Downloaded from genome.cshlp.org on June 16, 2026 . Published by Cold Spring Harbor Laboratory Press

Large-scale genomic sequencing of extraintestinal pathogenic *Escherichia coli* strains

Stephen J. Salipante, David J. Roach, Jacob O. Kitzman, et al.

Genome Res. published online November 4, 2014
Access the most recent version at doi:10.1101/gr.180190.114

P<P Published online November 4, 2014 in advance of the print journal.

Accepted Manuscript Peer-reviewed and accepted for publication but not copyedited or typeset; accepted manuscript is likely to differ from the final, published version.

Creative Commons License This article is distributed exclusively by Cold Spring Harbor Laboratory Press for the first six months after the full-issue publication date (see http://genome.cshlp.org/site/misc/terms.xhtml). After six months, it is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at http://creativecommons.org/licenses/by-nc/4.0/.

Email Alerting Service Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or **click here.**

Advance online articles have been peer reviewed and accepted for publication but have not yet appeared in the paper journal (edited, typeset versions may be posted when available prior to final publication). Advance online articles are citable and establish publication priority; they are indexed by PubMed from initial publication. Citations to Advance online articles must include the digital object identifier (DOIs) and date of initial publication.

To subscribe to *Genome Research* go to:
https://genome.cshlp.org/subscriptions

Published by Cold Spring Harbor Laboratory Press

Downloaded from genome.cshlp.org on June 16, 2026 . Published by Cold Spring Harbor Laboratory Press

Large-scale genomic sequencing of extraintestinal pathogenic *Escherichia coli* strains

Stephen J. Salipante[a,1,2], David J. Roach[b,1], Jacob O. Kitzman[b], Matthew W. Snyder[b], Bethany Stackhouse[b], Susan M. Butler-Wu[a], Choilee Lee[b], Brad T. Cookson[a,c], Jay Shendure[b,2]

Author affiliations: Departments of Laboratory Medicine[a], Genome Sciences[b], and Microbiology[c], University of Washington, Seattle, WA, 98195

[1]S.J.S and D.J.R. contributed equally to this work

[2]Corresponding authors:

Stephen Salipante, University of Washington, Box 357110, 1959 NE Pacific Street, Seattle, WA 98195-7110. Phone: (206) 598-6131 e-mail: stevesal@uw.edu

Jay Shendure, University of Washington, Box 355065, 3720 15th Ave NE, Seattle WA 98195-5065. Phone: (206) 685-3720 e-mail: shendure@uw.edu

Running Title: Large-scale clinical *E. coli* genomic sequencing

Keywords: Escherichia coli, genome wide association study, antibiotic resistance, virulence factor, molecular epidemiology, phylogeny, whole genome sequencing

Downloaded from genome.cshlp.org on June 16, 2026 . Published by Cold Spring Harbor Laboratory Press

Abstract:

Large-scale bacterial genome sequencing efforts to date have provided limited information on the most prevalent category of disease: sporadically acquired infections caused by common pathogenic bacteria. Here, we performed whole genome sequencing and *de novo* assembly of 312 blood- or urine-derived isolates of extraintestinal pathogenic (ExPEC) *Escherichia coli*, a common agent of sepsis and community-acquired urinary tract infections, obtained during the course of routine clinical care at a single institution. We find that ExPEC *E. coli* are highly genomically heterogeneous, consistent with pan-genome analyses encompassing the larger species. Investigation of differential virulence factor content and antibiotic resistance phenotypes reveals markedly different profiles among lineages and among strains infecting different bodily sites. We use high-resolution molecular epidemiology to explore the dynamics of infections at the level of individual patients, including identification of possible person-to-person transmission. Notably, a limited number of discrete lineages caused the majority of bloodstream infections, including one sub-clone (ST131-H30) responsible for 28% of bacteremic *E. coli* infections over a three-year period. We additionally use a microbial genome-wide-association study (GWAS) approach to identify individual genes responsible for antibiotic resistance, successfully recovering known genes but notably not identifying any novel factors. We anticipate that in the near future, whole genome sequencing of microorganisms associated with clinical disease will become routine. Our study reveals what kinds of information can be obtained from sequencing clinical isolates on a large scale, even well-characterized organisms such as *E. coli*, and provides insight into how this information might be utilized in a healthcare setting.

Downloaded from genome.cshlp.org on June 16, 2026 . Published by Cold Spring Harbor Laboratory Press

Introduction:

With the advent of high-throughput DNA sequencing technologies, it is becoming increasingly tractable to generate whole genome sequence data from large numbers of clinically-relevant bacterial isolates. However, most comparative genome sequencing efforts to date have focused on the biology and molecular epidemiology of organisms involved in disease outbreaks (Chin et al. 2011; Lieberman et al. 2011; Koser et al. 2012; Snitkin et al. 2012; Sanjar et al. 2014). Although illuminating, these studies have shed little light on the agents of bacterial disease that infect an overwhelming majority of patients: commonplace pathogens causing sporadically-acquired infections. Outbreaks represent the transmission of a single bacterial clone over a short period of time (Kennedy et al. 2010), providing a necessarily biased sampling that does not encompass the general properties of disease-causing organisms within a larger species. Relatedly, genomic studies of most bacteria are consistent with the distributed genome hypothesis, which proposes that the genetic content of a species is much larger than that of any single strain (Tettelin et al. 2005), necessitating sequencing of large numbers of unrelated clones in order to accurately catalog genetic variation (Rasko et al. 2008).

Escherichia coli is among the commonest clinical pathogens and is capable of causing a spectrum of disease both within the intestinal tract (intestinal pathogenic strains) and outside of it (extraintestinal pathogenic *E. coli*, or ExPEC). The most potentially destructive of these illnesses is bacterial invasion of the bloodstream: *E. coli* is the most common Gram-negative agent of sepsis, causing ~30% of all bacteremias and representing the tenth most common cause of death in industrialized nations (Martin et al. 2003; Jaureguy et al. 2008). Far more prevalent are *E. coli* urinary tract infections, which

Downloaded from genome.cshlp.org on June 16, 2026 . Published by Cold Spring Harbor Laboratory Press

encompass ~95% of all community-acquired cases (Lau et al. 2008; Manges et al. 2008). *E. coli* infections of either type incur significant morbidity and healthcare costs (Sannes et al. 2004; Lau et al. 2008; Ron 2010; Telli et al. 2010); regardless, only a handful of strains causing these diseases have been sequenced, and knowledge of ExPEC *E. coli* remains incomplete.

Here we performed large-scale whole genome sequencing and analysis of clinical isolates of extraintestinal pathogenic *E. coli*, obtained from routine diagnostic culture of peripheral blood or urine from patients within a single hospital system. These data enable a robust pan-genome analysis of ExPEC *E. coli*, high-resolution molecular epidemiological analysis, and genome-wide association studies for identifying antibiotic resistance genes.

Results:

Isolates: A total of 380 isolates were collected by the clinical Microbiology Laboratory of the University of Washington Medical Center. 288 uropathogenic *E. coli* (UPEC) were isolated from 277 patients (237 female, 36 male, 4 data not available) over a period of 5 months (**Supplemental Dataset 1**). Patients had an average age of 47 years (range newborn to 94 years), and two died during the health-care encounter that the isolate was obtained. Because of the lower incidence of *E. coli* bacteremia, 92 strains were isolated from blood culture of 47 patients (25 female, 22 male) over 3 years, representing all blood-borne *E. coli* isolates from our hospital over that period. Patients averaged 56 years of age (range newborn to 85), and 10 patients died during the health-care encounter that the isolate was obtained.

Pan-genome and core genome analysis: We initially explored the pan-genome composition of our strain collection. The “pan-genome” of a species refers to the full range

Downloaded from genome.cshlp.org on June 16, 2026 . Published by Cold Spring Harbor Laboratory Press

of non-orthologous genes that can be present in an organism, whereas the "core genome" comprises genes present in all representatives (Tettelin et al. 2005). *E. coli* is believed to have an "open" pan-genome marked by ongoing gene acquisition (Rasko et al. 2008), however, existing approximations are based on more limited numbers of previously sequenced strains (Rasko et al. 2008; Touchon et al. 2009; Kaas et al. 2012).

The pan-genome for 283 ExPEC strains passing quality control requirements for this analysis was estimated at 16,236 genes, or 14,877 genes after removing prophage and insertion sequence elements (**Supplemental Figure S1**). Substantial numbers of additional non-orthologous genes were identified with each strain sequenced, confirming an open species pan-genome even when analysis is confined to ExPEC strains. We predict a core genome size of 3,079 genes, or 3,039 genes after removal of prophages and insertion sequences (**Supplemental Figure S1**). These values compare favorably with previous pan-genome and core genome estimations (Kaas et al. 2012), although are somewhat lower and somewhat higher, respectively, than expected from a collection of this size, likely reflecting population structure among ExPEC *E. coli*. Comparatively, the core genome was most highly enriched for factors involved in basic cellular functions including DNA replication, cell wall synthesis, transcription, translation, and assorted amino acid metabolic and biosynthetic processes, while the pan-genome contained more genes related to metal-ion binding and virulence (predominantly flagellar proteins, capsular pathways, and secretion systems). All gene models from ExPEC *E. coli* isolates that were present in all, or nearly all, strains were also represented at a high frequency in commensals, suggesting that no specific gene is essential for ExPEC pathogenesis.

Downloaded from genome.cshlp.org on June 16, 2026 . Published by Cold Spring Harbor Laboratory Press

Phylogenomic analysis: We next investigated genomic and epidemiological relationships among isolates. After quality filtering, 312 *E. coli* isolates (221 UPEC and 91 bacteremia isolates) remained in this analysis. Phylogenomic reconstruction (Delsuc et al. 2005; Kumar et al. 2012) of genomic data was robust (**Fig. 1**), with likelihood values approaching 1 for most nodes (**Supplemental Figure S2**). Classification of strains to phylogenetic subgroups (Escobar-Paramo et al. 2006) and multilocus sequence types (MLST) (Tartof et al. 2005), current mainstays for *E. coli* classification, was performed through *in silico* analysis (**Fig. 1**, **Supplemental Dataset 2**).

All four of the common ExPEC phylogroups (Carlos et al. 2010) (A, B1, B2, and D) were represented, and comprised 8.3%, 8.7%, 65.7%, and 16.3% of the population, respectively, similar to reported distributions in other hospital systems (Clermont et al. 2000). Our analysis offers a more precise description of relationships among *E. coli* phylogroups (**Fig. 1A and Supplemental Figure S3)**, previously investigated through limited genetic information or using more qualitative methods (Lecointre et al. 1998; Johnson et al. 2006). Broadly, our data support the classification of phylogroups A and B1 as sister clades (Lecointre et al. 1998). Phylogroup B2, thought to be ancestral (Lecointre et al. 1998), forms a distinct clade that is divergent from the others. All phylogroups exhibited considerable genetic heterogeneity, and especially group D (**Supplemental Table S1**). Urine- and blood-derived isolates were extensively interleaved within the phylogeny, without significant enrichment for either infection type within any phylogroup (all comparisons $p \geq 0.05$, Fisher's exact test), and indicating that no clear phylogenomic division exists between ExPEC strains infecting these different sites.

Downloaded from genome.cshlp.org on June 16, 2026 . Published by Cold Spring Harbor Laboratory Press

Isolates from the same phylogroup, as determined by standard, combinatorial examination of a three-marker system (Escobar-Paramo et al. 2006), were closely related within the genomic phylogeny, with a few notable exceptions (**Fig. 1A**). Most strikingly, 4 strains classified as phylogroup D were most genomically similar to representatives of phylogroup B2 and formed a distinct clade, suggesting a common ancestry. As *E. coli* phylogroups reflect functional and evolutionary differences (Carlos et al. 2010), these outliers likely expose recent acquisition or loss of genetic material relevant to the phylogrouping marker system.

We looked for a possible correlation between the geographical distance separating *E. coli* isolates (as calculated between the centroids of patient home zip codes, average 441 km, range 0 – 7951 km) and the number of genomic differences among them. This analysis did not demonstrate a meaningful general correlation between genomic and geographical distance (not shown), suggesting that *E. coli* lineages are distributed relatively uniformly among members of our patient population.

We observed substantial clonal architecture. 285 isolates were distributed among 71 known MLST types (**Supplemental Dataset 2**), while 13 isolates demonstrated novel sequence types bearing undescribed MLST alleles or previously unreported allelic combinations (**Supplemental Table S2**). Six established sequence types accounted for just over half (51%) of the population: ST131, ST95, ST127, ST73, ST69, and ST393 (Adams-Sapper et al. 2013; Toval et al. 2014). Of these, ST131 (Price et al. 2013) and ST95 (Gibreel et al. 2012; Adams-Sapper et al. 2013) were most prevalent, comprising 16.1% (50/312) and 10.8% (34/312) of the overall population, respectively.

Downloaded from genome.cshlp.org on June 16, 2026 . Published by Cold Spring Harbor Laboratory Press

We also observed substantial population structure within individual MLST groups (**Fig. 1A**). Of note, we identified 2 different clades of ST131 isolates (**Fig. 2**): one small clade restricted to urinary infections and marked by fluoroquinolone sensitivity (7/7 isolates), the other isolated from both urine and blood and marked by an increased frequency of fluoroquinolone resistance (35/47 isolates). Focused investigation revealed that the more prevalent, fluoroquinolone resistant group corresponded to subclone H30, a recently emerged and highly pathogenic substrain (Colpan et al. 2013). A fraction of isolates from each H30 clade displayed extended-spectrum β-lactamase (ESBL) activity (**Fig. 2**): the most ancestral strains of the H30 subgroup exhibited concordant ESBL activity and fluoroquinolone resistance, whereas other scattered H30 isolates lacked one or more resistance phenotypes. This finding is consistent with the ST131-H30 multidrug resistance phenotype arising primarily through expansion of a single clone (Price et al. 2013), but also indicates that the phenotype has been lost by some descendants with measurable frequency (10 of the 47 isolates lacking ESBL activity, 1 isolate lacking fluoroquinolone resistance, and 10 having lost both).

Patient-level molecular epidemiology: We next examined the dynamics of *E. coli* infection within and among patients. 100 isolates resulted from longitudinal sampling of the same patients, with independent collections separated by several hours or up to two years. 18 urine-derived isolates were obtained from 9 patients (2 per patient) and 82 were isolated from the blood of 35 patients (90.1% of all blood-derived isolates).

We sought out subsets of strains that were virtually identical in terms of their genome sequences, which would suggest identity by descent. To first assess the magnitude of sequence artifacts introduced through sequencing, alignment, and variant calling,

Downloaded from genome.cshlp.org on June 16, 2026 . Published by Cold Spring Harbor Laboratory Press

technical replicates of four isolates were taken through library construction and sequencing in tandem. Replicates proved highly concordant, averaging 0.25±0.433 (average±standard deviation) pairwise differences (compared to 51,801 ± 29,207 differences in an all-by-all isolate comparison, **Supplemental Dataset 3**). Based on this level of uncertainty, we considered isolates genomically identical if they evidenced zero or 1 differentiating variant.

7 of 9 pairs of UPEC isolates were collected within three days of one another, and 6 of these pairs were genomically distinct (range of 502 to 69,681 pairwise differences). These findings are supported by different antibiotic resistance and hemolysis phenotypes in 4 cases, and indicate a high rate of polyclonal infections. Two pairs of UPEC isolates were obtained more than a month apart: one pair was genomically identical, supporting urinary tract colonization, while the other was not (66,059 pairwise differences), suggesting independent infections. We did not find instances of genomically identical UPEC isolates obtained from different patients.

In longitudinal samples from bacteremic patients, pairs of isolates obtained from an individual within 21 days universally comprised genomically identical strains, most likely evidencing cases of ongoing infection. The mean time between all paired samplings recovering the same clone was 12.7±16.3 (mean ± standard deviation) days (**Supplemental Figure S4**). Intriguingly, genomically identical strains could also be recovered over substantially longer periods of time: five independent samplings from patient 29 were performed over a five month period and yielded genomically identical multidrug resistant ST131-H30 isolates (**Fig. 1B**). Persistent *E. coli* bacteremia is rare, but may reflect repeated translocation of a pathogenic clone from the gut (Samet et al. 2013) or other colonized organs (Alsterlund et al. 2012; Gupta et al. 2013). In contrast, we found that pairs of

Downloaded from genome.cshlp.org on June 16, 2026 . Published by Cold Spring Harbor Laboratory Press

isolates obtained over longer periods of time (mean 251 days, standard deviation 261 days, **Supplemental Figure S4**) tended to represent genomically distinct isolates (**Fig. 1B**).

Of the 4,190 reported *E. coli* MLST groups, only 23 were recovered from blood infections. Moreover, only eight sequence types were cultured from multiple patients (**Fig. 1B**): combined, those lineages accounted for 76.6% (36/47) of blood infections in all patients over the study period. Included was the ST131-H30 lineage (21 independent isolates), collected from 28% (12/47) of bacteremic individuals. All 20 ST131-H30 blood isolates which underwent antibiotic resistance profiling, but only 54% (13/24) of those from urine, were fluoroquinolone resistant, supporting a distinct phenotypic profile associated with bacteremia (**Figure 2**). These results indicate that only a limited subset of closely related *E. coli* isolates were responsible for the majority of bacteremia cases.

The high prevalence of near-identical isolates in cases of bacteremia could reflect infection from high prevalence endemic strains (Manges et al. 2004; Manges et al. 2008), nosocomial transmission of *E. coli*, or some combination, warranting more detailed investigation of the blood-derived isolates (**Fig. 1B**). There was clinical evidence consistent with nosocomial transmission in one patient (patient 1, **Fig. 1B**). This immunocompromised individual became septic in the setting of graft versus host disease. Two isolates from phylogroup A, which we found to be genomically identical, were independently cultured from his blood over a period of 11 days, and the patient was transferred to an intensive care unit. Resolution of the infection was achieved and confirmed by 26 serially-negative blood cultures. Forty days after transfer, the patient again developed sepsis, and cultures recovered a multi-drug resistant *E. coli* strain we identified as ST131-H30 (phylogroup B2). Attempts at treatment were unsuccessful and the patient

Downloaded from genome.cshlp.org on June 16, 2026 . Published by Cold Spring Harbor Laboratory Press

227 died. At the time of this second infection and in the same hospital ward, patient 29 was
228 receiving treatment for *E. coli* bacteremia with an identical antibiotic resistance profile, and
229 which we similarly found to be an ST131-H30 strain. This evidence would be consistent
230 with nosocomial transmission from patient 29 to patient 1. Regardless, the ST131-H30
231 isolate from patient 1 differed from that of patient 29 by 355 genomic variants and
232 harbored 11 additional virulence factors. These genomic data unambiguously demonstrate
233 independent sources of infection, rather than nosocomial transmission.

234 Nevertheless, a strain genomically identical to the ST131-H30 isolate from patient
235 1, and with the same antibiotic resistance profile, was recovered from patient 6 over one
236 year later in a different hospital ward (**Fig. 1B**), convincingly supported by robust sequence
237 coverage of both isolates (>45X read depth). Although the epidemiological link in this
238 instance, if any, is unknown, sharing of *E. coli* strains among close contacts is documented
239 (Foxman et al. 1997; Johnson et al. 2008).

240 Although this strategy identifies true bacterial clones, which are by definition
241 genomically identical, some degree of clonal diversification may occur within patients
242 (Walker et al. 2013) or during the course of an outbreak (Lindsay 2014). We therefore
243 expanded our search for potential transmission events to include pairs of strains harboring
244 ≤ 15 genomic differences (**Supplemental Dataset 3, Supplemental Figure S5**), an amount
245 of divergence expected to accumulate over ~6,750-15,000 bacterial generations (Barrick et
246 al. 2009; Lee et al. 2012), and evaluated isolates with respect to temporal association. Five
247 pairs of strains qualified under this definition. Two pairs of those isolates were
248 distinguishable by distinct antibiotic resistance phenotypes, and thus were unlikely to
249 represent direct transmissions. However, the remaining paired comparisons comprised a

Downloaded from genome.cshlp.org on June 16, 2026 . Published by Cold Spring Harbor Laboratory Press

trio of UPEC strains (upec-61, -106, and -249) collected within 3 months of each other from a geographically constrained area (**Supplemental Table S3**), and exhibiting the same pan-antibiotic sensitive phenotype. Given robust sequence coverage of all three isolates (>49X read depth), we speculate that the three strains are epidemiologically linked, although contact among these patients cannot be known to us.

Distribution of virulence factors and antibiotic resistance phenotypes: Virulence factors (VFs) play an important role in conferring selective advantages to, and defining pathogenicity profiles of, *E. coli* strains (Nowrouzian et al. 2006; Ramos et al. 2010). Accordingly, disease-associated phylogroups of *E. coli* have a higher prevalence of VFs and antibiotic resistance than commensals (Picard et al. 1999; Price et al. 2013), and some subgroups are enriched for distinct subsets of VFs (Nowrouzian et al. 2006). To more fully explore the distribution of factors among ExPEC phylogroups, blood- and urine-derived isolates, and MLST groups which are under clinical selective pressure as human pathogens, we cataloged the prevalence of known VFs and antibiotic resistance phenotypes within groups of *E. coli* defined at these population levels (**Fig. 3, Supplemental Datasets 4 and 5**) and explored statistically significant differences in their enrichment or depletion. We also considered differences among these groups after accounting for population structure (Price et al. 2006) (**Fig. 3, Supplemental Datasets 4 and 5**) in order to identify factors which may have been acquired or lost independently within lineages multiple times, rather than inherited from a single common ancestor.

As expected, isolates from phylogroup B2 demonstrated a high frequency of carriage for the greatest number of VFs (Johnson et al. 1991; Picard et al. 1999), while phylogroup A strains displayed the lowest prevalence of VFs (**Fig. 3A**). Adhesins,

Downloaded from genome.cshlp.org on June 16, 2026 . Published by Cold Spring Harbor Laboratory Press

273 particularly of the *ecp* and *fim* gene families, were the most prevalent VF across all
274 phylogroups. After correcting for population structure, few differences in VF content
275 between populations remained statistically significant, suggesting that most differences
276 among phylogroups reflect patterns of descent. The major exceptions were several iron
277 utilization genes (*iuc* and *ybt* families), toxin *tsh*, and protectin *traT*, which were significant
278 after accounting for strain relatedness and implies ongoing acquisition of these genes in
279 phylogroups B2 and D. Interestingly, differences between the related B1 and B2
280 phylogroups for several of these factors were only significant after accounting for
281 population structure, possibly reflecting convergent gene acquisition.

282 The prevalence of virulence factor genes observed in isolates from blood or urine
283 was similar overall (**Fig. 3B**), consistent with our observations about phylogenetic lineages
284 being able to infect both bodily sites. After population structure correction, most
285 statistically significant differences between these groups represented only minor
286 dissimilarities in overall VF prevalence, however, a handful of genes differed in prevalence
287 by at least a factor of 1.5. Five such genes were enriched in blood-derived strains: invasin
288 *traJ*, toxins *sat* and *tosA*, capsule *papG*, and adhesion *papA* Notably, *papA* has a known
289 role in urinary colonization (Lindberg et al. 1987; Denich et al. 1991; Johnson et al. 2000),
290 but its preferential enrichment in blood isolates implies importance of this VF outside of
291 uncomplicated uropathogenesis(Johnson et al. 2000). As expected, urinary-derived isolates
292 were enriched for members of the *auf* adhesion family (Buckles et al. 2004; Kaper et al.
293 2004), and toxin vat (Spurbeck et al. 2012). Differences in the prevalence of resistance to
294 10/23 antibiotics were also significant independently of population structure, in all cases at
295 higher prevalence in blood-derived isolates.

Downloaded from genome.cshlp.org on June 16, 2026 . Published by Cold Spring Harbor Laboratory Press

MLST groups that were recovered most frequently from our patients evidenced high prevalence for a large number of VFs compared to non-dominant MLST groups, especially pronounced for group ST127 (**Fig. 3C**). Much like our analysis at the phylogroup level, almost no statistically significant differences among these groups were evident after correcting for population structure, indicating that the innate virulence gene repertoire of individual groups almost entirely reflects heredity from a common ancestor.

De novo identification of antibiotic resistance factors. It is now becoming appreciated that large numbers of microbial genome sequences can be used for robust genome wide association studies (GWAS), enabling the detection of genetic factors underlying phenotypic variation (Falush and Bowden 2006; Farhat et al. 2013; Sheppard et al. 2013; Alam et al. 2014; Laabei et al. 2014). Here, in light of the open nature of the *E. coli* pan-genome, we observed a significant number of novel sequences present neither in reference genomes nor previous isolates with each additional strain that was sequenced. It was consequently not possible to comprehensively or effectively perform GWAS at single nucleotide resolution, nor in any way that relied on the use of a reference genome. As an alternative, we elected to examine associations at the level of discrete coding sequences that were identified in *de novo* assemblies.

We took this approach to identifying transmissible antibiotic resistance determinants within our study cohort –i.e., single gene factors conveying a phenotype that could be spread through a population via plasmids or other mobile elements. For each isolate, we determined the presence or absence of predicted genes found across the collection, then assessed the statistical significance of differences in the frequency that each gene was found in antibiotic resistant and susceptible strains. As before, we performed a

Downloaded from genome.cshlp.org on June 16, 2026 . Published by Cold Spring Harbor Laboratory Press

principal components analysis correction to account for population structure (Price et al. 2006).

With the exception of drugs from the fluoroquinolone class (which predominantly arise from chromosomal point mutation (Morgan-Linnell et al. 2009)), known resistance factors were strongly associated with a resistant phenotype for each antibiotic (**Supplemental Table S4**, *p*-values of $10^{-2.05}$ to $10^{-12.2}$, mean of $10^{-5.1}$**).** Transposases, conjugation factors, transcription factors, and plasmid maintenance factors were also highly associated with antibiotic resistance, consistent with physical linkage to mobile elements. After correcting for correlation with known resistance genes, we examined the most significant genes identified for each drug and carried seventeen genes forward for functional characterization in a laboratory strain. Bacteria transformed with known resistance factors (5/5) exhibited expected gains in antibiotic resistance; however, none of the potentially novel factors (0/17) conferred any detectable influence on resistance levels (**Supplemental Table S5**).

Discussion:

Given sustained decreases in the cost of high-throughput sequencing, we are approaching a time when it will be possible for clinical laboratories to sequence all clinical bacterial isolates, even as routine standard of care (Didelot et al. 2012; Schatz and Phillippy 2012). Here, we have attempted to provide an early glimpse as to how this kind of data can be utilized in a healthcare setting, and to demonstrate what kinds of information can be readily obtained from performing bacterial sequencing of clinical isolates on a large scale.

With respect to population structure, we found no evidence of a phylogenomic division between strains infecting either the blood or urinary tracts of our patients (**Fig.**

Downloaded from genome.cshlp.org on June 16, 2026 . Published by Cold Spring Harbor Laboratory Press

342 **1A**). This contrasts with studies of human-derived and environmental *E. coli* (Luo et al.
343 2011), and suggests that specific ExPEC *E. coli* lineages are, in general, not restricted to
344 invasion of one of these bodily sites or the other. The distribution of isolates across *E. coli*
345 phylogroups was consistent with earlier reports; however, we identified several instances of
346 strains being misclassified to the incorrect phylogroup based on a standard, three marker
347 classification system (Escobar-Paramo et al. 2006), presumably due to unexpected loss or
348 gain of relevant genetic material. Although apparently rare, the potential for such events to
349 mislead phylogrouping analysis should be acknowledged, and argues for a comprehensive,
350 genomic approach to phylogroup determination. Whole genome sequencing also offers
351 substantially higher resolution than conventional strain typing approaches, revealing
352 population structure underlying MLST groups including the dominant ST131 clade (**Fig.**
353 **1A**). Interestingly, phylogenomic analysis suggests that multidrug resistance in subclone
354 ST131-H30 is unstable, and can be lost by some descendants over time (**Fig. 2**).

355 Most of the statistically significant differences in VF content and antibiotic
356 resistance phenotypes which distinguish population-level groups of *E. coli* do not persist
357 after taking population structure into account (**Fig. 3**), arguing that most reflect inheritance
358 by descent from an ancestral strain and are a product of population structure alone.
359 Nevertheless, a small subset of factors do remain significant after correction for population
360 structure, suggesting that at least some genes have undergone multiple instances of
361 independent acquisition or loss within lineages, and potentially identifying them as
362 important to specific lineage groups or routes of infection.

363 Patient-to-patient transmission of *E. coli* appears to be infrequent among the set of
364 patients we examined. Anecdotal exploration of one case of suspected nosocomial

Downloaded from genome.cshlp.org on June 16, 2026 . Published by Cold Spring Harbor Laboratory Press

transmission was ruled out in light of genomic data. Nevertheless, we found evidence for a pair of genomically indistinguishable isolates and a group of three closely related strains that were shared across multiple patients. Although the trio of related isolates were obtained within several weeks of one another, collection of the paired isolates representing a true genomic clone occurred more than a year apart and in the absence of a clear epidemiological link, perhaps indicating an environmental reservoir or a dominant strain within the larger community (Manges et al. 2008). Indeed, rates of nosocomial transmission of lineages as assayed by conventional, lower-resolution typing technologies (Hilty et al. 2012) may have led to an overestimation of the frequency of such events, as has been observed for other bacteria (Miller et al. 2014; SenGupta et al. 2014). However, given our finding of multiple polyclonal *E. coli* infections, it should be noted that our conclusions may be influenced by incomplete clinical sampling of the multiple strains present in some infections (Lindsay 2014), reflecting a limitation of current clinical microbiological procedures. However, that we were able to detect potential transmission events without detailed biogeographic information or patient histories argues for the power of large-scale and unbiased sampling of clinical isolates as a means to monitor bacterial transmission. This genomic approach may provide unexpected advantages over existing molecular epidemiological techniques with lower throughput and resolution, potentially revealing outbreaks with unconventional properties (such as those spreading slowly and indolently), in addition to detailed population-level trends and direct transmission events.

Lastly, that our GWAS-type analysis was able to identify known antibiotic resistance factors, but in experimental assays failed to identify novel antibiotic resistance determinants suggests the possibility that, in this well-studied organism, all single-gene

Downloaded from genome.cshlp.org on June 16, 2026 . Published by Cold Spring Harbor Laboratory Press

antibiotic resistance factors present at reasonable prevalence have been previously identified. Polygenic causes of resistance may exist, and the factors we have identified in this study may contribute to multifactorial modes of resistance, however, dissecting such potentially complex pathways is outside the scope of our current work. Regardless, a GWAS-type approach based exclusively on genomic data and strain phenotypes robustly identified known antibiotic resistance genes, validating the general strategy as a means to catalog genes underlying other traits of interest, and in other microbial organisms.

The ability to perform whole genome surveys of bacteria without bias and at the scale of entire health care networks has the potential to provide in-depth information about many aspects of bacterial pathogens. As this study has demonstrated, single data sets of this nature enable more comprehensive and multifactorial examination of even well-characterized pathogens like *E. coli*, both in a clinical context and from a more basic science perspective.

Downloaded from genome.cshlp.org on June 16, 2026 . Published by Cold Spring Harbor Laboratory Press

Methods:

Samples and functional strain characterization: All isolates were identified and typed by the Microbiology Laboratory at the University of Washington Medical Center (Seattle, WA), using routine clinical practices. Antibiotic resistance was assessed using a combination of Kirby-Bauer antibiotic testing and automated MIC drug testing (Sensititre system, *TREK* Diagnostic Systems). Use of specimens was approved by the University of Washington Human Subjects Review Committee.

Library preparation and Sequencing: DNA was extracted using Wizard Genomic DNA Purification Kit (Promega). Shotgun sequencing libraries were prepared using the Nextera system V1 (Epicentre BioTechnologies), PCR amplified using FailSafe E PCR mix (Epicentre), monitored by real-time PCR, and removed when exponential growth of product was first observed. Libraries were purified using Agencourt AMPure XP (Beckman Coulter). Pools of 96 uniquely indexed samples were sequenced using an Illumina HiSeq 2000 with 101 bp paired-end chemistries.

Core and Pan-genome analysis: Reads were adaptor trimmed and subjected to *de novo* genome assembly using ABySS (version 1.3.5) (Simpson et al. 2009), using *k*-mer values (range 20-48) empirically determined to maximize contiguity on a per-sample basis (**Supplemental Dataset 7**). Contigs less than 500 bp in length were discarded as likely misassemblies. The mean N50 statistic for all genomes was 183.6±97.9 Kb (**Supplemental Figure S6**). Gene predictions were made using Glimmer3.02b (Delcher et al. 2007). A "meta-reference" was next constructed to represent all unique coding sequences (CDS) in all strains. CDSs were extracted from 53 fully sequenced *E. coli* reference genomes (**Supplemental Dataset 8**) and were first clustered using CD-HIT v4.6 (Li and Godzik

Downloaded from genome.cshlp.org on June 16, 2026 . Published by Cold Spring Harbor Laboratory Press

2006) (arguments -n 3 -c 0.8 -G 1 -aL 0.8 -aS 0.8 -B 1) to de-duplicate sequences ≥80% identical. Experimental gene predictions were compared to the de-duplicated reference CDS using BLASTP (Altschul et al. 1990) and sequences with ≥90% identity and ≥33% coverage to a reference CDS were discarded. Remaining gene predictions were de-duplicated as before and merged with the reference CDS to form the final meta-reference. Meta-reference sequences were functionally annotated using DAVID v6.7 (Huang da et al. 2009), genes classes found in the pan-genome and core genome were tabulated, and classes with the greatest-fold enrichment in comparing the two sets were evaluated. BLASTX was used to search *de novo* assemblies against the meta-reference, and a CDS was considered present in a strain if ≥80% of the CDS was covered by an alignment and protein-level identity was ≥80%.

Pan-genome estimates were performed for sequences of ≥75 amino acids in length. We found that assemblies with an N50 statistic of $<5x10^4$ bp did not reliably contain a full complement of essential *E. coli* genes (Hashimoto et al. 2005), so we limited our analysis to the 283 strains passing this cutoff. 2,000 different random input orders of genomes were performed (Touchon et al. 2009) for a subset size 1 to 282, and quartiles calculated for each. Estimations were performed against the complete meta-reference and after removing likely phage sequences and insertion sequences, identified by BLAST search against a prophage database as described (Zhou et al. 2011). For each gene, the highest number of strains for which the gene was present in ≥95% of isolates (Kaas et al. 2012) was calculated over 2,000 different random input orders. Individual genes were counted as part of the core genome for all numbers of strains up to and including this number of strains. For comparative studies of pathogenic and commensal strains, the meta-reference was

Downloaded from genome.cshlp.org on June 16, 2026 . Published by Cold Spring Harbor Laboratory Press

subjected to BLASTX analysis against commensal reference genomes (Hall et al. 2013) as above, and sequences shared between paired queries were flagged.

Molecular epidemiology: Adaptor-trimmed sequence reads were aligned to *E. coli* K12 MG1665 (Genbank ID: 556503834) using BWA (v0.6.2) (Li and Durbin 2009) and SAMtools (v0.1.19) (Li et al. 2009), yielding a mean coverage depth of 39.6±25.5 (**Supplemental Figure S4**). Single nucleotide variant calling was performed using SAMtools, and variants supported by fewer than 10 reads or a likelihood score of <200 marked as "unknown" data. A total of 446,152 unique variant sites were found across all isolates. To filter out low quality genomes, isolates with ambiguous calls at 80% or more of total variant sites were excluded from phylogenetic reconstructions. Approximately-maximum-likelihood phylogenetic trees were made using FastTree 2.1 (Price et al. 2010).

Isolates were assigned to phylogenetic subgroups based on a three genetic marker system (Escobar-Paramo et al. 2006), using a BLAST search against *de novo* assemblies to register presence or absence. Classification of isolates assigned to a different phylogroup than phylogenomically related isolates was confirmed by examining depth of short reads against aligned to each of the three genetic markers. Sequence types were assigned by a BLAST search of assembled genomes to identify perfect matches against known MLST fragments from an established database (http://mlst.ucc.ie/, accessed 1/22/14). The pattern of MLST types for each locus was compared to reported sequence types. Strains for which one or more MLST loci could not be identified (14 isolates) or those bearing new locus sequences were unassigned. Assignment of strains as ST131-H30 was based on exact BLAST match to a partial *fimH* 30 allele (Colpan et al. 2013) (GenBank ID: 268639126).

Downloaded from genome.cshlp.org on June 16, 2026 . Published by Cold Spring Harbor Laboratory Press

Characterization of known virulence and antibiotic resistance factors. VF reference sequences were identified through a combination of the Virulence Factor Database (Chen et al. 2012) and primary literature review (**Supplemental Dataset 9**). In all strains the presence of each VF was assessed using a BLAST search as above. VF and antibiotic resistance phenotypes were assessed within major phylogroups and sequence types containing ten or more isolates. Technical replicates and isolates failing quality control for the pan-genome analysis were excluded. Statistical association between presence or absence of each factor (VF or resistance phenotype) and isolate classification was assessed using a logistic regression framework in R 3.1.1 (Team 2014) for MLST or phylogroup membership, or for blood- or urine-derived isolates. The strength of association for factors perfectly predicted by isolate classification was assessed with a Bayesian logistic regression model (package *arm*) with independent Cauchy priors, mean 0, and scale 5/2 for each coefficient. Only factors meeting a significance threshold of $p < 0.05$ in any of the three classification models (MLST, phylogroup, or blood vs. urine source) were carried forward for post-hoc pairwise testing. Differences in factor distribution or resistance phenotype for all pairwise comparisons within a classification scheme were evaluated using Tukey's HSD method (package *lsmeans*). As this method is statistically conservative, significant pairwise comparisons were identified using raw *p*-values. To account for possible lack of independence between isolates due to underlying population structure, the analysis was repeated using an additive logistic regression framework with three additional covariates: the first three principal components resulting from the decomposition of the matrix of presence or absence genotypes of each CDS across all isolates (described below).

Downloaded from genome.cshlp.org on June 16, 2026 . Published by Cold Spring Harbor Laboratory Press

Genome Wide Association Analysis for Antibiotic Resistance Phenotypes. BLASTX was used to search *de novo* assemblies against the meta-reference, and the presence or absence every CDS ≥ 75 amino acids in length from the meta-reference was assessed using alignment and coverage metrics as before. A logistic regression model was implemented in R 3.1.1 as above. To control for population structure, the matrix of presence or absence for each CDS and for each isolate was decomposed into principal components. Q-Q plots were evaluated by manual review (**Supplemental Figure S7**) and the first three principal components of the matrix were empirically determined as optimal and were retained as covariates in the model. To identify independently associated factors, known associations for each drug resistance phenotype were used as covariates. The 20 CDS with the most significant *p*-values were considered for each antibiotic. CDS corresponding to repetitive elements, transposons, insertion sequences, plasmid support machinery, resistance factors for other drugs or unrelated biochemical pathways, and CDS occurring at ≥15 % frequency in the antibiotic sensitive population were excluded to limit spurious associations due to linkage. Genes of interest (**Supplemental Table S5)** were cloned into pET-9a or pET-3a expression vectors (Novagen) using HD Infusion kit (ClonTech). Transformed *E. coli* BL21(DE3) (NEB) were induced using 0.3mM IPTG and subjected to antibiotic resistance testing by E-test (bioMérieux).

Data Access:

Sequence data generated for this study have been submitted to the NCBI Sequence Read Archive (SRA; http://www.ncbi.nlm.nih.gov/sra) under study accession number SRP042632.

Acknowledgements:

Downloaded from genome.cshlp.org on June 16, 2026 . Published by Cold Spring Harbor Laboratory Press

This work was supported by developmental research project grant 5U54AI057141-08REV from the National Institutes of Allergy and Infectious Diseases (NIAID)/Northwest Regional Center of Excellence for Biodefense and Emerging Infectious Diseases (NWRCE).

Disclosure Declarations:

The authors declare no competing financial interests or other conflicts of interest.

Downloaded from genome.cshlp.org on June 16, 2026 . Published by Cold Spring Harbor Laboratory Press

Figure Legends:

Figure 1. Whole genome phylogenetic tree of ExPEC E. coli isolates. A) Approximate maximum likelihood phylogeny showing the population structure of ExPEC *E. coli*. Isolates cultured from blood are represented as red terminal nodes and those cultured from urine are shown in blue. Colored ring denotes annotation of major *E. coli* phylogroups. Seven isolates assigned to phylogroups that are inconsistent with their phylogenomic placement are indicated with colored bars internal to this ring. The outermost ring (black) indicates groups of MLST sequence types. Sequence types with at least two representatives are numbered. The group corresponding to subclone S131-H30 is indicated. B) Approximate maximum likelihood phylogeny of blood isolates, only. Isolates are labeled according to the patient of origin and the relative day of collection (in red, ranging from day 0 for patient 43 to day 1184 for patient isolate 3_2). In instances where multiple isolates were obtained from the same patient, the order in which specimens were recovered is indicated by an underscore and a number. Patients for which multiple, genomically distinct strains were identified are highlighted. Isolates from patients 1 and 29 are indicated by blue text. The group corresponding to subclone ST131-H30 is indicated. Scale bars are expressed in changes per site for both panels.

Figure 2. Whole genome phylogenetic tree of ST131 isolates. The ST131-H30 subgroup is indicated, and is marked by a high prevalence of fluoroquinolone resistance and extended-spectrum β-lactamase (ESBL) activity. Node color indicates the relevant drug resistance phenotype (white circle indicates missing data). Nodes supported by log likelihood values below 0.8 are marked with a black circle. Scale bar is expressed in changes per site.

Downloaded from genome.cshlp.org on June 16, 2026 . Published by Cold Spring Harbor Laboratory Press

Figure 3. Proportion and relative enrichment of virulence factors and antibiotic resistance phenotypes carried by isolates in distinct groups. Rows correspond to individual VFs (top) or antibiotic resistance phenotypes (Abx, bottom). VFs are grouped by class (Tox, toxin; Prot, protectin; Iron, iron metabolism; Tr, transporter; Cap, capsule; Inv, invasin; Adh, adhesin; Mi, miscellaneous). Columns correspond to categories of isolates grouped according to different classification schemes. Prevalence of factors within each category is shown at left for each panel (blue heatmap). Raw *p*-values from all possible pairwise comparisons of factor prevalence between is shown at right for each panel (green heatmap), with the specific pairwise comparison indicated above each column. *p*-values were obtained after correcting for inferred population structure (left, labeled in red with "Corrected" or "C") or without such correction (right, labeled in red "Uncorrected" or "U"). A) Comparison of *E. coli* phylogroups. B) Comparison of isolates obtained from blood (B) and urine (U). C) Comparison of the six most prevalent MLST groups (sequence type numbers are indicated) and a seventh category encompassing all other MLST groups ("O").

Downloaded from genome.cshlp.org on June 16, 2026 . Published by Cold Spring Harbor Laboratory Press

561 **References:**

562 Adams-Sapper S, Diep BA, Perdreau-Remington F, Riley LW. 2013. Clonal composition
563 and community clustering of drug-susceptible and -resistant Escherichia coli
564 isolates from bloodstream infections. *Antimicrobial agents and chemotherapy*
565 **57**(1): 490-497.
566 Alam MT, Petit RA, 3rd, Crispell EK, Thornton TA, Conneely KN, Jiang Y, Satola SW,
567 Read TD. 2014. Dissecting vancomycin-intermediate resistance in staphylococcus
568 aureus using genome-wide association. *Genome biology and evolution* **6**(5): 1174-
569 1185.
570 Alsterlund R, Axelsson C, Olsson-Liljequist B. 2012. Long-term carriage of extended-
571 spectrum beta-lactamase-producing Escherichia coli. *Scandinavian journal of*
572 *infectious diseases* **44**(1): 51-54.
573 Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment
574 search tool. *Journal of molecular biology* **215**(3): 403-410.
575 Barrick JE, Yu DS, Yoon SH, Jeong H, Oh TK, Schneider D, Lenski RE, Kim JF. 2009.
576 Genome evolution and adaptation in a long-term experiment with Escherichia coli.
577 *Nature* **461**(7268): 1243-1247.
578 Buckles EL, Bahrani-Mougeot FK, Molina A, Lockatell CV, Johnson DE, Drachenberg
579 CB, Burland V, Blattner FR, Donnenberg MS. 2004. Identification and
580 characterization of a novel uropathogenic Escherichia coli-associated fimbrial gene
581 cluster. *Infection and immunity* **72**(7): 3890-3901.
582 Carlos C, Pires MM, Stoppe NC, Hachich EM, Sato MI, Gomes TA, Amaral LA, Ottoboni
583 LM. 2010. Escherichia coli phylogenetic group determination and its application in

Downloaded from genome.cshlp.org on June 16, 2026 . Published by Cold Spring Harbor Laboratory Press

584 the identification of the major animal source of fecal contamination. *BMC*
585 *microbiology* **10**: 161.
586 Chen L, Xiong Z, Sun L, Yang J, Jin Q. 2012. VFDB 2012 update: toward the genetic
587 diversity and molecular evolution of bacterial virulence factors. *Nucleic acids*
588 *research* **40**(Database issue): D641-645.
589 Chin CS, Sorenson J, Harris JB, Robins WP, Charles RC, Jean-Charles RR, Bullard J,
590 Webster DR, Kasarskis A, Peluso P et al. 2011. The origin of the Haitian cholera
591 outbreak strain. *The New England journal of medicine* **364**(1): 33-42.
592 Clermont O, Bonacorsi S, Bingen E. 2000. Rapid and simple determination of the
593 Escherichia coli phylogenetic group. *Applied and environmental microbiology*
594 **66**(10): 4555-4558.
595 Colpan A, Johnston B, Porter S, Clabots C, Anway R, Thao L, Kuskowski MA,
596 Tchesnokova V, Sokurenko EV, Johnson JR et al. 2013. Escherichia coli sequence
597 type 131 (ST131) subclone H30 as an emergent multidrug-resistant pathogen
598 among US veterans. *Clinical infectious diseases : an official publication of the*
599 *Infectious Diseases Society of America* **57**(9): 1256-1265.
600 Delcher AL, Bratke KA, Powers EC, Salzberg SL. 2007. Identifying bacterial genes and
601 endosymbiont DNA with Glimmer. *Bioinformatics* **23**(6): 673-679.
602 Delsuc F, Brinkmann H, Philippe H. 2005. Phylogenomics and the reconstruction of the
603 tree of life. *Nature reviews Genetics* **6**(5): 361-375.
604 Denich K, Blyn LB, Craiu A, Braaten BA, Hardy J, Low DA, O'Hanley PD. 1991. DNA
605 sequences of three papA genes from uropathogenic Escherichia coli strains:

Downloaded from genome.cshlp.org on June 16, 2026 . Published by Cold Spring Harbor Laboratory Press

606 evidence of structural and serological conservation. *Infection and immunity* **59**(11):
607 3849-3858.
608 Didelot X, Bowden R, Wilson DJ, Peto TE, Crook DW. 2012. Transforming clinical
609 microbiology with bacterial genome sequencing. *Nature reviews Genetics* **13**(9):
610 601-612.
611 Escobar-Paramo P, Le Menac'h A, Le Gall T, Amorin C, Gouriou S, Picard B, Skurnik D,
612 Denamur E. 2006. Identification of forces shaping the commensal Escherichia coli
613 genetic structure by comparing animal and human isolates. *Environmental*
614 *microbiology* **8**(11): 1975-1984.
615 Falush D, Bowden R. 2006. Genome-wide association mapping in bacteria? *Trends in*
616 *microbiology* **14**(8): 353-355.
617 Farhat MR, Shapiro BJ, Kieser KJ, Sultana R, Jacobson KR, Victor TC, Warren RM,
618 Streicher EM, Calver A, Sloutsky A et al. 2013. Genomic analysis identifies targets
619 of convergent positive selection in drug-resistant Mycobacterium tuberculosis.
620 *Nature genetics* **45**(10): 1183-1189.
621 Foxman B, Zhang L, Tallman P, Andree BC, Geiger AM, Koopman JS, Gillespie BW,
622 Palin KA, Sobel JD, Rode CK et al. 1997. Transmission of uropathogens between
623 sex partners. *J Infect Dis* **175**(4): 989-992.
624 Gibreel TM, Dodgson AR, Cheesbrough J, Fox AJ, Bolton FJ, Upton M. 2012. Population
625 structure, virulence potential and antibiotic susceptibility of uropathogenic
626 Escherichia coli from Northwest England. *The Journal of antimicrobial*
627 *chemotherapy* **67**(2): 346-356.

Downloaded from genome.cshlp.org on June 16, 2026 . Published by Cold Spring Harbor Laboratory Press

628 Gupta SK, Nanda V, Malviya P, Jacobs N, Naheed Z, Joseph T. 2013. An Unusual Case of
629 Early Onset Persistent Escherichia coli Septicemia Associated with Endocarditis.
630 *AJP reports* **3**(2): 105-106.
631 Hall BG, Cardenas H, Barlow M. 2013. Using complete genome comparisons to identify
632 sequences whose presence accurately predicts clinically important phenotypes. *PloS*
633 *one* **8**(7): e68901.
634 Hashimoto M, Ichimura T, Mizoguchi H, Tanaka K, Fujimitsu K, Keyamura K, Ote T,
635 Yamakawa T, Yamazaki Y, Mori H et al. 2005. Cell size and nucleoid organization
636 of engineered Escherichia coli cells with a reduced genome. *Molecular*
637 *microbiology* **55**(1): 137-149.
638 Hilty M, Betsch BY, Bogli-Stuber K, Heiniger N, Stadler M, Kuffer M, Kronenberg A,
639 Rohrer C, Aebi S, Endimiani A et al. 2012. Transmission dynamics of extended-
640 spectrum beta-lactamase-producing Enterobacteriaceae in the tertiary care hospital
641 and the household setting. *Clinical infectious diseases : an official publication of*
642 *the Infectious Diseases Society of America* **55**(7): 967-975.
643 Huang da W, Sherman BT, Lempicki RA. 2009. Systematic and integrative analysis of
644 large gene lists using DAVID bioinformatics resources. *Nature protocols* **4**(1): 44-
645 57.
646 Jaureguy F, Landraud L, Passet V, Diancourt L, Frapy E, Guigon G, Carbonnelle E,
647 Lortholary O, Clermont O, Denamur E et al. 2008. Phylogenetic and genomic
648 diversity of human bacteremic Escherichia coli strains. *BMC genomics* **9**: 560.
649 Johnson JR, Goullet P, Picard B, Moseley SL, Roberts PL, Stamm WE. 1991. Association
650 of carboxylesterase B electrophoretic pattern with presence and expression of

Downloaded from genome.cshlp.org on June 16, 2026 . Published by Cold Spring Harbor Laboratory Press

651 urovirulence factor determinants and antimicrobial resistance among strains of
652 Escherichia coli that cause urosepsis. *Infection and immunity* **59**(7): 2311-2315.
653 Johnson JR, Owens K, Gajewski A, Clabots C. 2008. Escherichia coli colonization patterns
654 among human household members and pets, with attention to acute urinary tract
655 infection. *J Infect Dis* **197**(2): 218-224.
656 Johnson JR, Owens KL, Clabots CR, Weissman SJ, Cannon SB. 2006. Phylogenetic
657 relationships among clonal groups of extraintestinal pathogenic Escherichia coli as
658 assessed by multi-locus sequence analysis. *Microbes and infection / Institut Pasteur*
659 **8**(7): 1702-1713.
660 Johnson JR, Stell AL, Scheutz F, O'Bryan TT, Russo TA, Carlino UB, Fasching C, Kavle
661 J, Van Dijk L, Gaastra W. 2000. Analysis of the F antigen-specific papA alleles of
662 extraintestinal pathogenic Escherichia coli using a novel multiplex PCR-based
663 assay. *Infection and immunity* **68**(3): 1587-1599.
664 Kaas RS, Friis C, Ussery DW, Aarestrup FM. 2012. Estimating variation within the genes
665 and inferring the phylogeny of 186 sequenced diverse Escherichia coli genomes.
666 *BMC genomics* **13**: 577.
667 Kaper JB, Nataro JP, Mobley HL. 2004. Pathogenic Escherichia coli. *Nature reviews*
668 *Microbiology* **2**(2): 123-140.
669 Kennedy AD, Porcella SF, Martens C, Whitney AR, Braughton KR, Chen L, Craig CT,
670 Tenover FC, Kreiswirth BN, Musser JM et al. 2010. Complete nucleotide sequence
671 analysis of plasmids in strains of Staphylococcus aureus clone USA300 reveals a
672 high level of identity among isolates with closely related core genome sequences.
673 *Journal of clinical microbiology* **48**(12): 4504-4511.

Downloaded from genome.cshlp.org on June 16, 2026 . Published by Cold Spring Harbor Laboratory Press

674 Koser CU, Holden MT, Ellington MJ, Cartwright EJ, Brown NM, Ogilvy-Stuart AL, Hsu
675 LY, Chewapreecha C, Croucher NJ, Harris SR et al. 2012. Rapid whole-genome
676 sequencing for investigation of a neonatal MRSA outbreak. *The New England*
677 *journal of medicine* **366**(24): 2267-2275.
678 Kumar S, Filipski AJ, Battistuzzi FU, Kosakovsky Pond SL, Tamura K. 2012. Statistics
679 and truth in phylogenomics. *Molecular biology and evolution* **29**(2): 457-472.
680 Laabei M, Recker M, Rudkin JK, Aldeljawi M, Gulay Z, Sloan TJ, Williams P, Endres JL,
681 Bayles KW, Fey PD et al. 2014. Predicting the virulence of MRSA from its genome
682 sequence. *Genome research* **24**(5): 839-849.
683 Lau SH, Reddy S, Cheesbrough J, Bolton FJ, Willshaw G, Cheasty T, Fox AJ, Upton M.
684 2008. Major uropathogenic Escherichia coli strain isolated in the northwest of
685 England identified by multilocus sequence typing. *Journal of clinical microbiology*
686 **46**(3): 1076-1080.
687 Lecointre G, Rachdi L, Darlu P, Denamur E. 1998. Escherichia coli molecular phylogeny
688 using the incongruence length difference test. *Molecular biology and evolution*
689 **15**(12): 1685-1695.
690 Lee H, Popodi E, Tang H, Foster PL. 2012. Rate and molecular spectrum of spontaneous
691 mutations in the bacterium Escherichia coli as determined by whole-genome
692 sequencing. *Proceedings of the National Academy of Sciences of the United States*
693 *of America* **109**(41): E2774-2783.
694 Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler
695 transform. *Bioinformatics* **25**(14): 1754-1760.

Downloaded from genome.cshlp.org on June 16, 2026 . Published by Cold Spring Harbor Laboratory Press

696 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G,
697 Durbin R, Genome Project Data Processing S. 2009. The Sequence Alignment/Map
698 format and SAMtools. *Bioinformatics* **25**(16): 2078-2079.
699 Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of
700 protein or nucleotide sequences. *Bioinformatics* **22**(13): 1658-1659.
701 Lieberman TD, Michel JB, Aingaran M, Potter-Bynoe G, Roux D, Davis MR, Jr., Skurnik
702 D, Leiby N, Lipuma JJ, Goldberg JB et al. 2011. Parallel bacterial evolution within
703 multiple patients identifies candidate pathogenicity genes. *Nature genetics* **43**(12):
704 1275-1280.
705 Lindberg F, Lund B, Johansson L, Normark S. 1987. Localization of the receptor-binding
706 protein adhesin at the tip of the bacterial pilus. *Nature* **328**(6125): 84-87.
707 Lindsay JA. 2014. Evolution of Staphylococcus aureus and MRSA during outbreaks.
708 *Infection, genetics and evolution : journal of molecular epidemiology and*
709 *evolutionary genetics in infectious diseases* **21**: 548-553.
710 Luo C, Walk ST, Gordon DM, Feldgarden M, Tiedje JM, Konstantinidis KT. 2011.
711 Genome sequencing of environmental Escherichia coli expands understanding of
712 the ecology and speciation of the model bacterial species. *Proceedings of the*
713 *National Academy of Sciences of the United States of America* **108**(17): 7200-7205.
714 Manges AR, Dietrich PS, Riley LW. 2004. Multidrug-resistant Escherichia coli clonal
715 groups causing community-acquired pyelonephritis. *Clinical infectious diseases :*
716 *an official publication of the Infectious Diseases Society of America* **38**(3): 329-334.

Downloaded from genome.cshlp.org on June 16, 2026 . Published by Cold Spring Harbor Laboratory Press

717 Manges AR, Tabor H, Tellis P, Vincent C, Tellier PP. 2008. Endemic and epidemic
718 lineages of Escherichia coli that cause urinary tract infections. *Emerging infectious*
719 *diseases* **14**(10): 1575-1583.
720 Martin GS, Mannino DM, Eaton S, Moss M. 2003. The epidemiology of sepsis in the
721 United States from 1979 through 2000. *The New England journal of medicine*
722 **348**(16): 1546-1554.
723 Miller RM, Price JR, Batty EM, Didelot X, Wyllie D, Golubchik T, Crook DW, Paul J,
724 Peto TE, Wilson DJ et al. 2014. Healthcare-associated outbreak of meticillin-
725 resistant Staphylococcus aureus bacteraemia: role of a cryptic variant of an
726 epidemic clone. *The Journal of hospital infection* **86**(2): 83-89.
727 Morgan-Linnell SK, Becnel Boyd L, Steffen D, Zechiedrich L. 2009. Mechanisms
728 accounting for fluoroquinolone resistance in Escherichia coli clinical isolates.
729 *Antimicrobial agents and chemotherapy* **53**(1): 235-241.
730 Nowrouzian FL, Adlerberth I, Wold AE. 2006. Enhanced persistence in the colonic
731 microbiota of Escherichia coli strains belonging to phylogenetic group B2: role of
732 virulence factors and adherence to colonic cells. *Microbes and infection / Institut*
733 *Pasteur* **8**(3): 834-840.
734 Picard B, Garcia JS, Gouriou S, Duriez P, Brahimi N, Bingen E, Elion J, Denamur E. 1999.
735 The link between phylogeny and virulence in Escherichia coli extraintestinal
736 infection. *Infection and immunity* **67**(2): 546-553.
737 Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. 2006. Principal
738 components analysis corrects for stratification in genome-wide association studies.
739 *Nature genetics* **38**(8): 904-909.

Downloaded from genome.cshlp.org on June 16, 2026 . Published by Cold Spring Harbor Laboratory Press

740 Price LB, Johnson JR, Aziz M, Clabots C, Johnston B, Tchesnokova V, Nordstrom L,
741 Billig M, Chattopadhyay S, Stegger M et al. 2013. The epidemic of extended-
742 spectrum-beta-lactamase-producing Escherichia coli ST131 is driven by a single
743 highly pathogenic subclone, H30-Rx. *mBio* **4**(6): e00377-00313.
744 Price MN, Dehal PS, Arkin AP. 2010. FastTree 2--approximately maximum-likelihood
745 trees for large alignments. *PloS one* **5**(3): e9490.
746 Ramos NL, Saayman ML, Chapman TA, Tucker JR, Smith HV, Faoagali J, Chin JC,
747 Brauner A, Katouli M. 2010. Genetic relatedness and virulence gene profiles of
748 Escherichia coli strains isolated from septicaemic and uroseptic patients. *European*
749 *journal of clinical microbiology & infectious diseases : official publication of the*
750 *European Society of Clinical Microbiology* **29**(1): 15-23.
751 Rasko DA, Rosovitz MJ, Myers GS, Mongodin EF, Fricke WF, Gajer P, Crabtree J,
752 Sebaihia M, Thomson NR, Chaudhuri R et al. 2008. The pangenome structure of
753 Escherichia coli: comparative genomic analysis of E. coli commensal and
754 pathogenic isolates. *Journal of bacteriology* **190**(20): 6881-6893.
755 Ron EZ. 2010. Distribution and evolution of virulence factors in septicemic Escherichia
756 coli. *Int J Med Microbiol* **300**(6): 367-370.
757 Samet A, Sledzinska A, Krawczyk B, Hellmann A, Nowicki S, Kur J, Nowicki B. 2013.
758 Leukemia and risk of recurrent Escherichia coli bacteremia: genotyping implicates
759 E. coli translocation from the colon to the bloodstream. *European journal of*
760 *clinical microbiology & infectious diseases : official publication of the European*
761 *Society of Clinical Microbiology* **32**(11): 1393-1400.

Downloaded from genome.cshlp.org on June 16, 2026 . Published by Cold Spring Harbor Laboratory Press

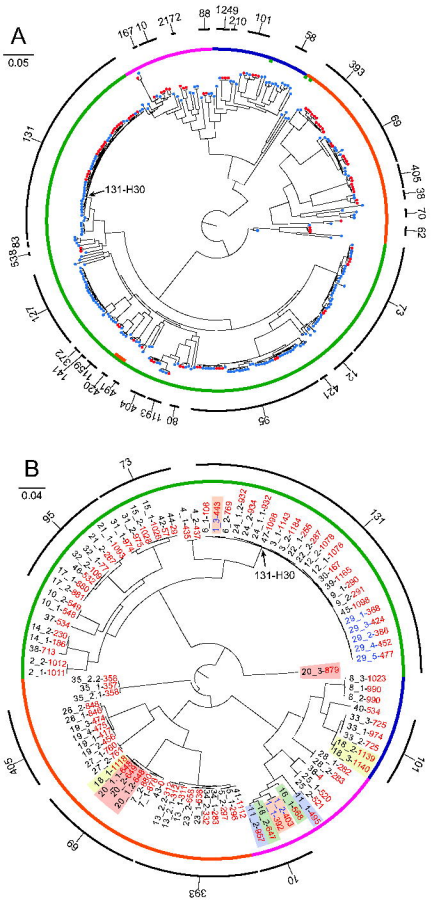
762 Sanjar F, Hazen TH, Shah SM, Koenig SS, Agrawal S, Daugherty S, Sadzewicz L, Tallon
763 LJ, Mammel MK, Feng P et al. 2014. Genome Sequence of Escherichia coli
764 O157:H7 Strain 2886-75, Associated with the First Reported Case of Human
765 Infection in the United States. *Genome announcements* **2**(1).
766 Sannes MR, Kuskowski MA, Owens K, Gajewski A, Johnson JR. 2004. Virulence factor
767 profiles and phylogenetic background of Escherichia coli isolates from veterans
768 with bacteremia and uninfected control subjects. *J Infect Dis* **190**(12): 2121-2128.
769 Schatz MC, Phillippy AM. 2012. The rise of a digital immune system. *GigaScience* **1**(1): 4.
770 SenGupta DJ, Cummings LA, Hoogestraat DR, Butler-Wu SM, Shendure J, Cookson BT,
771 Salipante SJ. 2014. Whole-Genome Sequencing for High-Resolution Investigation
772 of Methicillin-Resistant Staphylococcus aureus Epidemiology and Genome
773 Plasticity. *Journal of clinical microbiology* **52**(8): 2787-2796.
774 Sheppard SK, Didelot X, Meric G, Torralbo A, Jolley KA, Kelly DJ, Bentley SD, Maiden
775 MC, Parkhill J, Falush D. 2013. Genome-wide association study identifies vitamin
776 B5 biosynthesis as a host specificity factor in Campylobacter. *Proceedings of the*
777 *National Academy of Sciences of the United States of America* **110**(29): 11923-
778 11927.
779 Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I. 2009. ABySS: a parallel
780 assembler for short read sequence data. *Genome research* **19**(6): 1117-1123.
781 Snitkin ES, Zelazny AM, Thomas PJ, Stock F, Henderson DK, Palmore TN, Segre JA.
782 2012. Tracking a Hospital Outbreak of Carbapenem-Resistant Klebsiella
783 pneumoniae with Whole-Genome Sequencing. *Science translational medicine*
784 **4**(148): 148ra116.

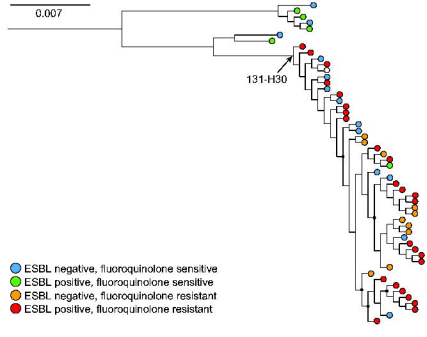
Downloaded from genome.cshlp.org on June 16, 2026 . Published by Cold Spring Harbor Laboratory Press

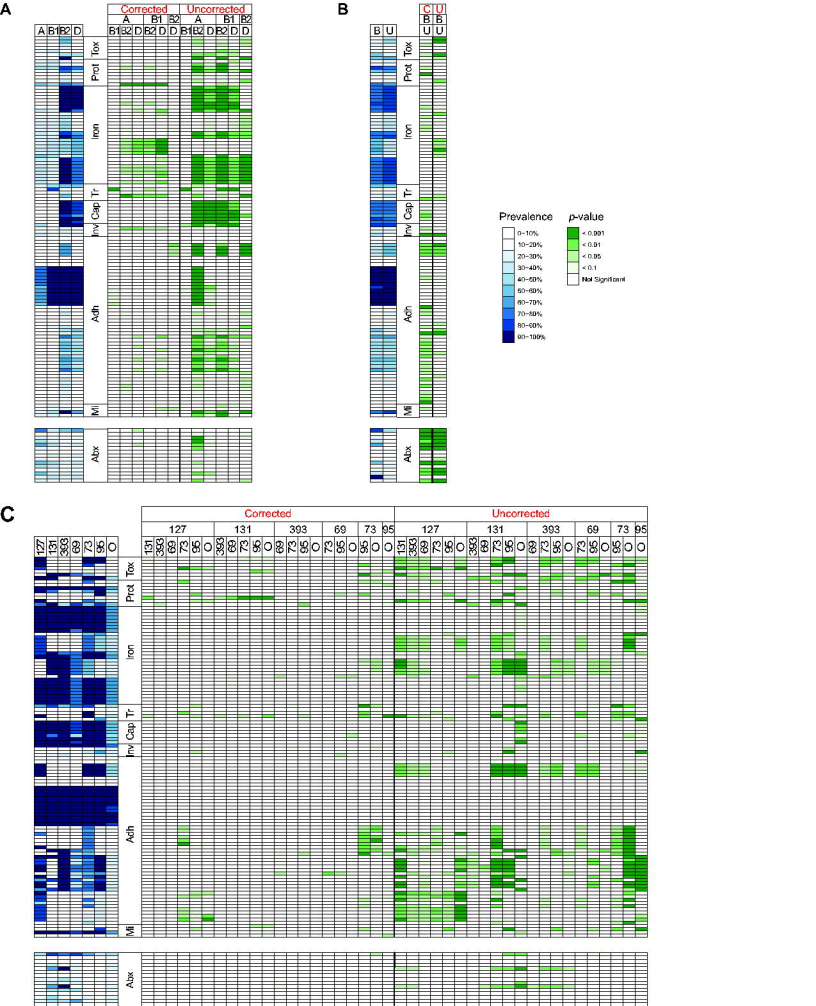
785 Spurbeck RR, Dinh PC, Jr., Walk ST, Stapleton AE, Hooton TM, Nolan LK, Kim KS,
786 Johnson JR, Mobley HL. 2012. Escherichia coli isolates that carry vat, fyuA, chuA,
787 and yfcV efficiently colonize the urinary tract. *Infection and immunity* **80**(12):
788 4115-4122.
789 Tartof SY, Solberg OD, Manges AR, Riley LW. 2005. Analysis of a uropathogenic
790 Escherichia coli clonal group by multilocus sequence typing. *Journal of clinical*
791 *microbiology* **43**(12): 5860-5864.
792 Team RC. 2014. R: A Language and Environment for Statistical Computing. R Foundation
793 for Statistical Computing, Vienna, Austria.
794 Telli M, Guiral E, Martinez JA, Almela M, Bosch J, Vila J, Soto SM. 2010. Prevalence of
795 enterotoxins among Escherichia coli isolates causing bacteraemia. *FEMS Microbiol*
796 *Lett* **306**(2): 117-121.
797 Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli SV,
798 Crabtree J, Jones AL, Durkin AS et al. 2005. Genome analysis of multiple
799 pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-
800 genome". *Proceedings of the National Academy of Sciences of the United States of*
801 *America* **102**(39): 13950-13955.
802 Touchon M, Hoede C, Tenaillon O, Barbe V, Baeriswyl S, Bidet P, Bingen E, Bonacorsi S,
803 Bouchier C, Bouvet O et al. 2009. Organised genome dynamics in the Escherichia
804 coli species results in highly diverse adaptive paths. *PLoS genetics* **5**(1): e1000344.
805 Toval F, Kohler CD, Vogel U, Wagenlehner F, Mellmann A, Fruth A, Schmidt MA, Karch
806 H, Bielaszewska M, Dobrindt U. 2014. Characterization of Escherichia coli Isolates

Downloaded from genome.cshlp.org on June 16, 2026 . Published by Cold Spring Harbor Laboratory Press

807 from Hospital Inpatients or Outpatients with Urinary Tract Infection. *Journal of*
808 *clinical microbiology* **52**(2): 407-418.
809 Walker TM, Ip CL, Harrell RH, Evans JT, Kapatai G, Dedicoat MJ, Eyre DW, Wilson DJ,
810 Hawkey PM, Crook DW et al. 2013. Whole-genome sequencing to delineate
811 Mycobacterium tuberculosis outbreaks: a retrospective observational study. *The*
812 *Lancet infectious diseases* **13**(2): 137-146.
813 Zhou Y, Liang Y, Lynch KH, Dennis JJ, Wishart DS. 2011. PHAST: a fast phage search
814 tool. *Nucleic acids research* **39**(Web Server issue): W347-352.
815
816


A
0.05
167
10
2172
88
1249
1210
101
58
383
69
405
38
70
62
73
421
74
95
80
1193
404
497
1420
1158
1372
141
121
538
83
131
131-H30
B
0.04
73
131
101
10
393
69
405
98
131-H30
20_3-879
Isolate Source
Blood
Urine
Phylogroup
B2
D
A
B1


0.007
131-H30
ESBL negative, fluoroquinolone sensitive
ESBL positive, fluoroquinolone sensitive
ESBL negative, fluoroquinolone resistant
ESBL positive, fluoroquinolone resistant


A
Corrected
Uncorrected
A
B1
B2
D
Tox
Prot
Iron
Tr
Cap
Inv
Adh
Mi
Abx
B
C
U
E
Prevalence
0–10%
10–20%
20–30%
30–40%
40–50%
50–60%
60–70%
70–80%
80–90%
90–100%
p-value
< 0.001
< 0.01
< 0.05
< 0.1
Not Significant
C
127
131
393
69
73
95
O