Title:
Genetic features of antimicrobial drug-susceptible extraintestinal pathogenic *Escherichia coli* pandemic sequence type 95

Yuan Hu Allegretti, Reina Yamaji, Sheila Adams-Sapper, Lee W Riley

Keywords: *Escherichia coli*, ExPEC, ST95, drug resistance, pan-genome-wide association study

Abstract
Extraintestinal pathogenic *Escherichia coli* (ExPEC) belonging to multilocus sequence type 95 (ST95) is one of the most geographically widespread ExPEC lineages causing bloodstream (BSI) and urinary tract infections (UTI). In contrast to other widespread ExPEC sequence types, a large proportion ST95 strains remains susceptible to all antimicrobial agents used to treat BSI or UTI. We aimed to identify genomic features of ST95 associated with its relatively high drug-susceptible frequency. We analyzed whole genomes of 1749 ST95 isolates, 80 from patients with BSI or UTI in Northern California, and 1669 isolates from the Enterobase database. We first compared whole-genome sequences (WGS) of 887 drug-susceptible strains and 862 strains resistant to one or more drugs (defined genotypically as strains harboring drug-resistance genes annotated in the ResFinder database) to identify genetic features associated with strains devoid of drug-resistance genes. We then conducted a pan-genome-wide association study on human clinical isolates of ST95, which included 553 UTI and BSI ST95 isolates. We found 44 accessory genes to be significantly associated with ST95 strains lacking drug resistance genes. Fifteen of these were not found in any of the WGSs of ST131 ExPEC strains, which are frequently multidrug-resistant. These genes were annotated to encode transporter or transfer systems and DNA repair polymerases. A large proportion of ST95 strains may have evolved to adapt to antibiotic-imposed stresses without acquiring drug-resistance genes.

Importance
Despite the increasing prevalence of antibiotic-resistant *Escherichia coli* strains that cause urinary tract and bloodstream infections, a major pandemic lineage of extraintestinal pathogenic *E. coli* (ExPEC) ST95 has a comparatively low frequency of drug resistance. We compared genomes of 1749 ST95 isolates to identify genetic features that may explain why most strains of ST95 resist becoming drug-resistant. Identification of such genomic features could contribute to the development of novel strategies to prevent the spread of antibiotic-resistant genes and devise new measures to control antibiotic-resistant infections.

1. Introduction

Antimicrobial drug resistance is one of the most serious global public health concerns of our time, threatening the effective prevention and treatment of infectious diseases (1, 2). In particular, several drug-resistant Gram-negative bacterial species belonging to Enterobacteriaceae are designated by the Centers for Disease Control and Prevention (CDC) and the World Health Organization (WHO) as "urgent-threat" or "priority 1, critical" pathogens in need of new treatment and research (2, 3). Among Enterobacteriaceae, *Escherichia coli* (*E. coli*) is the most frequent cause of common extraintestinal infections such as urinary tract infections (UTI) and bloodstream infections (BSI) (4). *E. coli* strains that cause such infections are referred to as extraintestinal pathogenic *E. coli* (ExPEC) (5). The prevalence and incidence of infections caused by ExPEC strains resistant to a wide spectrum of antimicrobial agents, including extended-spectrum beta-lactam (ESBL) drugs, fluoroquinolones, trimethoprim-sulfamethoxazole, and aminoglycosides have been rapidly increasing worldwide (6–12).

Epidemiological studies have shown that more than 50% of community and healthcare-associated UTI and BSI reported globally are caused by ExPEC strains belonging to a limited set of lineages as defined by multilocus sequence typing (MLST) (13, 14). These include sequence types ST10, ST69, ST73, ST95, ST127, ST131, and ST393 (15, 16). Except for ST95, large proportions of these lineages are resistant to antimicrobial agents used to treat UTI or BSI. In particular, ST131, the most commonly reported ExPEC lineage worldwide are frequently multidrug resistant (17–19).

Although ST95 strains belong to a small group of ExPEC lineages that cause more 50% of UTI and BSI cases in most regions of the world (10, 15), its drug-resistance frequency is characteristically lower relative to the other common ExPEC multilocus sequence types (14, 20–23). We previously reported that the carriage of a UTI89-like plasmid called pUTI89* and *fimH* type is associated with the pan-susceptible phenotype of ST95 (21, 22). However, the particular genomic features associated with this drug-susceptible phenotype remain unclear. Here, we generated a whole-genome sequences (WGS) dataset comprised of ST95 isolates from North California collection and Enterobase database, and compared the WGSs of susceptible ST95 strains to those of ST95 strains that carry drug-resistance genes to identify genetic features associated with drug susceptibility. We believe that this analysis may help shed new light on the compensatory fitness advantage of ST95 under antimicrobial drug stress without acquiring drug-resistance genes.

2. Results

2.1 Frequency and type of drug resistance genes among ST95 strains

We performed WGS analysis of 36 ST95 isolates that were previously collected from UTI patients in Northern California (14) and compared the sequences to those of 44 BSI ST95 isolates from Northern California (21) we previously reported, and 1669 ST95 WGSs deposited in the Enterobase database (24). We queried the carriage of recognized drug resistance genes against ResFinder and found that 837 (50%) of 1669 Enterobase ST95 isolates and 50 (62%) of 80 Northern California isolates did not contain any recognized resistance genes. Among the remaining 862 ST95 isolates, 112 recognized drug-resistant genes for 12 different classes of antibiotics were identified (Supplementary Table 2). Of these, 25 genes for 7 different classes of antibiotics were present in 80 clinical ST95 isolates from Northern California (Table 1). These classes of antimicrobial agents included aminoglycosides, beta-lactams, phenicols, trimethoprim, sulfonamide, tetracycline, and macrolides. Genes associated with fosfomycin, lincosamide, colistin, quinolone, and rifampicin resistance were not observed among any of the ST95 isolates from Northern California. In the period when the total annual submission of ST95 isolates to Enterobase was more than 50 (2010-2017), the prevalence of strains carrying drug-resistant genes has remained relatively constant (33%-59%). When stratified by BSI or UTI, 86 (38%) of 225 BSI isolates and 132 (53%) of 248 UTI isolates were drug-resistant based on the ResFinder database.

Among the sequences from the 80 Northern California isolates, we found extended-spectrum beta-lactamase (ESBL) genes $bla_{CTX\text{-}M\text{-}14}$, $bla_{CTX\text{-}M\text{-}15}$, $bla_{TEM\text{-}1a}$, $bla_{TEM\text{-}1b}$, and $bla_{TEM\text{-}1c}$ in 1 (1.3 %), 1 (1.3 %), 3 (3.9 %), 21 (27 %), and 1 (1.3 %) of 80 isolates, respectively. The same set of ESBL genes was present in 29 (1.7 %), 44 (2.6 %), 18 (1.1 %), 335 (20.1 %), and 148 (8.9 %) of 1669 isolates among the sequences deposited in the Enterobase database, respectively. The temporal distribution of drug resistance genes in all the 1749 ST95 isolates is shown in Supplementary Table 3.

2.2 Pan-genome analysis

A total of 35,134 genes constituted the pan-genome of 1749 ST95 WGSs. Of 35,134 genes, 3,588 (10.2%) were shared among more than 95% of the isolates (core genes) and 31,546 (89.5%) were distributed among subsets of the isolates (accessory genes). Of the latter, 29,522 genes were found in <15% of the isolates (cloud genes). Of the cloud genes, 13,970 (40%) genes were unique to only one isolate.

Among the 862 drug-resistant isolates, 28,388 genes constituted the pan-genome. Of 28,388 genes, 3,616 were core genes and 24,722 were accessory genes . Among the 887 drug-susceptible isolates, 26,425 constituted the pan-genome. Of these, 3,645 were core genes and 22,780 were accessory genes.

2.3 Phylogenetic and principal component analyses

Maximum-likelihood phylogenies of the SNP alignment of core genomes and the presence and absence of accessory genomes were obtained from FastTree (25) with the GTR model (Figure 1). Clustering based on *fimH* types was observed but we did not observe a strong correlation of drug resistance with phylogenetic trees based on the SNPs of core genes and the presence and absence of accessory genes.

Principal component analysis (PCA) was also conducted to visually assess variables accounting for the variance. Cluster patterns based on the variance of SNPs and accessory genes were associated with the carriage of plasmid pUTI89* and with *fimH* types 18, 27, 30, 41, and 54 in core genomes and accessory genomes, respectively (Figure 2). However, the cluster patterns were not observed with drug resistance or pan-susceptibility.

2.4 Genes associated with drug-susceptible ST95 strains

First, we conducted a univariate analysis with fisher's exact test of 31,546 accessory genes on the entire collection of 1749 ST95 isolates to identify genes uniquely associated with drug-susceptible strains. Due to the number of tests, Bonferroni adjusted P-value was used to determine a statistical significance cut-off. We found that 884 accessory genes were significantly associated with drug susceptibility, and 1567 genes were significantly associated with drug resistance (Supplementary Table 4). We also conducted a multivariate analysis adjusting for the carriage of pUTI89* and *fimH* types to control for confounding due to their previously reported significant association with drug-susceptibility (21). We found that 154 accessory genes were significantly associated with drug susceptibility, and 441 genes were significantly associated with drug resistance (Supplementary Table 4).

We then conducted a genome-wide analysis with 261 UTI ST95 isolates and 292 BSI ST95 isolates (80 from Northern California and 473 from the Enterobase ST95 collection) based on the metadata information. We limited the analysis to human BSI and UTI isolates (553 isolates in total) to preclude potential bias or errors that could result from including ST95 samples collected from unknown sources or sources other than human extraintestinal infections (e.g., environment, food animals, human non-infection sources) indicated in the Enterobase database. Genome-wide association analysis was done at the level of single nucleotide polymorphisms (SNP) among core genes as determined with SNP-sites (26), and at the level of gene content as determined with Roary (27). A logistic multivariate regression model was conducted to identify genomic features significantly associated with drug susceptibility. In this model, the carriage of pUTI89* and *fimH* types was adjusted to control for confounding (21).

We found 44 accessory genes significantly associated with strains that lacked recognized drug-resistance genes included in the ResFinder database (Table 2). We found 13 SNPs on core genes and 145 accessory genes associated with drug resistance (Supplementary Table 5). The correlation coefficient matrix of these significantly associated genes is shown in Figure 3.

Of the 44 genes associated with drug susceptible strains, 8 were annotated with Prokka as *alkA* (DNA-3-methyladenine glycosylase 2), *cia* (colicin), *dinG*_1 (3'-5' exonuclease), *livK* (ABC transporter substrate-binding protein LivK), *maX_3* (PTS sugar transporter subunit IIA), *mltC* (membrane-bound lytic murein transglycosylase MltC), *traD* (type IV conjugative transfer system coupling protein TraD), and *umuD_2* (translesion error-prone DNA polymerase V autoproteolytic subunit). The other 36 genes were annotated by Prokka as encoding hypothetical proteins of unknown function. The sequences of these 36 hypothetical proteins were subjected to BLAST search. Of these, 30 were annotated in the NCBI database. Eight genes were annotated to encode transfer system or transporter-related proteins and 5 genes were annotated as phage genes. The annotated functions of 44 genes are shown in Table 2.

2.5 Distribution of genes associated with pan-susceptible ST95 strains

Of 44 accessory genes significantly associated with strains lacking antimicrobial resistance genes, 13 were found on pUTI89*, 19 were on non-pUTI89* plasmids, and 15 were located in the chromosome (Table 2). Three genes were found both on pUTI89* plasmid and chromosomes.

We also sought these 44 gene sequences in ExPEC strains belonging to ST131, which are largely resistant to one or more antimicrobial agents. Of these, 29 had a match ranging from 1 to 28 among 7458 *Escherichia coli* O25b:H4-ST131 WGSs (Supplementary Table 7). Thus, 15 genes associated with ST95 strains lacking drug-resistance genes were not found in drug-resistant ST131 genomes. These genes included *cia*, *manX_3*, and 13 hypothetical proteins shown in Table 2. Of these 15 genes, 13 were significantly associated with drug susceptibility among the entire 1749 ST95 collections as well (Table 2). The two genes not associated were *manX_3* and group_13166, which were annotated as PTS sugar transporter subunit IIA and PTS sugar transporter subunit IIB, respectively.

3. Discussion

We identified genomic features associated with strains lacking drug resistance genes of ExPEC ST95 lineage isolated from BSI and UTI patients, as well as other sources. We analyzed 1749 ST95 strains--1669 ST95 from Enterobase (downloaded on March 12, 2020) and 80 ST95 isolates from UTI or BSI patients in Northern California. We also conducted a genome-wide association study with 553 ST95 BSI and UTI human isolates. The sources of other isolates submitted to Enterobase include soil, water, vegetables, animals other than humans, and human samples other than urine or blood (e.g., stool or wound). Our previous study of WGSs of 86 isolates of ST95 from BSI patients in the United States found that the carriage of UTI89 like plasmid (pUTI89*) or plasmid sequences and *fimH6* was significantly associated with drug susceptible strains (21). We also reported that in Northern California, large proportions of ST95 strains remained susceptible to most of the drugs we tested compared to other major pandemic ExPEC lineages despite its high prevalence among patients with community-acquired UTI, (14). We hypothesized that ST95 strains can maintain competitive advantage under antibiotic stress without acquiring drug-resistance genes among the human intestinal microbiota.

Of the 1749 ST95 WGSs, 862 (49%) carried one or more drug-resistance genes. Among these, 94 (36%) of 261 UTI ST95 isolates and 154 (53%) of 292 BSI ST95 isolates carried drug-resistance genes, respectively. We observed a lower prevalence of drug-resistant ST95 isolates in Northern California compared to those in the Enterobase database (38% and 50%, respectively). This may be due to a submission bias to Enterobase. For example, drug-resistant ST95 strains may have been more likely to be sequenced and deposited in Enterobase. The Northern California isolates were population-based samples (14, 22).

Principal component analysis and phylogenetic analysis based on the SNPs of core genes and the presence and absence of accessory genes were conducted to identify potential confounding factors associated with clustering. In both analyses, carriage of pUTI89* and *fimH* types were strongly associated with clustering. This suggests that the carriage of pUTI89* and the *fimH* types explain the majority of genomic variation we extracted from the pangenome analysis of the ST95 WGSs, and that these two factors needed to be controlled in assessing the genetic features associated with the presence or absence of drug-resistant genes in conducting a pan-genome-

wide association analysis. Clustering by *fim*H type is consistent with our previous findings (21), and also offers more evidence for *fimH* typing as a useful method for genotyping ExPEC (28).

We used multivariate logistic models to identify genomic features significantly associated with drug pan-susceptibility while controlling for *fimH* types and the carriage of pUTI89*. We limited the analysis to WGSs of those isolates from well-documented human BSI or UTI cases (553 isolates in total) to preclude potential bias that could result from including ST95 WGSs from unknown sources or non-human and non-human infection sources. We initially found 44 accessory genes associated with drug susceptibility. We did not identify any SNPs in the core genes significantly associated with drug-susceptible strains. Of these 44 genes, 15 were not found in any of 7458 *E. coli* ST131 WGSs deposited in the NCBI nucleotide database. ST131 strains are frequently multidrug resistant (15, 29). The other 29 genes showed 1 to 28 matches with the ST131 genome sequences, but these matches represented less than 1% of the ST131 WGSs, whereas they represented 10-87% of the ST95 WGSs (Table 2).

Thirteen of 15 genes remained associated with ST95 strains lacking drug-resistance genes among the entire collection of WGSs we analyzed (Table 2). The two genes that lacked association among the entire collection of 1749 WGSs were *manX_3* and group_13166, which were annotated as PTS sugar transporter subunit IIA and PTS sugar transporter subunit IIB, respectively.

Annotation of these 15 genes not found in any of the ST131 strains in the NCBI nucleotide database revealed that two genes were annotated as PTS sugar transporter subunit IIA and IIB, respectively. Previous reports revealed that PTS sugar transporters are involved in the penetration of polar antibiotics (30, 31). Three genes were annotated as phage genes. The other genes encoded colicin, malate dehydrogenase, DNA breaking-rejoining protein, surface exclusion protein, and peptidase M14. Five encoded hypothetical proteins, functions unknown.

Acquisition and maintenance of accessory genes by *E. coli* must provide a subpopulation of *E. coli* strains an adaptive advantage in an ecological niche. ST95 strains may have evolved compensatory niche-adaptive mechanisms to remain viable in a variety of ecological niches (human and other animal intestines, environment) despite remaining susceptible to antibiotics. One such mechanism is the role of amino acid transporters reported to contribute to metabolic adaptation of uropathogenic *E. coli* (UPEC) in the urinary tract (32).

A limitation of this study is that the drug resistance phenotype of 1669 ST95 WGS obtained from the Enterobase database was predicted from the presence of drug resistance genes annotated in ResFinder and not by experimental phenotyping tests. Although genotypic resistance has been shown to have 100% sensitivity and specificity in predicting phenotypic resistance in *E. coli* by some studies (33, 34), we cannot rule out the possibility of misclassifying drug resistance in this study. Nevertheless, we found strong associations of multiple accessory genes with the absence of recognized drug-resistance genes in ST95 strains.

Another limitation is the lack of consistent meta-data in the Enterobase database. Although the Enterobase database had data on 1669 ST95 isolates, the majority of them lacked information on the source of ST95 strains, which led to the reduction of sample size for the genome-wide

association study (n=553). Lastly, submission bias of ST95 to Enterobase could have affected the prevalence of drug-resistant strains in our collection. The prevalence of drug-resistant ST95 strains in this collection may over-estimate the actual prevalence of drug-resistant ST95 in the human population.

In conclusion, we identified several genes associated with drug pan-susceptibility in ExPEC ST95 strains, which suggests that the ST95 lineage may have evolved to compensate for its susceptibility to antibiotics by acquiring unique fitness genes that enable a subpopulation of them to survive in multiple environmental niches without having to gain drug resistance. Experimental confirmation of these findings is needed to support this proposal, which is feasible with the relatively small number of unique genes in pan-susceptible ST95 strains we identified.

4. Materials and Methods

4.1 Sample collection in Northern California

As previously described by Yamaji et al. (14) and Adams-Sapper et al. (22), we consecutively collected urine samples from patients with symptoms of UTI attending our university outpatient health service between September 2016 and May 2017, and all BSI isolates from inpatients admitted to San Francisco General Hospital (SFGH) between July 2007 and September 2010. All *E. coli* isolates were genotyped by multilocus sequence typing (MLST) based on the seven-gene scheme described at the PubMLST website (https://pubmlst.org/bigsdb?db=pubmlst_mlst_seqdef) (35). The seven-gene PCR products were sequenced by MiSeq System (Illumina) at the Functional Genomics Laboratory at University of California at Berkeley. All *E. coli* subtyped as ST95 were selected for WGS analysis. We analyzed 36 ST95 isolates from patients with UTI and 44 ST95 isolates from patients with BSI.

4.2 Short read genome sequencing

Genomic DNA was extracted from all strains with the Qiagen blood and tissue DNeasy kit. Library preparation for the MiSeq platform following a standard protocol for Illumina-compatible libraries was conducted at Functional Genomics Laboratory in Berkeley. Final libraries were sequenced via a 300-bp paired-end run on a MiSeq instrument with V3 chemistry and standard Illumina analysis software at The California Institute for Quantitative Biosciences at UC Berkeley (QB3-Berkeley).

4.3 Genome assembly

MiSeq reads were screened and trimmed based on length and quality with BBDUK version 1.0 under the default setting (http://jgi.doe.gov/data-and-tools/bb-tools/). The trimming process also removed residual adapter sequences. Quality check of individual FASTQ files was conducted with FastQC (36) and all of the newly sequenced WGS data passed the quality check. *De novo* assembly of trimmed paired reads for the 80 libraries was performed with Unicycler version 0.4.8 under the setting "--min_fasta_length 500" to remove contigs less than 500bp (37). The quality of *de novo* assembly was checked with QUAST version 5.0.2 under the default setting (38).

4.4 Complementary dataset

We obtained 1669 assembled whole genomes of ST95 and associated metadata from Enterobase (24) accessed on March 12, 2020. Information on source type, source details, collection year,

collection continent, and collection country of these whole genomes of ST95 is shown in Supplementary Table 1.

4.5 Bioinformatics analysis

Using the ResFinder database, we identified antimicrobial drug resistance (AMR) genes among the 1749 assembled genomes of ST95. We defined drug-resistant strains as those having at least one recognized drug resistance gene, and pan-susceptible strains as those not having any recognized drug resistance gene. Plasmid incompatibility group and replicon types were identified with the PlasmidFinder database (https://bitbucket.org/genomicepidemiology/plasmidfinder_db/src/master/, accessed April 2020) (39). *FimH* types were identified with the FimTyper database (https://bitbucket.org/genomicepidemiology/fimtyper/src/master/, accessed April 2020) (40). All identifications were done with Abricate version 1.0.1(41) with a 95% identity threshold across the reference sequences.

Annotation was performed on all the assembled genomes with Prokka version 1.14.0 under the default setting (42). Genes annotated as encoding hypothetical proteins were further analyzed on BLAST (accessed April 2021) (43). A pangenome of the entire data set was constructed with Roary version 3.13.0 (27) with a 95% identity cutoff. Here, genes present in 95% or more of the cohort isolates were defined as core and constituting the core portion of the pangenome, and genes present in less than 95% were defined as accessory and constituting the accessory portion of the pangenome. Accessory genes present in <15% of the cohort isolates were defined as cloud genes. A concatenated core CDS alignment ("core_gene_alignment.aln") was obtained from the Roary default output, and we extracted single nucleotide polymorphism (SNP) information with snp-sites with the default option (26). Phylogenetic trees were constructed with FastTree (25) with maximum likelihood method with the GTR model based on the SNP alignment generated with snp-sites and the presence and absence of accessory genes from the Roary default output ("accessory_binary_genes.fa"). Visualization was done with iToL version 6.1.2 (http://itol.embl.de).

4.6 Statistical analysis

All statistical analyses were conducted with R version 3.6.3 (44). A P-value of <0.05 was considered statistically significant. We used the prcomp function for principal component analysis. The association of a specific accessory gene with drug resistance was first tested with fisher's exact test without adjusting for any potential confounding factors with a cut-off at Bonferroni adjustment P-value, and then tested with multivariate analysis with the glm function adjusted for the carriage of UTI89 like plasmid and *fimH* types. For the regression analysis, 553 ST95 isolates from BSI and UTI patients (473 from Enterobase and 80 from Northern California) were used to remove potential bias or errors that could occur from including ST95 samples collected from non-human or non-disease sources. Relationships between each set of significantly associated genetic features were assessed with correlation coefficients. All graphs were plotted with ggplot2 (45).

4.7 Distribution of genes associated with pan-susceptibility

Once we identified ExPEC ST95 genes significantly associated with susceptibility to antimicrobial drugs, we analyzed the distribution and location of the gene in other ExPEC STs by BLAST (accessed June 2021) (43). We targeted ST131 because of their high frequency of

drug resistance and the number of WGSs in the database (18, 46). We sought for these ST95 genes in (1) the ST131 chromosome, plasmid, or phages and (2) ST131 whole genome sequences deposited in the NCBI database. NCBI has 127 complete whole genome sequences of *Escherichia coli* O25b:H4-ST131 and 7331 plasmid or short genome sequences of *Escherichia coli* O25b:H4-ST131 deposited.

4.8 Accession numbers

The whole-genome shotgun sequence results described here have been deposited in DDBJ/ENA/GenBank under the accession numbers shown in Supplementary Table 6. Genome sequences have been deposited in NCBI BioProject under accession number PRJNA763994.

Acknowledgement

We thank Holly Berryman Stern and the other staff members of the Tang Center of University of California, Berkeley (UC Berkeley)-affiliated health care service for their time and support of this project. This work was supported by Centers for Disease Control and Prevention investments to combat antibiotic resistance under BAA 200-2016-91939. The contents of this work are solely the responsibility of the authors and do not necessarily represent the official views of the Centers for Disease Control and Prevention. The funder of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report. The corresponding author had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Figures

Figure 1

Maximum-likelihood phylogeny of 1749 ST95 isolates constructed with FastTree and visualized with iTOL

A: Phylogeny based on the core genome.

B: Phylogeny based on the presence and absence of accessory genes

Figure 2

Principle component analysis of 1749 ST95 isolates based on (A) core genome and (B) presence and absence of accessory genome

Figure 3

The correlation coefficient matrix of ST95 genes significantly associated with ST95 strains lacking drug-resistance genes. This illustrates which groups of genes associated with drug susceptibility or drug resistance are linked.

Tables

Table 1

Frequency of antimicrobial resistance genes among *E. coli* isolates belonging to ST95 registered in Enterobase database and obtained in Northern California respectively, identified from ResFinder.

Table 2

Genes significantly associated with ST95 strains lacking drug-resistance genes. Annotation, location, and odds ratio with 95% confidence interval are shown for UTI and BSI isolates. These ST95 genes that had a matching sequence in *Escherichia coli* O25b:H4-ST131 WGSs from NCBI database are shown. Prevalence of these genes among 1749 ST95 WGSs and their association with susceptible strains are also shown.

Supplementary Table 1
Information on source type, source details, collection year, collection continent, and collection country of whole genomes of ST95 downloaded from Enterobase.

Supplementary Table 2
Drug resistance genes identified from the whole collection of 1749 ST95 from Resfinder.

Supplementary Table 3
Frequency of drug resistance among ST95 isolates registered in the Enterobase database between 1947 and 2019 and ST95 isolates from Northern California between 2013 and 2017, by year, stratified by isolation information (BSI or UTI). This illustrates the yearly distribution of drug resistance genes.

Supplementary Table 4
Univariate and multivariate analysis of the pangenome of 1749 ST95 whole genome sequences

Supplementary Table 5
Information of 13 SNPs on core genes and 145 accessory genes associated with drug resistance.

Supplementary Table 6
The accession numbers of submitted ST95 whole genome sequence on DDBJ/ENA/GenBank.

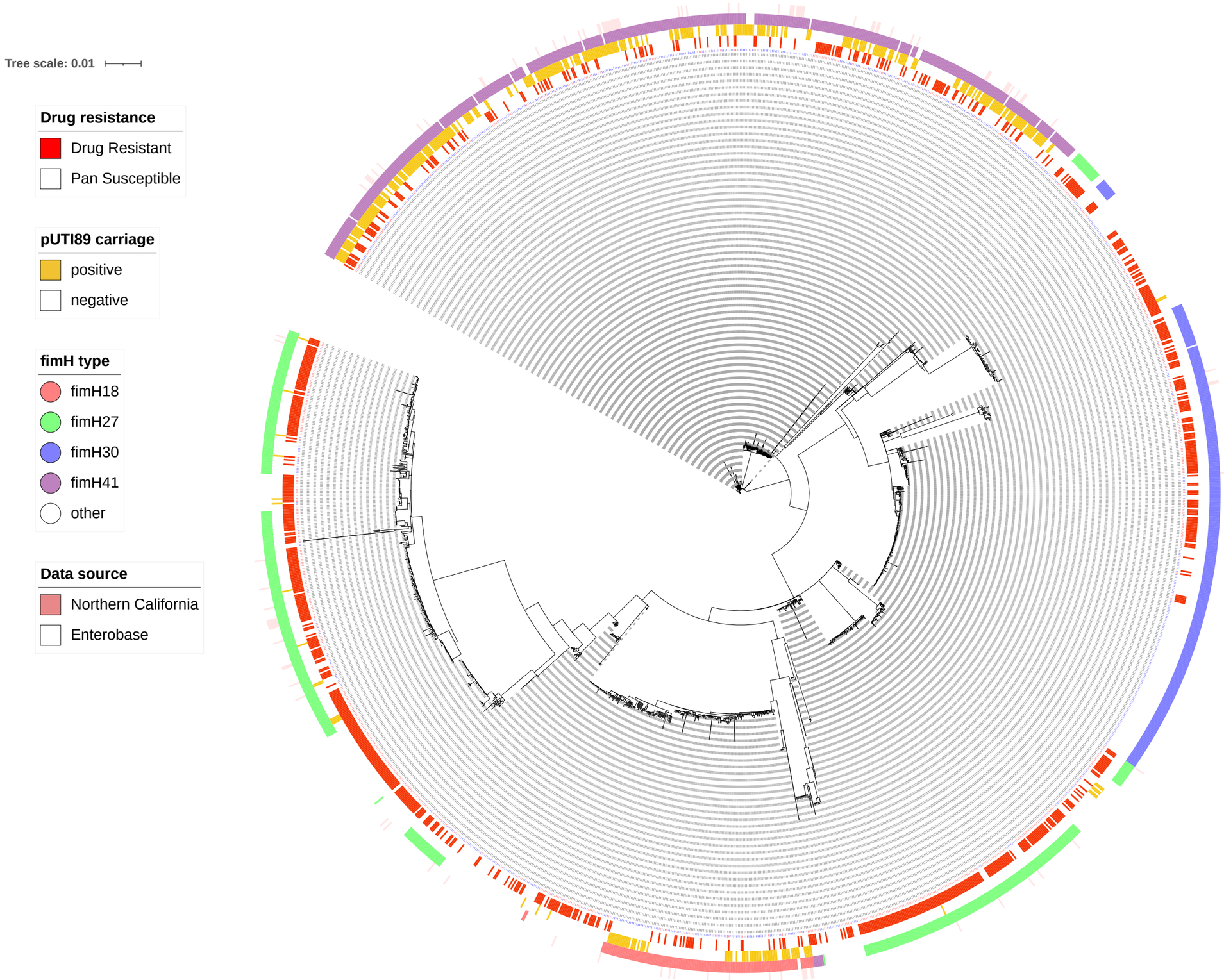
Supplementary Table 7
NCBI accession number and the number of matching regions of ST131 whole genome sequence that carries one or more genes of 44 ST95 genes found to be associated with ST95 strains lacking drug-resistance genes.

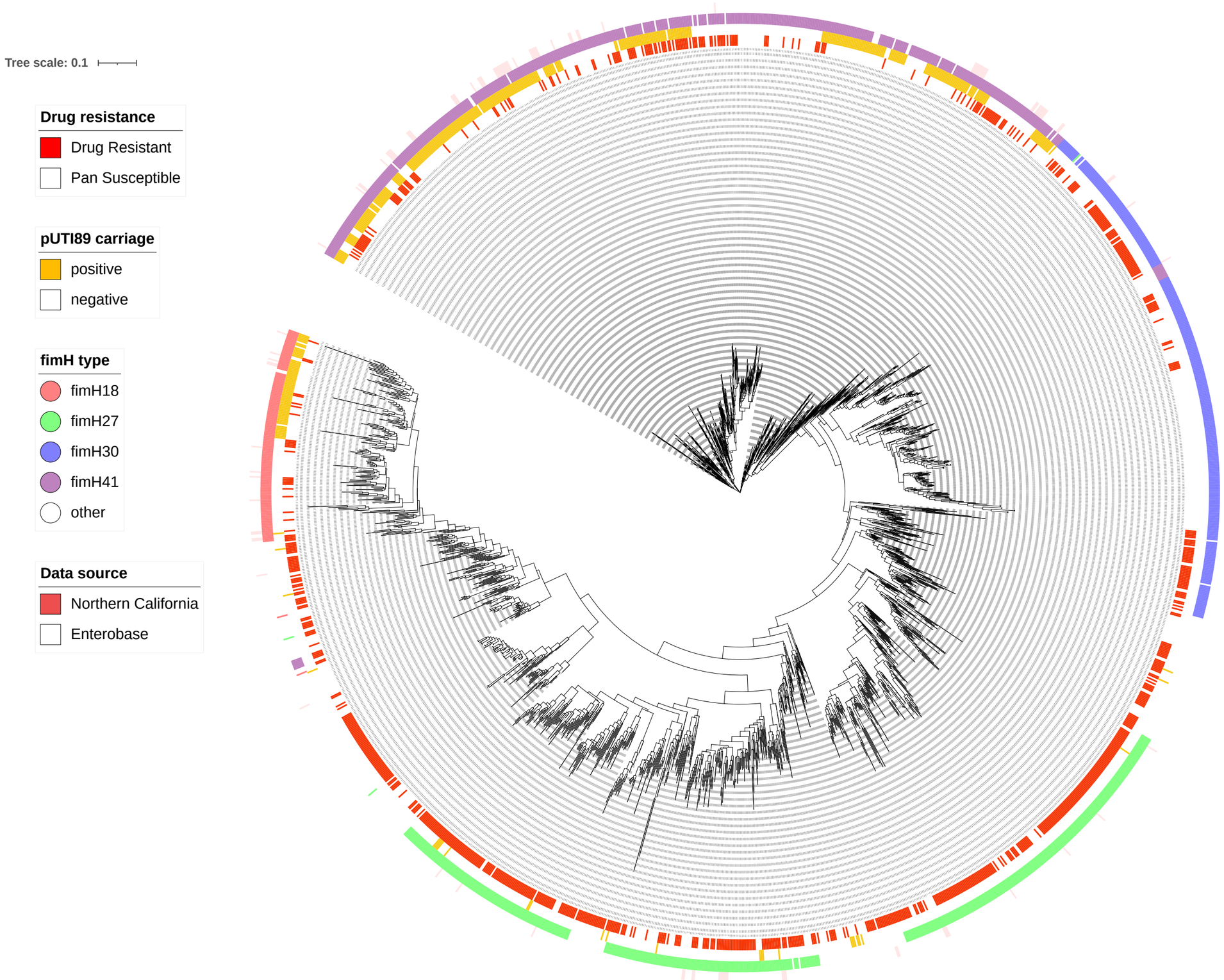
References

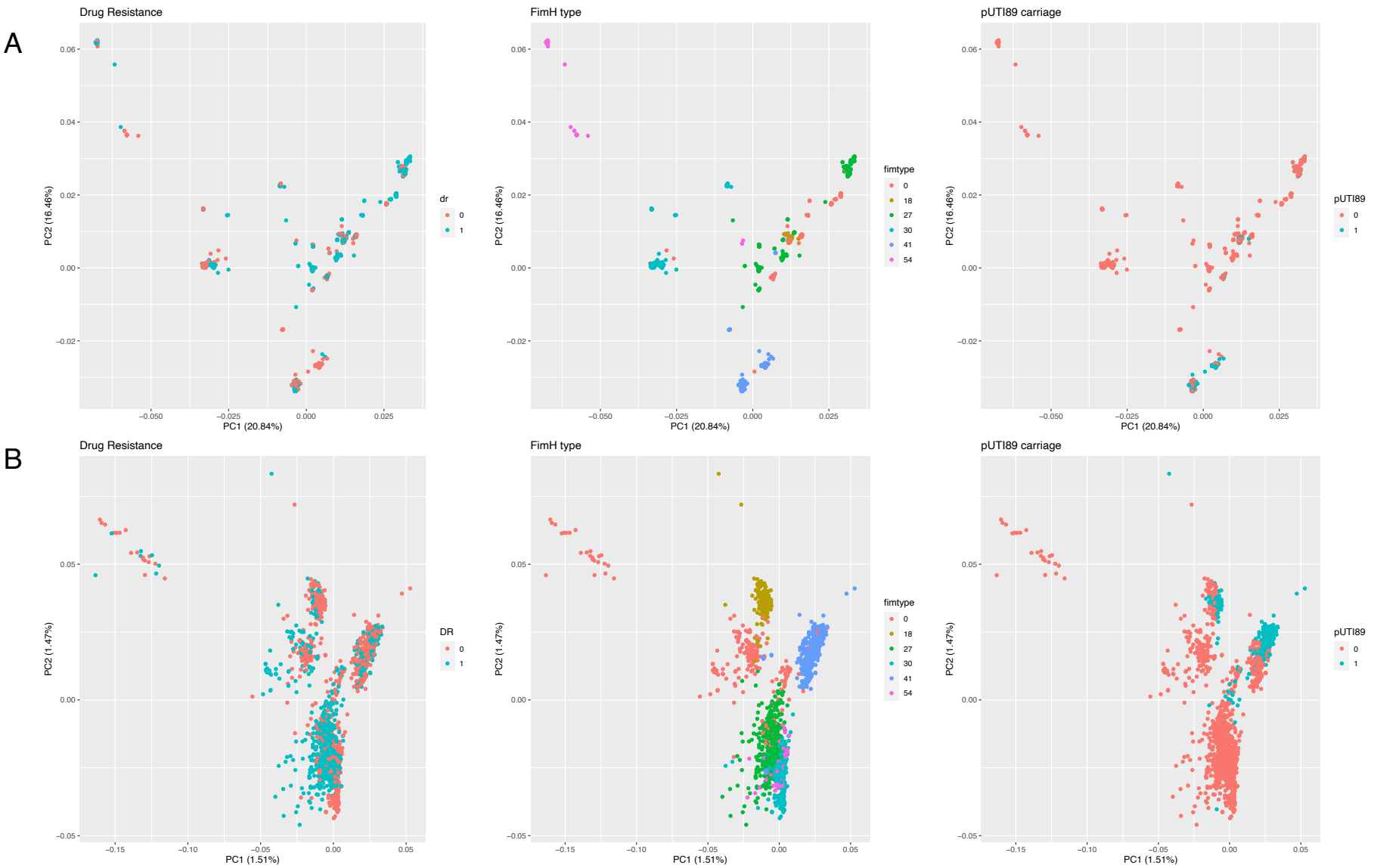
1. World Health Organization (WHO). 2015. Global Action Plan on Antimicrobial Resistance https://doi.org/10.1128/microbe.10.354.1.
2. World Health Organization (WHO). 2020. Global Antimicrobial Resistance and Use Surveillance System (GLASS) Report.
3. Centers for Disease Control and Prevention. 2019. Antibiotic resistance threats in the United States.
4. WHO (World Health Organisation). 2014. ANTIMICROBIAL RESISTANCE Global Report on Surveillance.
5. Johnson JR, Russo TA. 2002. Extraintestinal pathogenic Escherichia coli: “The other bad E coli.” J Lab Clin Med 139:155–162.
6. Thaden JT, Jr VGF, Sexton DJ, Anderson DJ. 2016. Increasing Incidence of Extended-Spectrum β-Lactamase- Producing Escherichia coli in Community Hospitals throughout the Southeastern United States. Infect Control Hosp Epidemiol 37:49–54.
7. Allocati N, Masulli M, Alexeyev MF, Di Ilio C. 2013. Escherichia coli in Europe: An overview. Int J Environ Res Public Health 10:6235–6254.
8. Yassin AK, Gong J, Kelly P, Lu G, Guardabassi L, Wei L, Han X, Qiu H, Price S, Cheng D, Wang C. 2017. Antimicrobial resistance in clinical Escherichia coli isolates from poultry and livestock, China. PLoS One 12:1–8.
9. Borges CA, Tarlton NJ, Riley LW. 2019. Escherichia coli from Commercial Broiler and Backyard Chickens Share Sequence Types, Antimicrobial Resistance Profiles, and Resistance Genes with Human Extraintestinal Pathogenic Escherichia coli. Foodborne Pathog Dis 16:813–822.
10. Manges AR, Geum HM, Guo A, Edens TJ, Fibke CD, Pitout JDD. 2019. Global Extraintestinal Pathogenic Escherichia coli (ExPEC) Lineages 32:1–25.
11. Hu Y, Matsui Y, Riley LW. 2020. Risk factors for fecal carriage of drug-resistant Escherichia coli□: a systematic review and meta-analysis. Antimicrob Resist Infect Control 3:1–12.
12. Matsumura Y, Noguchi T, Tanaka M, Kanahashi T, Yamamoto M, Nagao M, Takakura S, Ichiyama S, Yukimasa O, Higashiyama T, Sasano M, Watanabe H, Shimizu T, Hayashi A, Higuchi T, Fujita N, Komori T, Yamada Y, Ueda K, Yamamoto G, Moro K, Yabu S. 2017. Population structure of Japanese extraintestinal pathogenic Escherichia coli and its relationship with antimicrobial resistance. J Antimicrob Chemother 72:1040–1049.
13. Matsui Y, Hu Y, Rubin J, de Assis RS, Suh J, Riley LW. 2020. Multilocus sequence typing of Escherichia coli isolates from urinary tract infection patients and from fecal samples of healthy subjects in a college community. Microbiologyopen 9:1225–1233.
14. Yamaji R, Rubin J, Thys E, Friedman CR, Riley LW. 2018. Persistent pandemic lineages of uropathogenic Escherichia coli in a college community from 1999 to 2017. J Clin Microbiol 56:1–10.
15. Riley LW. 2014. Pandemic lineages of extraintestinal pathogenic Escherichia coli. Clin Microbiol Infec 20:380–390.
16. Manges AR, Geum HM, Guo A, Edens TJ, Fibke CD, Pitout JDD. 2019. Global Extraintestinal Pathogenic Escherichia coli (ExPEC) Lineages. Clin Microbiol Rev 32:e00135-18.

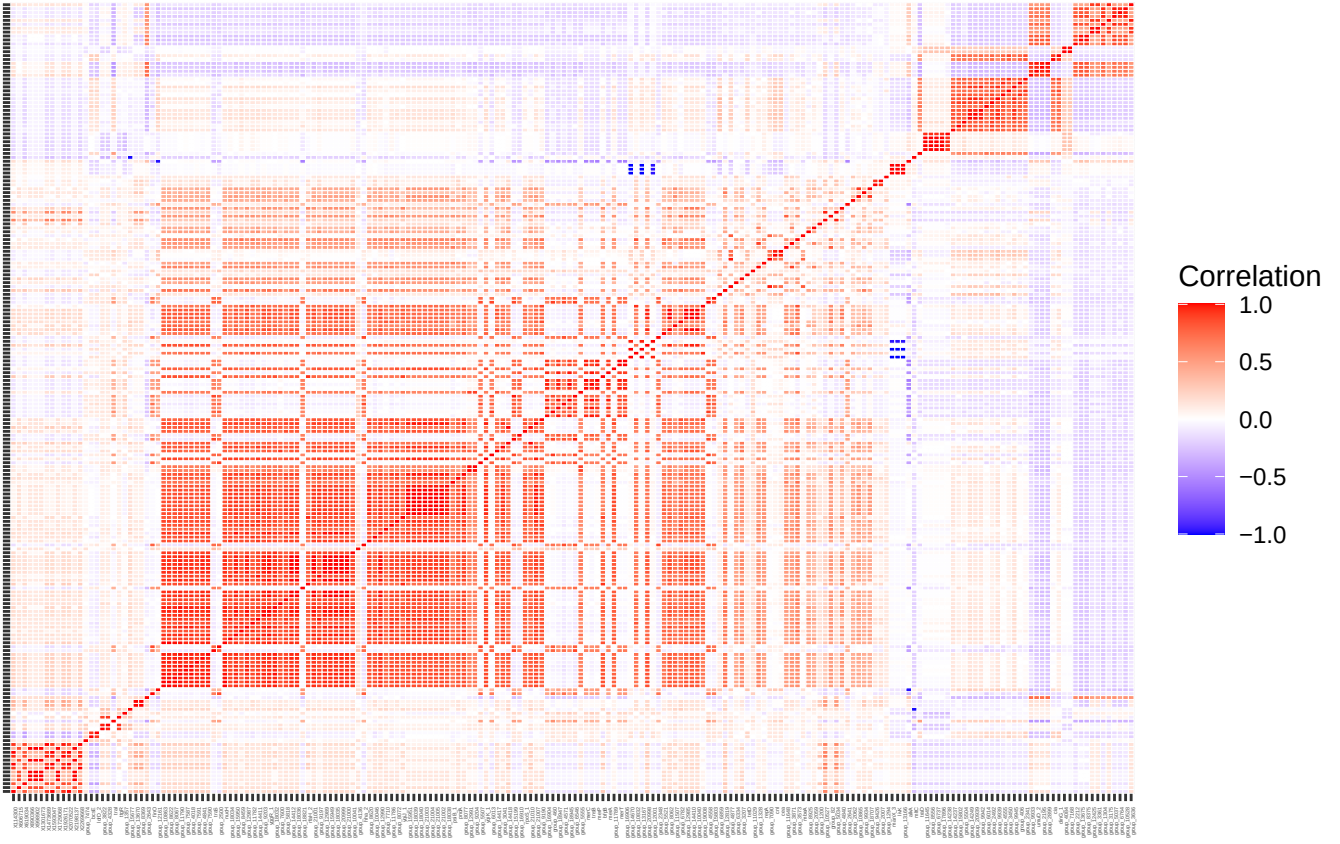
17. Li D, Reid CJ, Kudinha T, Jarocki VM, Djordjevic SP. 2020. Genomic analysis of trimethoprim-resistant extraintestinal pathogenic Escherichia coli and recurrent urinary tract infections. Microb Genomics 6:1–19.
18. Ludden C, Decano AG, Jamrozy D, Pickard D, Morris D, Parkhill J, Peacock SJ, Cormican M, Downing T. 2020. Genomic surveillance of escherichia coli ST131 identifies local expansion and serial replacement of subclones. Microb Genomics 6.
19. Kallonen T, Brodrick HJ, Harris SR, Corander J, Brown NM, Martin V, Peacock SJ, Parkhill J. 2017. Systematic longitudinal survey of invasive Escherichia coli in England demonstrates a stable population structure only transiently disturbed by the emergence of ST131. Genome Res 27:1437–1449.
20. McNally A, Kallonen T, Connor C, Abudahab K, Aanensen DM, Horner C, Peacock SJ, Parkhill J, Croucher NJ, Corander J. 2019. Diversification of colonization factors in a multidrug-resistant escherichia coli lineage evolving under negative frequency- dependent selection. MBio 10:1–19.
21. Stephens CM, Adams-sapper S, Sekhon M, Johnson JR, Riley LW. 2017. Genomic Analysis of Factors Associated with Low Prevalence of Antibiotic Resistance in Extraintestinal Pathogenic Escherichia coli Sequence Type 95 Strains. mSphere 2:e00390-16.
22. Adams-Sapper S, Diep BA, Perdreau-Remington F, Riley LW. 2013. Clonal composition and community clustering of drug-susceptible and -resistant Escherichia coli isolates from bloodstream infections. Antimicrob Agents Chemother 57:490–497.
23. Siu HL, Reddy S, Cheesbrough J, Bolton FJ, Willshaw G, Cheasty T, Fox AJ, Upton M. 2008. Major uropathogenic Escherichia coli strain isolated in the Northwest of England identified by multilocus sequence typing. J Clin Microbiol 46:1076–1081.
24. Zhou Z, Alikhan NF, Mohamed K, Fan Y, Achtman M. 2020. The EnteroBase user's guide, with case studies on Salmonella transmissions, Yersinia pestis phylogeny, and Escherichia core genomic diversity. Genome Res 30:138–152.
25. Price MN, Dehal PS, Arkin AP. 2010. FastTree 2 - Approximately maximum-likelihood trees for large alignments. PLoS One 5.
26. Page AJ, Taylor B, Delaney AJ, Soares J, Seemann T, Keane JA, Harris SR. 2016. SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. Microb genomics 2:e000056.
27. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, Fookes M, Falush D, Keane JA, Parkhill J. 2015. Roary: Rapid large-scale prokaryote pan genome analysis. Bioinformatics 31:3691–3693.
28. Dias RCS, Moreira BM, Riley LW. 2010. Use of fimH single-nucleotide polymorphisms for strain typing of clinical isolates of Escherichia coli for epidemiologic investigation. J Clin Microbiol 48:483–488.
29. Johnson JR, Johnston B, Clabots C, Kuskowski MA, Castanheira M. 2010. Escherichia coli Sequence Type ST131 as the Major Cause of Serious Multidrug□Resistant E. coli Infections in the United States . Clin Infect Dis 51:286–294.
30. Jeckelmann JM, Erni B. 2020. Transporters of glucose and other carbohydrates in bacteria. Pflugers Arch Eur J Physiol 472:1129–1153.
31. Garcia De Gonzalo C V., Denham EL, Mars RAT, Stülke J, Van Der Donk WA, Van Dijl JM. 2015. The phosphoenolpyruvate: Sugar phosphotransferase system is involved in

sensitivity to the glucosylated bacteriocin sublancin. Antimicrob Agents Chemother 59:6844–6854.
32. Mann R, Mediati DG, Duggin IG, Harry EJ, Bottomley AL. 2017. Metabolic adaptations of Uropathogenic E. coli in the urinary tract. Front Cell Infect Microbiol 7.
33. Su M, Satola SW, Read TD. 2019. Genome-based prediction of bacterial antibiotic resistance. J Clin Microbiol 57.
34. Zankari E, Hasman H, Kaas RS, Seyfarth AM, Agersø Y, Lund O, Larsen MV, Aarestrup FM. 2013. Genotyping using whole-genome sequencing is a realistic alternative to surveillance based on phenotypic antimicrobial susceptibility testing. J Antimicrob Chemother 68:771–777.
35. Tartof SY, Solberg OD, Manges AR, Riley LW. 2005. Analysis of a Uropathogenic Escherichia coli Clonal Group by Multilocus Sequence Typing. J Clin Microbiol 43:5860–5864.
36. Andrews S. 2010. FastQC: a quality control tool for high throughput sequence data.
37. Wick RR, Judd LM, Gorrie CL, Holt KE. 2017. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. PLoS Comput Biol 13:1–22.
38. Mikheenko A, Prjibelski A, Saveliev V, Antipov D, Gurevich A. 2018. Versatile genome assembly evaluation with QUAST-LG. Bioinformatics 34:i142–i150.
39. Carattoli A, Zankari E, Garci??-Fern??ndez A, Larsen MV, Lund O, Villa L, Aarestrup FM, Hasman H. 2014. In Silico detection and typing of plasmids using plasmidfinder and plasmid multilocus sequence typing. Antimicrob Agents Chemother 58:3895–3903.
40. Roer L, Tchesnokova V, Allesøe R, Muradova M, Chattopadhyay S, Ahrenfeldt J, Thomsen MCF, Lund O, Hansen F, Hammerum AM, Sokurenko E, Hasmana H. 2017. Development of a Web Tool for Escherichia coli Subtyping Based on fimH Alleles. J Clin Microbiol 55:2538–2543.
41. Seemann T. 2014. ABRicate.
42. Seemann T. 2014. Prokka: Rapid prokaryotic genome annotation. Bioinformatics 30:2068–2069.
43. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: Architecture and applications. BMC Bioinformatics 10:1–9.
44. R Core Team. 2018. R: A Language and Environment for Statistical Computing. Vienna, Austria.
45. Wickham H. 2016. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.
46. Totsika M, Beatson SA, Sarkar S, Phan MD, Petty NK, Bachmann N, Szubert M, Sidjabat HE, Paterson DL, Upton M, Schembri MA. 2011. Insights into a multidrug resistant Escherichia coli pathogen of the globally disseminated ST131 lineage: Genome analysis and virulence mechanisms. PLoS One 6.


Tree scale: 0.01
Drug resistance
Drug Resistant
Pan Susceptible
pUTI89 carriage
positive
negative
fimH type
fimH18
fimH27
fimH30
fimH41
other
Data source
Northern California
Enterobase


Tree scale: 0.1
Drug resistance
Drug Resistant
Pan Susceptible
pUTI89 carriage
positive
negative
fimH type
fimH18
fimH27
fimH30
fimH41
other
Data source
Northern California
Enterobase


A
Drug Resistance
FimH type
pUTI89 carriage
PC2 (16.48%)
PC1 (20.84%)
dr
fimtype
pUTI89
0
1
18
27
30
41
54
B
PC2 (1.47%)
PC1 (1.51%)
DR


Correlation
1.0
0.5
0.0
−0.5
−1.0

Antibiotics class	Resistance gene name	Enterobase Total (%)	North California Total (%)
Pan susceptible		837 (50)	50 (62)
Aminoglycoside	aac(3)-IId_1	36 (2.16)	2 (2.56)
	aadA2_1	39 (2.34)	2 (2.56)
	aadA5_1	33 (1.98)	4 (5.13)
	ant(2")-Ia_1	2 (0.12)	2 (2.56)
	ant(3")-Ia_1	38 (2.28)	1 (1.28)
	aph(3")-Ib_5	306 (18.33)	6 (7.69)
	aph(6)-Id_1	254 (15.22)	7 (8.97)
Beta-lactam	blaCTX-M-14_1	29 (1.74)	1 (1.28)
	blaCTX-M-15_1	44 (2.64)	1 (1.28)
	blaTEM-1A_1	18 (1.08)	3 (3.85)
	blaTEM-1B_1	335 (20.07)	21 (26.92)
	blaTEM-1C_1	148 (8.87)	1 (1.28)
Phenicol	catA1_1	16 (0.96)	2 (2.56)
	floR_2	17 (1.02)	1 (1.28)
Trimethoprim	dfrA12_8	33 (1.98)	2 (2.56)
	dfrA17_1	63 (3.77)	5 (6.41)
	dfrA5_1	153 (9.17)	2 (2.56)
Sulphonamide	sul1_2	1 (0.06)	1 (1.28)
	sul1_5	107 (6.41)	5 (6.41)
	sul2_2	278 (16.66)	4 (5.13)
	sul2_3	114 (6.83)	3 (3.85)
Tetracycline	tet(A)_6	211 (12.64)	8 (10.26)
	tet(B)_1	1 (0.06)	1 (1.28)
	tet(B)_2	107 (6.41)	5 (6.41)
Macrolide	mph(A)_2	42 (2.52)	3 (3.85)

Gene/SNP	Annotation	location	UTI (n=261) OR [95% CI]	P-value	BSI (n=292) OR [95% CI]	P-value	Entire ST95 (n=1749) OR [95% CI]	P-value	Frequency (%)	ST131
alkA	DNA-3-methyladenine glycosylase 2	chromosome	5.5 [1.7-20.4]	0.007	2.5 [1-6.6]	0.046	5.6 [4.2-7.5]	0	1420 (80)	5
cia	colicin	plasmid	4.8 [1.8-13.9]	0.003	2.6 [1.1-6.2]	0.023	2 [1.5-2.6]	0	318 (18)	0
dinG_1	3'-5' exonuclease	plasmid	5.3 [2-15.4]	0.001	4.7 [1.9-12.3]	0.001	2.3 [1.7-3]	0	311 (18)	1
livK	leucine-specific binding protein precursor, high-affinity branched-chain amino acid ABC transporter substrate-binding protein LivK	chromosome	3.8 [1.2-13.5]	0.03	4 [1.5-12]	0.008	0.9 [0.7-1.2]	0.357	1546 (87)	5
manX_3	PTS sugar transporter subunit IIA	chromosome	3.8 [1.2-13.5]	0.03	3.9 [1.5-11.8]	0.009	0.9 [0.6-1.2]	0.320	1547 (87)	0
mltC	membrane-bound lytic murein transglycosylase MltC	chromosome	2.1 [1-4.4]	0.042	2.1 [1-4.4]	0.043	NA [NA-36.1]	0.481	1336 (75)	1
traD	type IV conjugative transfer system coupling protein TraD	plasmid	4.1 [1.9-9.2]	0	2.2 [1.2-4.3]	0.014	NA [NA-36.1]	0.481	760 (43)	1
umuD_2	translesion error-prone DNA polymerase V autoproteolytic subunit	pUTI89	11.2 [3.6-38.5]	0	4.2 [1.6-10.9]	0.003	5.3 [4-7.1]	0	371 (21)	1
group_406	conjugal transfer mating pair stabilization protein TraG	plasmid	6.1 [2.5-16.1]	0	4.5 [2.2-9.7]	0	2.1 [1.6-2.7]	0	379 (21)	22
group_1304	hypothetical protein	pUTI89	43.5 [7.9-1000]	0	9.7 [2.5-62.5]	0.004	11.1 [6.7-19.4]	0	196 (11)	28
group_2195	DUF4113 domain-containing protein	plasmid	12 [4-40]	0	6.7 [1.9-24.4]	0.003	7.6 [5.5-10.5]	0	462 (26)	0
group_2212	DNA repair protein UmuC	pUTI89	18.9 [5.5-90.9]	0	5.8 [2-19.6]	0.002	7 [4.8-10.3]	0	264 (15)	2
group_2341	incFII family plasmid replication initiator RepA	plasmid	3.9 [1.6-10.2]	0.004	3.4 [1.5-7.5]	0.003	2 [1.6-2.5]	0	377 (21)	25
group_2492	conjugal transfer pilus acetylase TraX	plasmid	4.9 [2-12.7]	0.001	6.8 [3.3-15.2]	0	1.9 [1.5-2.4]	0	423 (24)	25
group_2899	phage integrase family protein	pUTI89	21.7 [6.2-83.3]	0	6.7 [1.9-24.4]	0.003	8 [5.8-11.3]	0	352 (20)	28
group_3175	DUF945 domain-containing protein	pUTI89	6.6 [1.7-43.5]	0.017	4.1 [1.1-19.6]	0.048	5.1 [3.5-7.8]	0	191 (11)	23
group_3361	transglycosylase SLT domain-containing protein	pUTI89	6.6 [1.9-30.3]	0.006	3.3 [1.1-11.2]	0.044	5.2 [3.6-7.7]	0	218 (12)	23
group_3450	fertility inhibition protein FinO	plasmid	3.4 [1.4-8.8]	0.009	3 [1.4-6.4]	0.005	2 [1.5-2.5]	0	384 (22)	25
group_3636	type IV conjugative transfer system protein TraL	pUTI89	6.2 [1.6-41.7]	0.022	5.6 [1.3-38.5]	0.036	7.6 [4.7-12.7]	0	169 (10)	22
group_4084	phage tail protein	chromosome	6.3 [1.2-50]	0.041	4.2 [1.4-13.3]	0.012	1.7 [1.3-2.1]	0	311 (18)	1
group_4555	TrbJ conjugal transfer protein	plasmid	6.4 [2.5-17.5]	0	4.5 [2.1-9.7]	0	2 [1.6-2.6]	0	395 (22)	12
group_5037	IS21 family transposase	pUTI89	23.3 [4.3-500]	0.003	9.3 [2.3-62.5]	0.005	13.2 [7.7-24.3]	0	191 (11)	5
group_5504	nuclease-like protein	plasmid	4.8 [2-12.2]	0.001	4.7 [2.4-9.6]	0	1.8 [1.4-2.2]	0	495 (28)	9
group_6014	IncF conjugal transfer surface exclusion protein TraT	plasmid	4 [1.7-9.4]	0.001	4.5 [2.2-9.3]	0	1.9 [1.5-2.4]	0	446 (25)	23
group_6737	phage minor tail protein G	chromosome	2.6 [1.2-5.9]	0.015	2.5 [1.2-5.3]	0.018	1.3 [1.1-1.6]	0.010	795 (45)	0
group_6764	hypothetical protein	plasmid	4.8 [1.6-17.5]	0.008	4 [1.3-15.6]	0.027	9.8 [6-16.8]	0	187 (11)	28
group_7168	tail fiber domain-containing protein	chromosome	6.3 [1.2-50]	0.041	4.3 [1.4-15.2]	0.016	1.8 [1.4-2.3]	0	305 (17)	1
group_8375	ISL3-like element ISEc53 family transposase	pUTI89, chromosome	16.4 [4.1-111.1]	0	8.9 [2.5-41.7]	0.002	11.6 [7.3-19]	0	238 (13)	4
group_8556	phage portal protein	chromosome	2.6 [1.2-5.9]	0.015	2.5 [1.2-5.4]	0.014	1.3 [1.1-1.6]	0.010	603 (34)	0
group_9933	DinI-like family protein	pUTI89, chromosome	10.8 [3.3-38.5]	0	4.8 [1.7-14.1]	0.003	5.4 [4-7.2]	0	373 (21)	1
group_9939	malate dehydrogenase	plasmid	3 [1.3-7.2]	0.011	4.2 [2-9.3]	0	1.5 [1.2-1.9]	0	400 (23)	0
group_9942	hypothetical protein	plasmid	4.1 [1.8-10]	0.001	3.8 [1.9-7.8]	0	1.5 [1.2-1.8]	0	459 (26)	0
group_9945	hypothetical protein	plasmid	4.1 [1.8-10.2]	0.001	3.3 [1.5-7.5]	0.003	1.3 [1-1.6]	0.032	382 (22)	0
group_10528	ATP-binding protein	pUTI89	19.6 [3.7-333.3]	0.005	16.1 [3-333.3]	0.009	16 [8.8-31.8]	0	183 (10)	1
group_11645	DNA breaking-rejoining protein	chromosome	2.6 [1.2-5.9]	0.015	2.5 [1.2-5.3]	0.018	1.3 [1.1-1.6]	0.009	610 (34)	0
group_12425	PAS domain-containing protein	pUTI89	4.5 [1.4-17.2]	0.016	9.2 [2.2-62.5]	0.007	4.7 [3.3-6.8]	0	226 (13)	7
group_13166	PTS sugar transporter subunit IIB	chromosome	3.8 [1.2-13.5]	0.03	3.7 [1.5-10.5]	0.009	0.9 [0.6-1.2]	0.322	1545 (87)	0
group_13675	ClbS/DfsB family four-helix bundle protein	pUTI89, chromo	10 [2.9-47.6]	0.001	9.3 [2.7-43.5]	0.001	11.5 [7.4-18.8]	0	246 (14)	4
group_14237	surface exclusion protein	plasmid	3.1 [1.4-7]	0.005	4.9 [2.5-10.1]	0	1.7 [1.4-2.1]	0	528 (30)	0
group_14238	phage tail assembly protein T	chromosome	2.6 [1.2-6]	0.017	2.3 [1.1-5.2]	0.033	1.4 [1.1-1.7]	0.003	542 (31)	0
group_14249	hypothetical protein	plasmid	3.7 [1.6-8.8]	0.002	4.5 [2.2-9.2]	0	1.6 [1.3-2]	0	485 (27)	0
group_15802	replication regulatory protein RepA	plasmid	3.2 [1.4-7.7]	0.009	5.2 [2.6-10.8]	0	1.5 [1.2-1.9]	0	522 (29)	4
group_17896	DUF2190 family protein	chromosome	2.6 [1.2-5.9]	0.015	2.5 [1.2-5.4]	0.014	1.3 [1.1-1.6]	0.009	597 (34)	0
group_20939	peptidase M14	plasmid	3.2 [1.4-7.7]	0.006	4.5 [2.2-9.2]	0	1.6 [1.3-2]	0	485 (27)	0