Figure S6A). However, plasmidic MGEs exclusively identified in urinary isolates exhibited a trend toward a narrower host range compared with those found in the gut (Figure S6A, ANOVA p = 0.053, Tukey post hoc test versus gut-exclusive p = 0.053, versus dual-habitat p = 0.057). Moreover, these MGEs were significantly less likely to be mapped to common gut residents, including *Salmonella enterica*, *Citrobacter freundii*, or *Enterobacter cloacae* (Figure S6B, Fisher's exact test, FDR corrected p < 0.05), indicating that gut-associated plasmidic MGEs are more likely to be shared with other gut residents.

Contrary to the high intrahabitat dissimilarity of lineage MGE profiles in urinary colonizers (Figure 5C), we observed homogeneous within-habitat MGE carriage in dual and gut-colonizing lineages. In gut-colonizing lineages, heterogeneity of MGE carriage was significantly elevated across habitats compared to within-habitat, as well as significantly larger compared to dual colonizers (Figure 5C, n = 87 lineages, two-way ANOVA p ≤ 1.57e−5, Tukey post hoc p < 0.001 and p = 0.014, respectively, Data S5). These results suggest that multihabitat selection in dual colonizers may stabilize the MGE pool across habitat boundaries. Urinary isolates' MGE pools were significantly smaller compared tointestinal isolates' pools (Figure 5D, n = 87 lineages, two-way ANOVA p = 0.042). Moreover, we found that habitat-specific genes from metabolic, antibiotic resistance, and virulence-associated functional categories were mapped to MGEs exclusively present in urinary or stool isolates (Figures 5E and 5F). These observations suggest that mobilization of key functions associated with adaptation to either habitat, such as iron acquisition or nitrogen compound uptake in the urinary tract (Figure 4D), may play a key role in UPEC niche adaptation.

Interestingly, the association of MGEs with ARGs resulted in a pool of ''hidden'' ARGs not observed in the DxU isolate but present in other isolates of the same lineage (Figure S7). Isolates harboring ''hidden'' ARGs frequently showed concordant variation in their replicon profile compared to the DxU isolate (66/78 cases, 84.6%), corroborating differential resistance plasmid carriage as a potential driver of within-lineage plasticity of ARGs.

Decreased MGE richness is associated with rUTI in gut-colonizing UPEC lineages

Based on our observation of decreased urinary MGE richness, we hypothesized that MGEs may hamper urinary fitness of gut-adapted UPEC lineages, resulting in an inverse relationship between MGE richness and the likelihood of a lineage to cause a rUTI during the follow-up period. In fact, we found that gut colonizer lineages causing rUTI exhibited significantly lower average MGE richness per isolate compared to their non-rUTI counterparts (Figure 6A, n = 43 lineages, Welch's t test, FDR corrected p = 0.001). Notably, no such relationship was observed for dual colonizers (n = 26 lineages, Welch's t test, FDR corrected p = 0.884).

Despite considerable variability in the functional composition of their mobilized gene pool, no functional category was significantly enriched after correcting for multiple hypothesis testing in either rUTI or non-rUTI lineages (Figure S8A, n = 69 lineages, Fisher's exact test, all FDR corrected p > 0.05). However, we observed a trend toward lower mobilized ARG richness in rUTI lineages compared with non-rUTI lineages (Figures S8B and S8C, n = 69 lineages, Wilcoxon rank-sum test p = 0.055). We found no difference between the mobilized ARG richness of UPEC persistence types (Figures S8D and S8E, n = 87 lineages, Kruskal-Wallis p = 0.231).

To identify mobilized functions negatively impacting urinary fitness of gut-adapted UPEC lineages, we characterized the habitat association of each putative MGE for all gut colonizer lineages. We identified a large gut-specific MGE pool (238/457, 52.08%) absent from any urinary isolate. GOOA of genes present on these gut-specific MGEs identified 9 of 94 GO categories

(C) (Top) Two-sided histogram of within-lineage habitat-specific genes of dual (maroon) and gut (blue) colonizers. Urinary-specific genes are shown on the left. Gut-specific genes are shown on the right. (Bottom) Genes most frequently found to be urine (left) or gut (right) specific across lineages, normalized by the total number of gene-carrying lineages. Bar color corresponds to the colonization type a gene was found in as habitat-specific. Hypothetical genes are not shown.

(D) Overrepresented GO terms associated with urine specific genes of dual (top, maroon) or gut colonizers (bottom, blue). Bubble size corresponds to the number of habitat-specific genes in each GO term.

(E) Overrepresented GO terms associated with stool specific genes, using the same formatting as in (D).

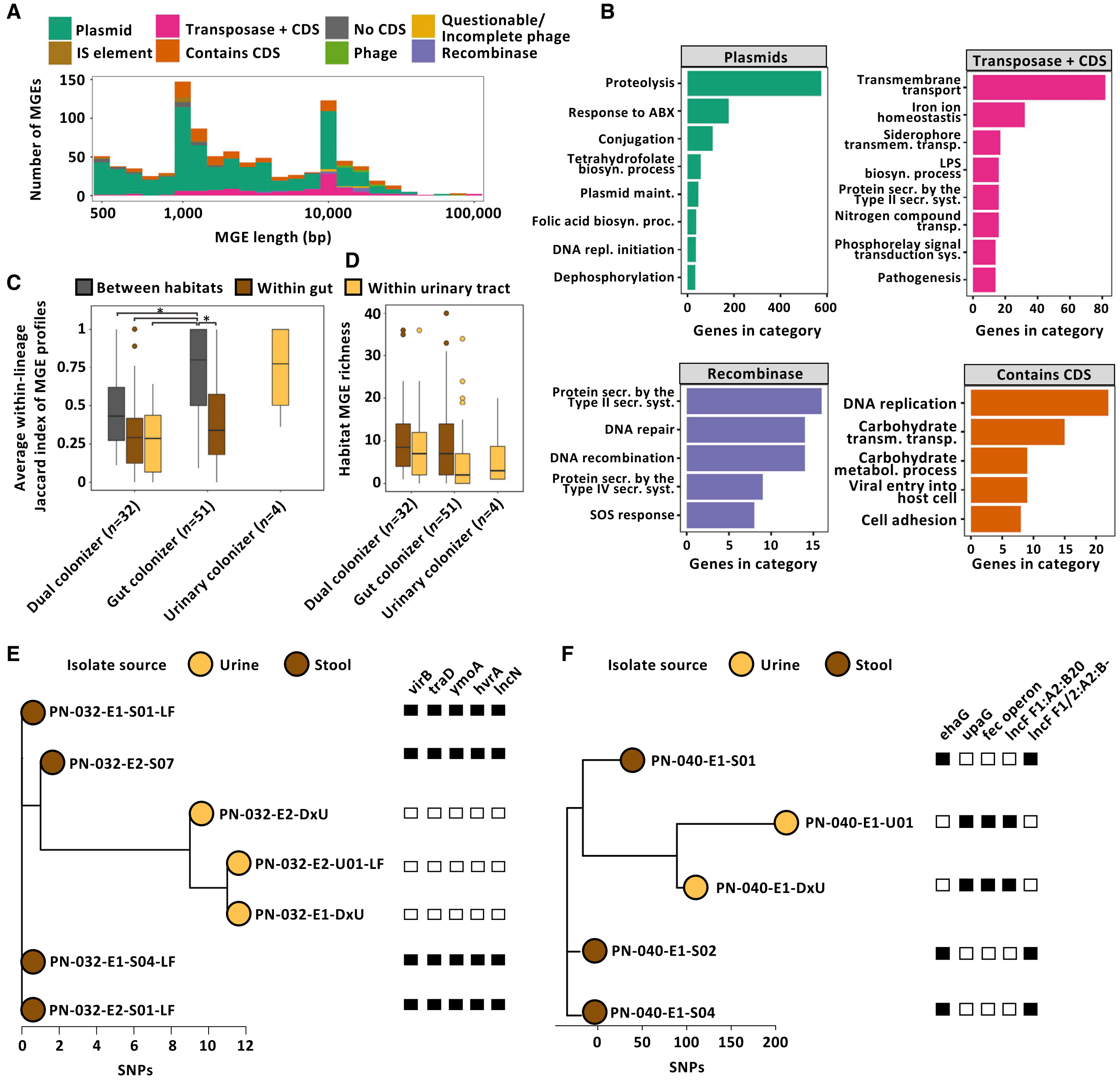


Figure 5. Mobile genetic elements drive niche-specific genomic plasticity of UPEC

(A) Visualization of within-lineage MGEs. Element length (log-scale) is plotted against element count. IS, insertion sequence; CDS, coding sequence.

(B) GO terms overrepresented in selected MGE subclasses.

(C) Boxplot of average within-lineage Jaccard distance based on MGE presence/absence data of same-lineage isolates between habitats (gray), within gut (brown), and within urine (yellow) grouped by colonization type. All comparisons are statistically significant (n = 87 lineages, two-way ANOVA p ≤ 1.57e−5, Tukey post hoc gut colonizer within-gut versus between habitats p < 0.001, gut colonizer between habitat versus dual colonizer between habitat p = 0.014).

(D) MGE richness is larger in gut compared to urine isolates (n = 87 lineages, two-way ANOVA p = 0.042). Outliers (outside 1.5× interquartile range) are depicted as points. Whiskers represent 1.5× interquartile range. Upper, middle, and lower box lines indicate 75th, 50th, and 25th percentiles, respectively.

(E) Unrooted phylogeny of lineage PN-040_1 based on SNP distances annotated with selected habitat-specific genes. Relative short-read coverage over selected, habitat-specific MGEs harboring depicted genes is shown.

(F) Unrooted phylogeny of lineage PN-004_1 based on SNP distances annotated with selected habitat-specific genes. Relative short-read coverage over selected, habitat-specific MGEs harboring depicted genes is shown.

significantly depleted in urinary isolates (Figure 6B, Fisher's exact test, FDR-corrected p value < 0.05, Data S3), including DNA-related, lipid biosynthetic, and type IV secretion system processes. Interestingly, although some gut-specific GO categories were absent from the MGE pool of rUTI-causing gut colonizers (e.g., antibiotic biosynthesis, tryptophan biosynthesis), these GO terms were in general not underrepresented in their MGE pool (Figure 6B).

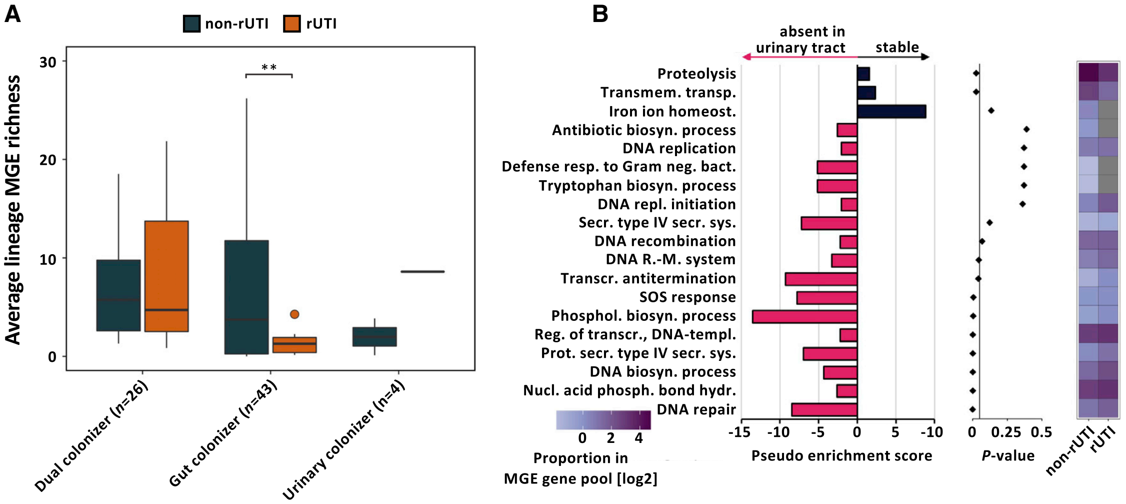


Figure 6. Gut-colonizing UPEC lineages causing rUTI exhibit decreased MGE richness

(A) MGE richness of lineages causing rUTI during the follow-up period and non-rUTI lineages parsed by colonization type (n = 73 lineages, Welch's t test, FDR-corrected gut colonizer p = 0.001, dual and urinary colonizer FDR corrected p > 0.05). Outliers (outside 1.5× interquartile range) are depicted as points. Whiskers represent 1.5× interquartile range. Upper, middle, and lower box lines indicate 75th, 50th, and 25th percentiles, respectively.

(B) (Left) Pseudoenrichment score of GO terms in the pool of MGEs absent or stable in urinary isolates of gut-colonizing UPEC lineages. Top 19 GO categories by p value are visualized. Pink bars indicate gene associated GO terms overrepresented in the urine instable MGE pool, and black bars indicate GO terms enriched in the pool of MGEs stable in urinary isolates. Pseudoenrichment score was calculated by adding one count to all GO categories. (Middle) p values for each GO category determined from overrepresentation analysis using hypergeometric distribution. (Right) Proportion of each visualized GO term in the MGE associated gene pool of rUTI and non-rUTI-causing lineages of gut-colonizing UPEC. Gray tiles indicate absence of a GO term in the MGE gene pool.

DISCUSSION

Invasion and colonization of the urinary from the gastrointestinal tract is the first step in the infectious cascade of the majority of UTIs caused by UPEC (Kaper et al., 2004). Although the affordable implementation of WGS in longitudinal cohort studies has uncovered adaptive patterns of various species to specific host environments (Didelot et al., 2016; Gatt and Margalit, 2021), the within-host pathoadaptation of multihabitat pathogens remains understudied. Here, we characterize the pathoadaptation of UPEC, one of the most common bacterial pathogens recovered from multiple body sites. Viewing UPEC within-host evolution in the context of their respective niche is key to understanding the origins of urovirulence in inherently intestinal *E. coli*, particularly in light of the lack of a unifying genomic signature of UPEC (Schreiber et al., 2017).

Our results support three distinct modes of UPEC persistence: exclusive persistence in the gastrointestinal tract (gut colonizer), persistence in both the gastrointestinal and urinary tracts (dual colonizer), and exclusive persistence in the urinary tract (urine colonizer). We find that these distinct patterns of persistence differentially shape UPEC within-host pathoadapation. Although development of antibiotic resistance is strongly selected for in all persisting UPEC lineages, as previously reported for other pathogens (Fajardo-Lubián et al., 2019; Khademi et al., 2019; Rossi et al., 2021), we find that distinct functions are under selection in gut and dual colonizers. Specifically, signatures of positive selection in distinct transport functions indicate that niche-specific adaptation directly impacts evolutionary trajectories of pathoadaptive traits (Tang and Saier, 2014). Further adaptation to multiple habitats diversifies allelic profiles of persisting UPEC lineages. Intriguingly, potential interhabitat transfer resulting in the influx of uroadaptive mutations back into gut populations may lower the fitness boundaries for urinary recolonization by intrinsically gut-adapted *E. coli*. Experimental evidence has shown that virulence factors critical for urocolonization are similarly beneficial in the intestinal reservoir (Chen et al., 2013; Russell et al., 2018; Spaulding et al., 2017), mitigating theoretical evolutionary trade-offs. These observations suggest that urovirulence may be a direct consequence of the generalist properties of the *E. coli* virulence repertoire (Brown et al., 2012), which is, as we show, fine-tuned by habitat-specific adaptations in the urinary tract.

Our observations support the hypothesis that persistent pathogen colonization requires within-lineage genotypic heterogeneity originating from both *in situ* adaptation as well as genomic plasticity (Hammond et al., 2020). The prevalence of habitat-restricted mutations and genomic plasticity between urine and stool isolates provides strong evidence that niche-specific adaptation dictates within-host evolution during UPEC persistence. We find that habitat-specific genes are associated with functions that increase *E. coli* fitness in the intestinal or urinary habitat, such as piliation, iron acquisition, nitrogen import, or anaerobic respiration (Elhenawy et al., 2021; Jones et al., 2011; Robinson et al., 2018; Spaulding et al., 2017). Persisting pathogen lineages require mechanisms that facilitate rapid rearrangements of large genomic regions to adapt to the distinct selective regimes of each habitat. Requirements for rapid genomic plasticity have been described for other pathogens, specifically during early stages of habitat colonization (Gabrielaite et al., 2020; Rau et al., 2012). Our results support the hypothesis that those genomic rearrangements are in part facilitated by MGEs (Sokurenko et al., 2006). Intriguingly, we observed that functions related to DNA repair were depleted in the MGE gene pool of urinary isolates from gut-adapted UPEC. This observation is consistent with the concept that stress-induced mutagenesis enables maladapted bacteria to adapt rapidly to their environment and may therefore be beneficial for gut-adapted UPEC lineages following urinary inoculation (Shee et al., 2011). Heterogeneous MGE carriage provides opportunistic pathogens with a unique mechanism to maintain fitness in multiple habitats. *In vitro* experiments have shown that complex environments result in discontinuous plasmid distribution in clonal populations, potentially resulting in fitness benefits under changing selection (Rodríguez-Beltrán et al., 2021; Slater et al., 2008, 2010). Our results support the hypothesis that MGE-mediated plasticity in bacterial populations is a key mechanism for habitat adaptation and may directly impact bacterial fitness upon habitat transition. Our data further suggest that a pool of gut-specific MGEs shared with other gut resident species may be lost in the urinary environment. Moreover, we find that gut-colonizing lineages causing rUTI during our follow-up period have significantly lower MGE richness compared with their non-rUTI counterparts, suggesting an inverse relationship between MGE richness and likelihood of rUTI in gut-adapted lineages of UPEC. Consistent with predictions from *in vitro* work (Harrison et al., 2018), the absence of a similar trend in dual colonizers suggests that multihabitat colonization stabilizes plasmid carriage under spatially heterogeneous selection, potentially via mechanisms like compensatory mutations (Hall et al., 2021; Harrison et al., 2015).

However, important questions remain to be investigated. This study could not address the topic of directionality and interhabitat transfer, the frequency of which may impact adaptative trajectories of persisting UPEC lineages. Moreover, given the apparent importance of genomic plasticity for UPEC fitness, localization of functions on either the chromosome or MGEs may determine the uropathogenic potential of intestinal *E. coli* lineages. The mosaic structure of plasmids poses the question that functions determine plasmid spread, evolution, and persistence in UPEC lineages. Although our study represents one of the largest genomic databases of UPEC to date, a number of patients were lost due to dropout, limiting the number of available isolates from follow-up episodes, specifically diagnostic isolates from outpatient settings. Similarly, our study lacked a representative number of lineages persisting exclusively in the urinary tract that are potentially uniquely adapted to the urinary environment. Large multiepisode sampling efforts from patients at risk for rUTI are required to support rarity of this persistence type and the novel genomic predictions of our study.

This study, harnessing an expansive, longitudinal patient cohort sampled at multiple habitats, provides a framework for future investigations, studying the role of both *in vivo* mutations and genomic plasticity in the within-host adaptation of bacterial pathogens across niches. Similar investigations in other species may reveal further mechanisms of colonization and aid targeted decolonization of persisting human pathogens.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.chom.2022.04.008.

ACKNOWLEDGMENTS

The authors thank Eric Keen, Drew J. Schwartz, Bejan Mahmud, Alaric D'Souza, Kimberley Sukhum, Manish Boolchandani, and Mary K. Hayden for providing scientific discussions and support. We thank the staff at the Edison Family Center for Genome Sciences and Systems Biology at Washington University School of Medicine: Bonnie Dee, Kathleen Matheny, and Keith Page for administrative support; Jessica Hoisington-Lopez and MariaLynn Crosby for managing the high-throughput sequencing core; and Eric Martin and Brian Koebbe for computational support. We thank Carrie Crook, Tony James, and Emily Reesey for providing study coordination support. This work was supported in part by awards to the authors from the U.S. Centers for Disease Control and Prevention Epicenter Prevention Program (grant U54CK000482; principal investigator, V.J.F.); to J.H.K. from the Longer Life Foundation (an RGA/Washington University partnership), the National Center for Advancing Translational Sciences (grants KL2TR002346 and UL1TR002345), and the National Institute of Allergy and Infectious Diseases (NIAID) (grant K23A1137321) of the National Institutes of Health (NIH); and to G.D. from NIAID (grant R01AI123394) and the Eunice Kennedy Shriver National Institute of Child Health and Human Development (grant R01HD092414) of NIH. R.T.'s research was funded by the Deutsche Forschungsgemeinschaft (DFG; German Research Foundation; grant 402733540). The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies.

AUTHOR CONTRIBUTIONS

Conceptualization, J.H.K., E.R.D., C.-A.D.B., G.D., R.T., and J.C.; resources, J.H.K., E.R.D., G.D., K.A.R., S.S., C.C., M.H.B., and E.L.S.; investigation, R.T., T.H., A.T., M.A.W., B.Wang, Z.H.I., S.R.S., A.W.B., K.R.F., B.X., B.Williams, P.C.-T., E.L., and J.H.K.; data curation, K.A.R. and R.T.; bioinformatics

and statistical analysis, R.T. and J.C.; writing – original draft, R.T. and J.C.; writing – review & editing, R.T., J.C., A.T., T.H., K.A.R., M.A.W., V.J.F., A.W.B., B.Williams, P.C.-T., E.L., C.-A.D.B., E.R.D., J.H.K., and G.D.; visualization, R.T., J.C., and A.T.; supervision, J.H.K., E.R.D., C.-A.D.B., and G.D.; project administration, K.A.R. and J.H.K.; funding acquisition, V.J.F., J.H.K., E.R.D., C.-A.D.B., and G.D.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: October 11, 2021
Revised: March 9, 2022
Accepted: April 14, 2022
Published: May 10, 2022

REFERENCES

Antipov, D., Hartwick, N., Shen, M., Raiko, M., Lapidus, A., and Pevzner, P.A. (2016). plasmidSPAdes: assembling plasmids from whole genome sequencing data. Bioinformatics *32*, 3380–3387.

Arndt, D., Grant, J.R., Marcu, A., Sajed, T., Pon, A., Liang, Y., and Wishart, D.S. (2016). PHASTER: a better, faster version of the PHAST phage search tool. Nucleic Acids Res. *44*, W16–W21.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., et al. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. *19*, 455–477.

Baym, M., Kryazhimskiy, S., Lieberman, T.D., Chung, H., Desai, M.M., and Kishony, R. (2015). Inexpensive multiplexed library preparation for megabase-sized genomes. PLoS One *10*, e0128036.

Beloin, C., Michaelis, K., Lindner, K., Landini, P., Hacker, J., Ghigo, J.M., and Dobrindt, U. (2006). The transcriptional antiterminator RfaH represses biofilm formation in *Escherichia coli*. J. Bacteriol. *188*, 1316–1331.

Biggel, M., Xavier, B.B., Johnson, J.R., Nielsen, K.L., Frimodt-Møller, N., Matheeussen, V., Goossens, H., Moons, P., and Van Puyvelde, S. (2020). Horizontally acquired papGII-containing pathogenicity islands underlie the emergence of invasive uropathogenic *Escherichia coli* lineages. Nat. Commun. *11*, 1–15.

Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114–2120.

Bricio-Moreno, L., Sheridan, V.H., Goodhead, I., Armstrong, S., Wong, J.K.L., Waters, E.M., Sarsby, J., Panagiotou, S., Dunn, J., Chakraborty, A., et al. (2018). Evolutionary trade-offs associated with loss of PmrB function in host-adapted pseudomonas aeruginosa. Nat. Commun. *9*, 2635.

Bronson, R.A., Gupta, C., Manson, A.L., Nguyen, J.A., Bahadirli-Talbott, A., Parrish, N.M., Earl, A.M., and Cohen, K.A. (2021). Global phylogenomic analyses of Mycobacterium abscessus provide context for non cystic fibrosis infections and the evolution of antibiotic resistance. Nat. Commun. *12*, 1–10.

Brown, S.P., Cornforth, D.M., and Mideo, N. (2012). Evolution of virulence in opportunistic pathogens: generalism, plasticity, and control. Trends Microbiol. *20*, 336–342.

Carattoli, A., Zankari, E., García-Fernández, A., Voldby Larsen, M., Lund, O., Villa, L., Møller Aarestrup, F., and Hasman, H. (2014). In silico detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing. Antimicrob. Agents Chemother. *58*, 3895–3903.

Chattopadhyay, S., Feldgarden, M., Weissman, S.J., Dykhuizen, D.E., Van Belle, G., and Sokurenko, E.V. (2007). Haplotype diversity in "source-sink" dynamics of *Escherichia coli* urovirulence. J. Mol. Evol. *64*, 204–214.

Chen, S.L., Wu, M., Henderson, J.P., Hooton, T.M., Hibbing, M.E., Hultgren, S.J., and Gordon, J.I. (2013). Genomic diversity and fitness of *E. coli* strains recovered from the intestinal and urinary tracts of women with recurrent urinary tract infection. Sci. Transl. Med. *5*, 184ra60.

Choi, U., and Lee, C.R. (2019). Distinct roles of outer membrane porins in antibiotic resistance and membrane integrity in *Escherichia coli*. Front. Microbiol. *10*, 953.

Coll, F., Harrison, E.M., Toleman, M.S., Reuter, S., Raven, K.E., Blane, B., Palmer, B., Kappeler, A.R.M., Brown, N.M., Török, M.E., et al. (2017). Longitudinal genomic surveillance of MRSA in the UK reveals transmission patterns in hospitals and the community. Sci. Transl. Med. *9*, 953.

Danecek, P., and McCarthy, S.A. (2017). BCFtools/csq: haplotype-aware variant consequences. Bioinformatics *33*, 2037–2039.

Darch, S.E., McNally, A., Harrison, F., Corander, J., Barr, H.L., Paszkiewicz, K., Holden, S., Fogarty, A., Crusz, S.A., and Diggle, S.P. (2015). Recombination is a key driver of genomic and phenotypic diversity in a Pseudomonas aeruginosa population during cystic fibrosis infection. Sci. Rep. *5*, 7649.

Didelot, X., Walker, A.S., Peto, T.E., Crook, D.W., and Wilson, D.J. (2016). Within-host evolution of bacterial pathogens. Nat. Rev. Microbiol. *14*, 150–162.

Dixon, P. (2003). VEGAN, a package of R functions for community ecology. J. Veg. Sci. *14*, 927–930.

Durrant, M.G., Li, M.M., Siranosian, B.A., Montgomery, S.B., and Bhatt, A.S. (2020). A bioinformatic analysis of integrative mobile genetic elements highlights their role in bacterial adaptation. Cell Host Microbe *27*, 140–153.e9.

Elhenawy, W., Hordienko, S., Gould, S., Oberc, A.M., Tsai, C.N., Hubbard, T.P., Waldor, M.K., and Coombes, B.K. (2021). High-throughput fitness screening and transcriptomics identify a role for a type IV secretion system in the pathogenesis of Crohn's disease-associated *Escherichia coli*. Nat. Commun. *12*, 2032.

Fajardo-Lubián, A., Ben Zakour, N.L., Agyekum, A., Qi, Q., and Iredell, J.R. (2019). Host adaptation and convergent evolution increases antibiotic resistance without loss of virulence in a major human pathogen. PLoS Pathog. *15*, e1007218.

Felsenstein, J. (1989). PHYLIP - phylogeny inference (Version 3.2). Cladistics *5*, 164–166.

Flores-Mireles, A.L., Walker, J.N., Caparon, M., and Hultgren, S.J. (2015). Urinary tract infections: epidemiology, mechanisms of infection and treatment options. Nat. Rev. Microbiol. *13*, 269–284.

Foxman, B. (2014). Urinary tract infection syndromes. Occurrence, recurrence, bacteriology, risk factors, and disease burden. Infect. Dis. Clin. North Am. *28*, 1–13.

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics *28*, 3150–3152.

Gabrielaite, M., Johansen, H.K., Molin, S., Nielsen, F.C., and Marvig, R.L. (2020). Gene loss and acquisition in lineages of pseudomonas aeruginosa evolving in cystic fibrosis patient airways. mBio *11*, 1–16.

Gatt, Y.E., and Margalit, H. (2021). Common adaptive strategies underlie Within-host evolution of bacterial pathogens. Mol. Biol. Evol. *38*, 1101–1121.

Götz, S., García-Gómez, J.M., Terol, J., Williams, T.D., Nagaraj, S.H., Nueda, M.J., Robles, M., Talón, M., Dopazo, J., and Conesa, A. (2008). High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Res. *36*, 3420–3435.

Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. Bioinformatics *29*, 1072–1075.

Hall, J.P.J., Wright, R.C.T., Harrison, E., Muddiman, K.J., Wood, A.J., Paterson, S., and Brockhurst, M.A. (2021). Plasmid fitness costs are caused by specific genetic conflicts enabling resolution by compensatory mutation. PLoS Biol. *19*, e3001225.

Hammond, J.A., Gordon, E.A., Socarras, K.M., Chang Mell, J.C., and Ehrlich, G.D. (2020). Beyond the pan-genome: current perspectives on the functional and practical outcomes of the distributed genome hypothesis. Biochem. Soc. Trans. *48*, 2437–2455.

Harrison, E., Guymer, D., Spiers, A.J., Paterson, S., and Brockhurst, M.A. (2015). Parallel compensatory evolution stabilizes plasmids across the parasitism-mutualism continuum. Curr. Biol. *25*, 2034–2039.

Harrison, E., Hall, J.P.J., and Brockhurst, M.A. (2018). Migration promotes plasmid stability under spatially heterogeneous positive selection. Proc. R. Soc. Lond. B *285*, 20180324.

Hibbing, M.E., Dodson, K.W., Kalas, V., Chen, S.L., and Hultgren, S.J. (2020). Adaptation of arginine synthesis among uropathogenic branches of the *Escherichia coli* phylogeny reveals adjustment to the urinary tract habitat. mBio *11*, 1–15.

Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S.K., Cook, H., Mende, D.R., Letunic, I., Rattei, T., Jensen, L.J., et al. (2019). EggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Res. *47*, D309–D314.

Jain, C., Rodriguez, R., L.M., Phillippy, A.M., Konstantinidis, K.T., and Aluru, S. (2018). High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat. Commun. *9*, 1.

Jia, B., Raphenya, A.R., Alcock, B., Waglechner, N., Guo, P., Tsang, K.K., Lago, B.A., Dave, B.M., Pereira, S., Sharma, A.N., et al. (2017). CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. Nucleic Acids Res. *45*, D566–D573.

Joensen, K.G., Tetzschner, A.M.M., Iguchi, A., Aarestrup, F.M., and Scheutz, F. (2015). Rapid and easy *in silico* serotyping of *Escherichia coli* isolates by use of whole-genome sequencing data. J. Clin. Microbiol. *53*, 2410–2426.

Jolley, K.A., Bray, J.E., and Maiden, M.C.J. (2018). Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. Wellcome Open Res. *3*, 124.

Jones, S.A., Gibson, T., Maltby, R.C., Chowdhury, F.Z., Stewart, V., Cohen, P.S., and Conway, T. (2011). Anaerobic respiration of *Escherichia coli* in the mouse intestine. Infect. Immun. *79*, 4218–4226.

Kaper, J.B., Nataro, J.P., and Mobley, H.L.T. (2004). Pathogenic *Escherichia coli*. Nat. Rev. Microbiol. *2*, 123–140.

Khademi, S.M.H., Sazinas, P., and Jelsbak, L. (2019). Within-host adaptation mediated by intergenic evolution in *Pseudomonas aeruginosa*. Genome Biol. Evol. *11*, 1385–1397.

Langmead, B., Wilks, C., Antonescu, V., and Charles, R. (2019). Scaling read aligners to hundreds of threads on general-purpose processors. Bioinformatics *35*, 421–432.

Larsen, M.V., Cosentino, S., Rasmussen, S., Friis, C., Hasman, H., Marvig, R.L., Jelsbak, L., Sicheritz-Pontén, T., Ussery, D.W., Aarestrup, F.M., et al. (2012). Multilocus sequence typing of total-genome-sequenced bacteria. J. Clin. Microbiol. *50*, 1355–1361.

Lees, J.A., Kremer, P.H.C., Manso, A.S., Croucher, N.J., Ferwerda, B., Serón, M.V., Oggioni, M.R., Parkhill, J., Brouwer, M.C., van der Ende, A., et al. (2017). Large scale genomic analysis shows no evidence for pathogen adaptation between the blood and cerebrospinal fluid niches during bacterial meningitis. Microb. Genomics *3*, e000103.

Letunic, I., and Bork, P. (2019). Interactive Tree of Life (iTOL) v4: recent updates and new developments. Nucleic Acids Res. *47*, W256–W259.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

Lieberman, T.D., Flett, K.B., Yelin, I., Martin, T.R., McAdam, A.J., Priebe, G.P., and Kishony, R. (2014). Genetic variation of a bacterial pathogen within individuals with cystic fibrosis provides a record of selective pressures. Nat. Genet. *46*, 82–87.

Lieberman, T.D., Michel, J.B., Aingaran, M., Potter-Bynoe, G., Roux, D., Davis, M.R., Skurnik, D., Leiby, N., LiPuma, J.J., Goldberg, J.B., et al. (2011). Parallel bacterial evolution within multiple patients identifies candidate pathogenicity genes. Nat. Genet. *43*, 1275–1280.

Lourenço, M., Ramiro, R.S., Güleresi, D., Barroso-Batista, J., Xavier, K.B., Gordo, I., and Sousa, A. (2016). A mutational hotspot and strong selection contribute to the order of mutations selected for during *Escherichia coli* adaptation to the gut. PLoS Genet. *12*, e1006420.

Madeira, F., Park, Y.M., Lee, J., Buso, N., Gur, T., Madhusoodanan, N., Basutkar, P., Tivey, A.R.N., Potter, S.C., Finn, R.D., et al. (2019). The EMBL-EBI search and sequence analysis tools APIs in 2019. Nucleic Acids Res. *47*, W636–W641.

Mann, R., Mediati, D.G., Duggin, I.G., Harry, E.J., and Bottomley, A.L. (2017). Metabolic adaptations of uropathogenic *E. coli* in the urinary tract. Front. Cell. Infect. Microbiol. *7*, 241.

Marvig, R.L., Sommer, L.M., Molin, S., and Johansen, H.K. (2015). Convergent evolution and adaptation of Pseudomonas aeruginosa within patients with cystic fibrosis. Nat. Genet. *47*, 57–64.

McGinnis, S., and Madden, T.L. (2004). BLAST: at the core of a powerful and diverse set of sequence analysis tools. Nucleic Acids Res. *32*, W20–W25.

Nielsen, K.L., Stegger, M., Godfrey, P.A., Feldgarden, M., Andersen, P.S., and Frimodt-Møller, N. (2016). Adaptation of *Escherichia coli* traversing from the faecal environment to the urinary tract. Int. J. Med. Microbiol. *306*, 595–603.

Osei Sekyere, J. (2018). Genomic insights into nitrofurantoin resistance mechanisms and epidemiology in clinical Enterobacteriaceae. Future Sci. OA *4*, FSO293.

Page, A.J., Cummins, C.A., Hunt, M., Wong, V.K., Reuter, S., Holden, M.T.G., Fookes, M., Falush, D., Keane, J.A., and Parkhill, J. (2015). Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics *31*, 3691–3693.

Page, A.J., Taylor, B., Delaney, A.J., Soares, J., Seemann, T., Keane, J.A., and Harris, S.R. (2016). SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. Microb. Genomics *2*, e000056.

Paradis, E., and Schliep, K. (2019). Ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics *35*, 526–528.

Parks, D.H., Imelfort, M., Skennerton, C.T., Hugenholtz, P., and Tyson, G.W. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. *25*, 1043–1055.

Poulsen, L.K., Licht, T.R., Rang, C., Krogfelt, K.A., and Molin, S. (1995). Physiological state of *Escherichia coli* BJ4 growing in the large intestines of streptomycin-treated mice. J. Bacteriol. *177*, 5840–5845.

Price, M.N., Dehal, P.S., and Arkin, A.P. (2009). Fasttree: computing large minimum evolution trees with profiles instead of a distance matrix. Mol. Biol. Evol. *26*, 1641–1650.

Rang, C.U., Licht, T.R., Midtvedt, T., Conway, P.L., Chao, L., Krogfelt, K.A., Cohen, P.S., and Molin, S. (1999). Estimation of growth rates of *Escherichia coli* BJ4 in streptomycin- treated and previously germfree mice by *in situ* rRNA hybridization. Clin. Diagn. Lab. Immunol. *6*, 434–436.

Rau, M.H., Marvig, R.L., Ehrlich, G.D., Molin, S., and Jelsbak, L. (2012). Deletion and acquisition of genomic content during early stage adaptation of Pseudomonas aeruginosa to a human host environment. Environ. Microbiol. *14*, 2200–2211.

Robinson, A.E., Heffernan, J.R., and Henderson, J.P. (2018). The iron hand of uropathogenic *Escherichia coli*: the role of transition metal control in virulence. Future Microbiol. *13*, 745–756.

Rodríguez-Beltrán, J., DelaFuente, J., León-Sampedro, R., MacLean, R.C., and San Millán, Á. (2021). Beyond horizontal gene transfer: the role of plasmids in bacterial evolution. Nat. Rev. Microbiol. *19*, 347–359.

Rossi, E., La Rosa, R., Bartell, J.A., Marvig, R.L., Haagensen, J.A.J., Sommer, L.M., Molin, S., and Johansen, H.K. (2021). Pseudomonas aeruginosa adaptation and evolution in patients with cystic fibrosis. Nat. Rev. Microbiol. *19*, 331–342.

Rozov, R., Brown Kav, A., Bogumil, D., Shterzer, N., Halperin, E., Mizrahi, I., and Shamir, R. (2017). Recycler: an algorithm for detecting plasmids from *de novo* assembly graphs. Bioinformatics *33*, 475–482.

Russell, C.W., Fleming, B.A., Jost, C.A., Tran, A., Stenquist, A.T., Wambaugh, M.A., Bronner, M.P., and Mulvey, M.A. (2018). Context-dependent requirements for FimH and other canonical virulence factors in gut colonization by extraintestinal pathogenic *Escherichia coli*. Infect. Immun. *86*, e00746-17.

Sarkar, S., Ulett, G.C., Totsika, M., Phan, M.D., and Schembri, M.A. (2014). Role of capsule and O antigen in the virulence of uropathogenic *Escherichia coli*. PLoS One *9*, e94786.

Schreiber, H.L., Spaulding, C.N., Dodson, K.W., Livny, J., and Hultgren, S.J. (2017). One size doesn't fit all: unraveling the diversity of factors and interactions that drive *E. coli* urovirulence. Ann. Transl. Med. *5*, 28.

Schwartz, D.J., Kalas, V., Pinkner, J.S., Chen, S.L., Spaulding, C.N., Dodson, K.W., and Hultgren, S.J. (2013). Positively selected FimH residues enhance virulence during urinary tract infection by altering FimH conformation. Proc. Natl. Acad. Sci. USA *110*, 15530–15537.

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. Bioinformatics *30*, 2068–2069.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. *13*, 2498–2504.

Shee, C., Gibson, J.L., Darrow, M.C., Gonzalez, C., and Rosenberg, S.M. (2011). Impact of a stress-inducible switch to mutagenic repair of DNA breaks on mutation in *Escherichia coli*. Proc. Natl. Acad. Sci. USA *108*, 13659–13664.

Sheppard, S.K., Guttman, D.S., and Fitzgerald, J.R. (2018). Population genomics of bacterial host adaptation. Nat. Rev. Genet. *19*, 549–565.

Siguier, P., Perochon, J., Lestrade, L., Mahillon, J., and Chandler, M. (2006). ISfinder: the reference centre for bacterial insertion sequences. Nucleic Acids Res. *34*, D32–D36.

Slater, F.R., Bruce, K.D., Ellis, R.J., Lilley, A.K., and Turner, S.L. (2008). Heterogeneous selection in a spatially structured environment affects fitness tradeoffs of plasmid carriage in pseudomonads. Appl. Environ. Microbiol. *74*, 3189–3197.

Slater, F.R., Bruce, K.D., Ellis, R.J., Lilley, A.K., and Turner, S.L. (2010). Determining the effects of a spatially heterogeneous selection pressure on bacterial population structure at the sub-millimetre scale. Microb. Ecol. *60*, 873–884.

Sokurenko, E.V., Feldgarden, M., Trintchina, E., Weissman, S.J., Avagyan, S., Chattopadhyay, S., Johnson, J.R., and Dykhuizen, D.E. (2004). Selection footprint in the FimH adhesin shows Pathoadaptive niche differentiation in *Escherichia coli*. Mol. Biol. Evol. *21*, 1373–1383.

Sokurenko, E.V., Gomulkiewicz, R., and Dykhuizen, D.E. (2006). Source-sink dynamics of virulence evolution. Nat. Rev. Microbiol. *4*, 548–555.

Spaulding, C.N., Klein, R.D., Ruer, S., Kau, A.L., Schreiber, H.L., Cusumano, Z.T., Dodson, K.W., Pinkner, J.S., Fremont, D.H., Janetka, J.W., et al. (2017). Selective depletion of uropathogenic *E. coli* from the gut by a FimH antagonist. Nature *546*, 528–532.

Su, C.C., Rutherford, D.J., and Yu, E.W. (2007). Characterization of the multidrug efflux regulator AcrR from *Escherichia coli*. Biochem. Biophys. Res. Commun. *361*, 85–90.

Supek, F., Bošnjak, M., Škunca, N., and Šmuc, T. (2011). REVIGO summarizes and visualizes long lists of gene ontology terms. PLoS One *6*, e21800.

Tang, F., and Saier, M.H. (2014). Transport proteins promoting *Escherichia coli* pathogenesis. Microb. Pathog. *71–72*, 41–55.

Thänert, R., Reske, K.A., Hink, T., Wallace, M.A., Wang, B., Schwartz, D.J., Seiler, S., Cass, C., Burnham, C.-A., Dubberke, E.R., et al. (2019). Comparative genomics of antibiotic-resistant uropathogens implicates three routes for recurrence of urinary tract infections. mBio *10*, e01977-19.

Weinstein, M.P. (2018). M100Ed29|Performance Standards for Antimicrobial Susceptibility Testing, Twenty-Ninth Edition (Clinical and Laboratory Standards Institute).

Weissman, S.J., Beskhlebnaya, V., Chesnokova, V., Chattopadhyay, S., Stamm, W.E., Hooton, T.M., and Sokurenko, E.V. (2007). Differential stability and trade-off effects of pathoadaptive mutations in the *Escherichia coli* FimH adhesin. Infect. Immun. *75*, 3548–3555.

Wielgoss, S., Schneider, D., Barrick, J.E., Tenaillon, O., Cruveiller, S., Chane-Woon-Ming, B., Médigue, C., and Lenski, R.E. (2011). Mutation rate inferred from synonymous substitutions in a long-term evolution experiment with *Escherichia coli*. G3 Genes Genomes Genet. *1*, 183–186.

Wilson, D.J.; CRyPTIC Consortium (2020). GenomegaMap: within-species genome-wide dN=dS estimation from over 10,000 genomes. Mol. Biol. Evol. *37*, 2450–2460.

Young, B.C., Wu, C.H., Gordon, N.C., Cole, K., Price, J.R., Liu, E., Sheppard, A.E., Perera, S., Charlesworth, J., Golubchik, T., et al. (2017). Severe infections emerge from commensal bacteria by adaptive evolution. Elife *6*, e30637.

Zankari, E., Hasman, H., Cosentino, S., Vestergaard, M., Rasmussen, S., Lund, O., Aarestrup, F.M., and Larsen, M.V. (2012). Identification of acquired antimicrobial resistance genes. J. Antimicrob. Chemother. *67*, 2640–2644.

Zhao, S., Lieberman, T.D., Poyet, M., Groussin, M., Xavier, R.J., Alm, E.J., Kauffman, K.M., and Gibbons, S.M. (2019). Adaptive evolution within gut microbiomes of healthy people article adaptive evolution within gut microbiomes of healthy people. Cell Host Microbe *25*, 656–667.e8.

Please cite this article in press as: Thänert et al., Persisting uropathogenic *Escherichia coli* lineages show signatures of niche-specific within-host adaptation mediated by mobile genetic elements, Cell Host & Microbe (2022), https://doi.org/10.1016/j.chom.2022.04.008

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological samples		
Stool samples from UTI patients	This paper	N/A
Urine samples from UTI patients	This paper	N/A
Critical commercial assays		
Hardy Diagnostic's ESBL agar	Hardy Diagnostics	Catalog #: G321
Hardy Diagnostic's MAC agar	Hardy Diagnostics	Catalog #: GA35
Mueller Hinton agar	Hardy Diagnostics	Catalog #: C6421
Kirby Bauer disk diffusion antibiotic disks	Hardy Diagnostics	N/A
Kirby Bauer disk diffusion antibiotic disks	Becton Dickinson	N/A
Blood agar plates	Hardy Diagnostics	Catalog #: GA50
QIAamp Bacteremia DNA kit	Qiagen	Catalog #: 12240-50
Nextera DNA Library Preparation Kit	Illumina	Catalog #: FC-131-1024
Deposited data		
Raw sequencing data for isolate whole genomes	This paper	NCBI SRA: PRJNA682246
Metadata of *E. coli* isolates sequenced for this study	This paper	See Data S1
Reference *E. coli* genomes	See Data S6	N/A
Software and algorithms		
Trimmomatic v.36	Bolger et al., 2014	https://github.com/usadellab/Trimmomatic
SPAdes v.3.11.0	Bankevich et al., 2012	https://github.com/ablab/spades
QUAST v5.0.2	Gurevich et al., 2013	http://quast.sourceforge.net
checkM v.1.0.13	Parks et al., 2015	https://github.com/Ecogenomics/CheckM
Prokka v.1.12	Seemann, 2014	https://github.com/tseemann/prokka
RGI-CARD v.5.1.0	Jia et al., 2017	https://github.com/arpcard/rgi
Resfinder v.4.0	Zankari et al., 2012	https://bitbucket.org/genomicepidemiology/resfinder/src/master/
mlst v2.11	Joensen et al., 2015	https://bitbucket.org/genomicepidemiology/mlst/src/master/
serotypefinder v2.0.1	Larsen et al., 2012	https://bitbucket.org/genomicepidemiology/serotypefinder/src/master/
Roary v3.8.0	Page et al., 2015	https://sanger-pathogens.github.io/Roary/
iTOL v.4	Letunic and Bork, 2019	https://itol.embl.de
FastTree v.2.1.10	Price et al., 2009	http://www.microbesonline.org/fasttree/
snp-sites v.2.4.0	Page et al., 2016	https://github.com/sanger-pathogens/snp-sites
fastANI v1.3	Jain et al., 2018	https://github.com/ParBLiSS/FastANI
Bowtie2 v.2.3.4	Langmead et al., 2019	http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
SAMtools v.1.9	Li et al., 2009	https://www.htslib.org/download/
BCFtools v.1.9	Danecek and McCarthy, 2017	https://www.htslib.org/download/
Ape package in R v.3.6.3	Paradis and Schliep, 2019	https://cran.r-project.org/web/packages/ape/index.html
PHYLIP v3.697	Felsenstein, 1989	https://evolution.genetics.washington.edu/phylip.html

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Snippy v4.3.8	N/A	https://github.com/tseemann/snippy
Genomegamap v1.0.1	Wilson, 2020	https://github.com/danny-wilson/genomegaMap
CD-HIT	Fu et al., 2012	http://weizhong-lab.ucsd.edu/cd-hit/
VEGAN package in R v.3.6.3	Dixon, 2003	https://cran.r-project.org/web/packages/vegan/index.html
blast2go	Götz et al., 2008	https://www.blast2go.com
REVIGO	Supek et al., 2011	http://revigo.irb.hr
Cytoscape	Shannon et al., 2003	https://cytoscape.org
PHASTER	Arndt et al., 2016	https://phaster.ca
plasmidSPAdes v.3.11.0	Antipov et al., 2016	https://github.com/ablab/spades
PlasmidFinder v.4.0	Carattoli et al., 2014	https://cge.cbs.dtu.dk/services/PlasmidFinder/
Recycler v.0.6.2	Rozov et al., 2017	https://github.com/Shamir-Lab/Recycler
ncbi-blast v.2.6.0+	McGinnis and Madden, 2004	https://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/
eggnog-mapper v.6.8	Huerta-Cepas et al., 2019	https://github.com/eggnogdb/eggnog-mapper
Clustal Omega	Madeira et al., 2019	https://www.ebi.ac.uk/Tools/msa/
MView	Madeira et al., 2019	https://www.ebi.ac.uk/Tools/msa/
ISfinder	Siguier et al., 2006	https://isfinder.biotoul.fr
Other		
MALDI-TOF MS	VITEK MS, bioMérieux	N/A
NextSeq 500 HighOutput platform	Illumina	N/A

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Gautam Dantas (dantas@wustl.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

Raw sequencing data has been deposited at the NCBI SRA database and are publicly available as of the date of publication. Accession numbers are listed in the key resources table. Relevant raw data and metadata can be found as supplemental information spreadsheets.

This paper does not report original code. We use well-established computational and statistical analysis software and packages. These are fully referenced in the STAR Methods section and key resources table.

Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Patient cohort

Subjects for this prospective, multi-center cohort study were recruited from patients with positive clinically indicated urine cultures at Barnes-Jewish Hospital/Washington University in St. Louis (WU), St. Louis, Missouri, Duke University Hospital (DK), Durham, North Carolina, the Hospital of the University of Pennsylvania (PN), Philadelphia, Pennsylvania and Rush University Medical Center (RH), Chicago, Illinois. This study was approved by the Washington University Human Research Protection Office as the single IRB; local IRB approval was obtained as necessary. Patients with a symptomatic UTI diagnosed and treated by a physician and a urine culture that yielded *E. coli* with one of the following resistances were included in the current analysis: (1) resistance to ciprofloxacin or levofloxacin, (2) resistance to any third generation cephalosporin, (3) resistance to ertapenem and susceptible to meropenem, imipenem, and/or doripenem, (4) resistance to >2 of the following antimicrobial classes: carbapenems, aminoglycosides, fluoroquinolones, fourth generation cephalosporins, piperacillin/tazobactam, or (5) identification of any of the following resistance mechanisms: ESBL, CRE, KPC, NDM-1, OXA-48, IMP, IMP-1, or VIM.

Patients were excluded if they were younger than 18 years, if more than one organism was detected by the clinical laboratory at or above the clinical significance threshold, had any chronic indwelling urinary device, or any medical or surgical condition leading to intestinal or urinary system disease or anatomic alteration. Written, informed consent was obtained from all patients. Patients age averaged 56.26 years (range: 18-94, median: 59). 93.5% of patients were female, and 6.50% of patients male. 58.54% of patients self-reported their race as White, and 37.40% as Black. 4.07% of patients reported their ethnicity as Hispanic. Pearson's chi-square tests indicated no significant association of age, gender, or race with UTI recurrence or UPEC colonization.

123 of 127 enrolled patients had at least one biological specimen yielding *E. coli* and were included in the current study. This total includes data from 12 patients enrolled at WU reported in a pilot study (Thänert et al., 2019). In total, 41 patients were enrolled at WU, 22 at DK, 12 at RH and 48 at PN.

METHOD DETAILS

Sample collection and processing

Enrolled subjects submitted stool and urine specimens to the study team at eleven sampling points over a 6-month follow-up period; enrollment (sampling point 01); the end of UTI antimicrobial treatment (02); days 3 (03), 7 (04), 14 (05), 30 (06), 60 (07), 90 (08), 120 (09), 150 (10), and 180 (11) post-treatment. If patients experienced rUTI during the 6-month follow-up period, they were invited to continue to participate with a new follow-up period. Visual schematic of the study design was created with BioRender.com. Samples were kept on ice immediately after production and during transport by courier. Upon arrival to the lab, samples were immediately cultured or prepared for long-term storage and frozen at -80 °C.

Stool and urine samples collected at sampling points 01, 02, 04, 06, and 11 were selectively cultured to assess asymptomatic uropathogen persistence. For stool culturing, ~1 g of stool sample was supplemented with an equal amount of PBS (w/v) and vortexed to homogenize the samples. Ten, 10-fold serial dilutions of the homogenate were prepared in PBS and 10μl of the first 10 dilutions were streaked on selective agar using a 10 μL calibrated loop. For urine culture, urines were directly plated onto selective agar using a 10 μL calibrated loop using a cross-streak pattern. After 20-30 hours of incubation, agar plates were examined for growth of the putative pathogen. Selective agars were selected to be specific to each patient's identified UPEC. MacConkey agar (MAC) supplemented with ciprofloxacin was used for ciprofloxacin-resistant *E. coli*, while ESBL *E. coli* was cultured on Hardy Diagnostic's ESBL agar and MAC agar supplemented with cefotaxime. A single, representative colony of each distinct colony morphology present on a given culture plate was selected for further processing and sequenced-based analysis. The identity of the cultured pathogens was confirmed using MALDI-TOF MS (VITEK MS, bioMérieux, Durham, NC, USA). Single colonies were diluted in TSB/glycerol and stored at -80°C for later sequencing-based and phenotypic analysis. If patients were unable to submit a specimen at a predetermined sampling point samples collected at the next closest available time point were selected for analysis. Additionally, pre-recurrence specimens of rUTI patients and time-matched samples from non-rUTI were further processed. Non-rUTI patients were matched to rUTI patients based on (1) colonization status (defined below) and (2) treatment antibiotic during the first episode.

Antimicrobial susceptibility testing

Antimicrobial susceptibility testing of pathogens was performed on Mueller Hinton agar (Hardy Diagnostics, Santa Maria, CA, USA) using Kirby Bauer disk diffusion with antibiotic disks purchased from Hardy Diagnostics (Santa Maria, CA, USA) and Becton Dickinson (Franklin Lakes, NJ, USA). Results were interpreted according to consensus-based medical laboratory standards as provided in the Clinical and Laboratory Standards Institute (CLSI) guidelines for antimicrobial susceptibility testing (Weinstein, 2018), which provide species-specific breakpoint definitions for determining susceptibility or resistance.

DNA extraction, short-read sequencing, and quality filtering

Isolates were streaked onto blood agar (Hardy Diagnostics, Santa Maria, CA, USA) and incubated at 35°C overnight. Genomic DNA was extracted using the QIAamp Bacteremia DNA kit (Qiagen, Germantown, MD, USA). Sequencing libraries from both isolate gDNA and fecal metagenomic DNA were prepared using the Nextera kit (Illumina, San Diego, CA, USA) (Baym et al., 2015). Libraries were pooled and sequenced (2 x150 bp) to a depth of ~2.5 million reads on the NextSeq 500 HighOutput platform (Illumina, San Diego, CA, USA). The resulting reads were trimmed of adapters using Trimmomatic v.36 (parameters: LEADING:10 TRAILING:10 SLIDINGWINDOW:4:15 MINLEN:60) (Bolger et al., 2014).

QUANTIFICATION AND STATISTICAL ANALYSIS

Isolate genome assembly and annotation

Draft genomes were assembled using SPAdes v.3.11.0 (parameters: -k 21,33,55,77 -careful) (Bankevich et al., 2012). The resulting scaffolds.fasta files were used for analysis. The quality of draft genomes was assessed by calculating assembly statistics using QUAST v5.0.2 and checkM v.1.0.13 (Gurevich et al., 2013; Parks et al., 2015). High-quality assemblies (<300 contigs, >90% of genome in contigs >1000bp, completeness >90%, contamination <5%) were annotated for open reading frames with Prokka v.1.12 (default parameters, contigs > 500 bp) (Seemann, 2014). Twenty-four publicly available *E. coli* genomes of known phylogroup were downloaded from NCBI to use as reference and annotated as described above (Data S6). These genomes were used to assign phylogroups to the isolates sequenced in this study based on core-genome relatedness to the set of references. ARGs were annotated *in silico* using RGI-CARD v.5.1.0 (95% identity, 100% coverage) and Resfinder v.4.0 (95% identity, 100% coverage) (Jia et al., 2017; Zankari et al., 2012).

Phylogenetic analysis and lineage definition

MLST were annotated *in silico* using mlst v2.11 (default parameter) and serotypes were assigned using serotypefinder v2.0.1 (parameters: -mp blast -l 0.8 -t 0.90) (Joensen et al., 2015; Larsen et al., 2012). Core-genome alignments were generated using Roary v3.8.0 (default parameters, -cd 100) (Page et al., 2015). For sequence type-specific phylogenetic analysis core-genomes were constructed using all isolates typed to ST 131 or 1193, respectively (Figure S2). To define lineages, all *E. coli* isolates from the same patient were used for core-genome construction. Patient-specific core-genome sizes are provided in Data S7. Newick trees of the core genome phylogenies were generated using FastTree v.2.1.10 (parameters: -gtr -nt) and visualized using iTOL v.4 (Letunic and Bork, 2019; Price et al., 2009).

To define *E. coli* lineages, patient-specific pairwise core-genome SNP distances were determined from the patient-specific Roary core-genome alignments via snp-sites v.2.4.0 (default parameter) (Page et al., 2016). Output files were converted into SNP distance matrices using custom R and python scripts. Based on the distribution of pairwise SNP distances (Figure S1; Data S7), *E. coli* lineages were herein defined to have <500 SNPs. Lineages were defined to be UPEC for the purpose of this study if they were isolated as the causative agent (DxU isolate) of a UTI. Pairwise ANI values between same-patient isolates were calculated using fastANI v1.3 (parameters: –fragLen 3,000, –minFraction 0.5) (Jain et al., 2018).

Determination of colonization patterns, lineage persistence, and rUTI causing UPEC

To understand colonization dynamics of UPEC and assess the impact of inter-habitat transfer on UPEC within-host adaptation, each UPEC lineage was categorized into one of four distinct persistence patterns: urinary tract colonization, intestinal colonization, dual, and uncolonized. Lineages were characterized as colonizing a given habitat (1) if the UPEC lineage was recovered from a habitat-specific specimen (stool/urine) at >1 collection point, or (2) if all habitat-specific specimens (stool/urine) from a UTI episode were positive for the UPEC lineage. DxU urine specimens were not considered for classification purposes. Lineages for which either type of specimen from their corresponding patient was unavailable were left unclassified. Lineages were further classified as rUTI if (1) the patient of isolation experienced a recurrence during the follow-up period and either (2) the same lineage was isolated as the DxU isolate of a rUTI or (3) no other lineage of *E. coli* was isolated at any point during follow-up. Lineages without follow-up DxU isolates or when multiple lineages of *E. coli* were isolated from a rUTI patient were left unclassified. Lineages from non-rUTI patients were classified as non-rUTI.

Characterization of within-lineage allelic diversity

To determine the allelic diversity between isolates from the same lineage, ‘‘pseudo-assemblies’’ were constructed for each UPEC lineage, as previously described (Thänert et al., 2019; Zhao et al., 2019). Equal proportions of reads from each isolate of a given lineage were pooled, assembled into a draft genome using SPAdes v.3.11.0 (parameters: -k 21,33,55,77 -careful), and annotated using Prokka v.1.12 (default parameters, contigs > 500 bp) (Bankevich et al., 2012; Seemann, 2014). These pseudo-assemblies were used as high-resolution reference genomes to characterize within-lineage allelic variation. Isolate reads were mapped to their respective pseudo-assemblies using Bowtie2 v.2.3.4 (parameters: -X 2000 –no-mixed –very-sensitive –n-ceil 0,0.01) (Langmead et al., 2019). SNPs and insertions/deletions were annotated using SAMtools v.1.9 and BCFtools v.1.9 (parameters: bcftools call -c -I ’DP>10 & QS>0.95’, bcftools view -i ’FQ<-85’) (Danecek and McCarthy, 2017; Li et al., 2009). SNPs were further filtered for major allele frequency >90% and gene presence in >60% of isolates from a given lineage, to exclude SNPs in potential MGEs. Mutated loci were mapped back to the reference GFF file (from Prokka) to identify corresponding coding sequences. Pairwise SNP distance matrices were used to construct unrooted lineage-specific phylogenetic trees, using the ape package in R v.3.6.3 (Paradis and Schliep, 2019). Time to last common ancestor (LCA) was estimated using median branch lengths of the resulting tree (determined via ape function ‘*edge.length*’) and dividing it by the estimated rate of *E. coli* evolution of 8.9×10^{-11} per base-pair per generation (Wielgoss et al., 2011), given an intestinal generation time of 80 minutes (Poulsen et al., 1995; Rang et al., 1999).

dMRCA estimation

To estimate dMRCA for each lineage, we generated parsimonious SNP trees using PHYLIP v3.697 (Felsenstein, 1989) to infer the ancestral sequence. VCF files resulting from within-lineage SNP characterization above were merged (bcftools merge – snps) including an isolate from the closest-related lineage according to ANI as an outgroup. The resulting VCF files were converted to ‘.phy’ format using the s_vcf2phylip.py script published by Ortiz et al on Github (https://github.com/edgardomortiz/vcf2phylip/blob/master/vcf2phylip.py). Files were used as input in the PHYLIP dnapars program (default parameters). Isolate dMRCA values were determined based on variable positions to the ancestral allele and used to calculate lineage averages. Lineage dMRCA values were compared between colonization types using Kruskal-Wallis with Dunn post hoc test. p-values were adjusted for multiple comparisons using the Benjamini-Hochberg method (FDR).

Permutation test for non-random distribution of mutations

To identify non-random parallel evolution in UPEC lineages separate permutation tests were implemented for the two main colonization types; gut colonizers (gut isolates only) and dual colonizers. Mutations were randomly distributed across the lineage-specific pseudo-reference assemblies (*i.e.*, if a lineage exhibited 10 SNPs total, 10 random SNPs were assigned in the genome). This process was repeated 1000 times for all lineages. The overall simulated distribution was used as the expected (neutral) distribution to test significance. The *P*-value was calculated as the top percentile of the neutral distribution at which the observed lineage count was present. To profile UPEC within-host adaptation, gut colonizers' pseudo-reference assemblies were generated using only gut isolate reads. To profile inter-habitat, within-lineage mutations, 71 urinary isolates from the 51 gut colonizing lineages were added and permutations were re-run.

Estimation of dN/dS

To determine signatures of positive selection at specific genes, isolate gene sequences were aligned using Snippy v4.3.8, using as a reference the corresponding pseudo-assembly.ffn file as annotated by Prokka v3.8.0. STOP codons were masked from the Snippy snps.consensus.fa output files using a custom script. dN/dS values for each gene's lineage-specific alignment were determined in Genomegamap v1.0.1 using the Maximum Likelihood estimation (Wilson, 2020). Overall dN/dS values for gene groups were estimated by generating a codon-based library of all possible mutations and calculating expected N/S ratios for each gene in the gene group. Overall dN/dS values were then calculated by summarizing the observed non-synonymous and synonymous mutations over all genes within the gene group. 95% confidence intervals were calculated by sampling from a binomial distribution as done previously (Zhao et al., 2019). Insertions/deletions as well as genes of plasmidic origin, due to their increased genetic variability (Rodríguez-Beltrán et al., 2021), were masked for group-wise dN/dS calculations.

Identification of within-lineage genomic plasticity

The accessory gene content of each UPEC lineage was identified based on a collapsed set of non-redundant genes. Therefore, clusters homologous genes were identified using CD-HIT (Fu et al., 2012), clustering translated gene sequences clustering at >90% amino acid identity. Within-lineage Jaccard dissimilarities (distances) of accessory gene content were calculated using the VEGAN package in R v.3.6.3 (Dixon, 2003). Average values for each lineage were used in comparisons. Dissimilarities of gene content were compared between colonization types, between and within habitat using ANOVA and Kruskal-Wallis with Dunn post hoc. *P*-values were adjusted for multiple comparisons using the Benjamini-Hochberg method (FDR).

GO overrepresentation analysis (GOOA)

To gain insights into the functions under selection during UPEC persistence, we annotated GO terms of genes with non-random mutational signatures (as per the permutation test above) or habitat-specific within-lineage abundance patterns using blast2go (Götz et al., 2008). We compared gene-set associated GO terms frequencies to their expected value as determined using a fully GO-annotated colonization-type specific background (*i.e.*, pangenome of each colonization type). To reduce redundancy in the GO term list associated with habitat-specific genes, we clustered overlapping GO terms using REVIGO prior to analysis allowing small similarity (<0.5) (Supek et al., 2011). Functional categories under selection during UPEC within-host persistence were identified using one-sided Fisher's exact test (hypergeometric distribution) in R v.3.6.3. *P*-values were adjusted for multiple comparisons using the Benjamini-Hochberg method (FDR). Fold-changes (enrichment scores) were calculated comparing observed vs expected values. For GO network analysis significant GOOA results were clustered semantically using REVIGO and visualized using Cytoscape (Shannon et al., 2003; Supek et al., 2011).

Comparison with published UPEC genomes

We downloaded raw reads for 703 UPEC genomes previously curated from multiple studies (Data S7) from NCBI (Biggel et al., 2020). We assembled genomes using SPAdes v.3.11.0 and assemblies using Prokka v.1.12 (default parameters). We extracted the amino acid sequences of OmpC and NsfA, found to be under positive selection and associated with the gain of phenotypic antibiotic resistance in this study, from all assemblies containing these genes. We queried the mutations (SNPs and INDELs) identified in this study against the set of reference sequences and extracted sequences from UPEC genomes containing the same mutations. We performed multiple sequence alignment between variable regions from our study and UPEC genomes using Clustal Omega and visualized alignments using MView (Madeira et al., 2019). OmpC and NfsA sequences from UTI89 were used as a reference.

MGE identification, annotation and characterization

We identified putative MGEs differentially abundant in isolates of the same lineage by aligning short reads to the pseudo-reference assembly. Candidate regions of at least 500bp length and <0.2X relative coverage in at least one isolate were considered for further analysis. Candidate MGEs in closed genomic proximity (<1 read pair - 300bp apart) were clustered to account for sporadic read mapping into conserved genomic regions interrupting continuous MGE identification. If candidate MGEs covered >90% of a contig in the pseudo-assembly, the whole contig was defined as a candidate MGE. Coverage for all putative within-lineage MGEs was determined for all isolates and a MGE presence/absence matrix was generated based on the average relative coverage for putative MGEs in each isolate's short read alignment. <0.2X relative coverage over the complete length of the MGE equaled absence and >0.8X relative coverage equaled presence in a given isolate. Intermediate values were defined to be unclear evidence of MGE presence/absence. Within-lineage similarity of isolate MGE profiles was assessed using Jaccard dissimilarities (distances) calculated using the VEGAN package in R v.3.6.3 (Dixon, 2003). Comparison of MGE profiles was performed using ANOVA with Tukey post hoc test and Welch's t-test. *P*-values were adjusted for multiple comparisons using the Benjamini-Hochberg method (FDR).

MGEs were annotated similarly to a previously published protocol for *de novo* MGE identification (Durrant et al., 2020). The pool of within-lineage MGEs was queried for prophages using PHASTER (Arndt et al., 2016). MGE contigs of plasmidic origin were identified combining replicon typing using 'Plasmid MLST' with mapping within-lineage MGE contigs to the complete pool of plasmidic contigs identified *de novo* in the draft assemblies of all isolate as previously described (Jolley et al., 2018; Thänert et al., 2019). This ''lineage-plasmidome'' was identified using plasmidSPAdes v.3.11.0 (parameters: –plasmid -k 21,33,55,77 –careful), Recycler v.0.6.2 (parameters: -k 77 -i True), and PlasmidFinder v.4.0 (parameters: -p enterobacteriaceae -k 95.00) (Antipov et al., 2016; Carattoli et al., 2014; Rozov et al., 2017). A non-redundant list of putative plasmidic contigs was validated against the NCBI plasmid database using ncbi-blast v.2.6.0+ (McGinnis and Madden, 2004). Contigs with >90% identity and >90% coverage of plasmid in the database were retained. This total ''lineage-plasmidome'' was annotated using Prokka v.1.12 (default parameters), the eggnog-mapper v.6.8 (parameters: -m diamond –query-cover 0.9), RGI-CARD v.5.1.0 (95% identity, 100% coverage), and Resfinder v.4.0 (95% identity, 100% coverage) (Huerta-Cepas et al., 2019; Jia et al., 2017; Seemann, 2014; Zankari et al., 2012). MGEs were determined to be plasmidic if they (1) had an exact replicon match in the Plasmid MLST database or (2) if they aligned to a contig of *de novo* identified plasmidic origin at >80% coverage and 99% identity using ncbi-blast v.2.6.0+ (McGinnis and Madden, 2004). Insertion sequences (IS) and transposases were identified in MGEs by blasting against the ISfinder database (Siguier et al., 2006). As the repetitive nature of IS frequently causes short-read assemblies to break, incomplete IS are often found at the edge of contigs. To account for this, IS were determined to be present if either (1) a partial IS match was identified at the edge of contig with >95% identity or (2) an IS was identified at >90% identity and >80% coverage. IS elements were defined as elements that only contained an IS/ Transposase and no other genes. Lastly, recombinases were identified in the Prokka annotations of the MGE pool.

Consistent with previous methods (Durrant et al., 2020), the final annotation for each MGE was assigned hierarchically from specific to general as follows; (1) Intact phages, (2) Plasmid, (3) IS element, (4) CDS+Transposase, (5) Recombinase, (6) Questionable/ Incomplete phage, (7) Contains CDS, and (8) No CDS. Habitat-specific genes were identified in the MGE pool using ncbi-blast v.2.6.0+ and determined to be present if (1) coverage >90% at 99% identity or (2) coverage >10% at 100% identify and the gene was determined to be located at the edge of a contig (McGinnis and Madden, 2004).

To reduce the likelihood of false positives, GOOA of mobilized functions between rUTI and non-rUTI lineages (Figure 6B) was performed after filtering out GO-terms present in less than 5% of all analyzed lineages. GO term overrepresentation in the mobilized gene pool of either rUTI or non-rUTI lineages was assessed using Fisher's exact test. *P*-values were adjusted for multiple comparisons using the Benjamini-Hochberg method (FDR). Pseudo enrichment scores were calculated comparing observed GO term abundances between compared groups adding the minimal value in the array as a pseudocount.

We further assessed MGE host ranges by aligning putative MGE contigs against the NCBI nucleotide database using ncbi-blast v.2.6.0+ (McGinnis and Madden, 2004), filtering for hits with >95% identity and 95% query coverage. Uncultured bacteria, eukaryotes, synthetic constructs/vectors, and mixed communities were filtered from the resulting hits. Taxa IDs were converted to species-level annotations and the number of species-level blast hits was summarized per MGE category. Statistical comparisons were performed using ANOVA and species under-represented in the urinary MGE pool were determined using one-sided Fisher's exact test. The 25 species most abundant in the blast hitlist were considered for statistical analysis. *P*-values were corrected for multiple-hypothesis testing using the Benjamini-Hochberg method (FDR).

General statistical approaches

Statistical comparisons were performed using ANOVA with Tukey post hoc, Kruskal-Wallis with Dunn post hoc, Welch's t-test, and Fisher's exact test as outlined above. Parametric or nonparametric tests were selected for a given comparison based on whether the underlying data approximated a normal distribution (Shapiro-Wilk test). When multiple-hypothesis were investigated, *P*-values were corrected for multiple-hypothesis testing using the Benjamini-Hochberg method (FDR). *P*-values <0.05 were considered 'significant'. Statistical details, including the statistical test used for each comparison, the number of observations (*n*), definition of center, dispersion and precision can be found in the results section, the figure legends, and figures.