# Evolution of extended-spectrum β-lactamase-producing ST131 *Escherichia coli* at a single hospital over 15 years

Shu-Ting Cho[1], Emma G. Mills[1], Marissa P. Griffith[1,2], Hayley R. Nordstrom[1], Christi L. McElheny[1], Lee H. Harrison[1,2], Yohei Doi[1], Daria Van Tyne[1,3]*

[1]Division of Infectious Diseases, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA

[2]Microbial Genomic Epidemiology Laboratory, Center for Genomic Epidemiology, University of Pittsburgh, Pittsburgh, PA, USA

[3]Center for Evolutionary Biology and Medicine, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA

*Corresponding author: Daria Van Tyne, vantyne@pitt.edu

**Keywords**: *Escherichia coli*, ST131, antimicrobial resistance, comparative genomics, mobile genetic elements

**Abstract**

*Escherichia coli* belonging to sequence type ST131 constitute a globally distributed pandemic lineage that causes multidrug-resistant extra-intestinal infections. ST131 *E. coli* frequently produce extended-spectrum β-lactamases (ESBLs), which confer resistance to many β-lactam antibiotics and make infections difficult to treat. We sequenced the genomes of 154 ESBL-producing *E. coli* clinical isolates belonging to the ST131 lineage from patients at the University of Pittsburgh Medical Center (UPMC) between 2004 and 2018. Isolates belonged to the well described ST131 clades A (8%), B (3%), C1 (33%), and C2 (54%). An additional four isolates belonged to another distinct subclade within clade C and encoded genomic characteristics that have not been previously described. Time-dated phylogenetic analysis estimated that the most recent common ancestor (MRCA) for all clade C isolates from UPMC emerged around 1989, consistent with previous studies. We identified multiple genes potentially under selection in clade C, including the cell wall assembly gene *ftsI*, the LPS biosynthesis gene *arnC*, and the yersiniabactin uptake receptor *fyuA.* Diverse ESBL genes belonging to the $bla_{CTX\text{-}M}$, $bla_{SHV}$, and $bla_{TEM}$ families were identified; these genes were found at varying numbers of loci and in variable numbers of copies across isolates. Analysis of ESBL flanking regions revealed diverse mobile elements that varied by ESBL type. Overall, our findings show that ST131 subclades C1 and C2 dominated and were stably maintained among patients in the same hospital and uncover possible signals of ongoing adaptation within the clade C ST131 lineage.

## Introduction

*Escherichia coli* sequence type (ST) 131 is a globally distributed extra-intestinal pathogenic *E. coli* (ExPEC) lineage that causes bloodstream and urinary tract infections[1]. ST131 isolates commonly exhibit multidrug resistance and often produce extended-spectrum β-lactamases (ESBLs), which give them the ability to resist therapy with many β-lactam antibiotics including expanded-spectrum cephalosporins[2]. The emergence and global spread of ESBL-producing *E. coli* raise serious issues for clinical management.

Prior studies have shown that the *E. coli* ST131 population can be separated into three major phylogenetic clades[3]. Typing of the *fimH* locus has been traditionally used to classify isolates into clade A (*fimH*41), clade B (*fimH*22), and clade C (*fimH*30). Isolates belonging to clade A have been mostly found in Asia, whereas clade C isolates dominate in the United States[4]. The clade C population has further diverged into the nested subclades C1 (*fimH*30R) and C2 (*fimH*30Rx), with isolates in both subclades encoding mutations in the *gyrA* and *parC* genes that confer resistance to fluoroquinolones. Most isolates in the C2 subclade carry the ESBL gene $bla_{\text{CTX-M-15}}$, while isolates in the C1 subclade often carry $bla_{\text{CTX-M-27}}$[5]. ESBL genes are frequently maintained on mobile genetic elements (MGEs)[6], which are often carried on plasmids but can also be integrated into the chromosome[7].

Here we survey the genomic diversity and evolution of ESBL-producing ST131 *E. coli* isolates at a single medical center in the Pittsburgh area over a 15-year period. We describe the distribution of subclades and the diversity of ESBL-encoding MGEs, as well as the evolution of clade C isolates specifically, at our hospital. Our results suggest that a diverse ST131 *E. coli* population circulates in our facility, from which we periodically sampled. We also found evidence that distinct ST131 subpopulations have persisted in our hospital for over a decade, suggesting that multiple subclades are stably maintained in this setting.

## Results

### The ESBL-producing *E. coli* ST131 population at UPMC is dominated by clade C

To survey the genomic diversity of ESBL-producing ST131 *E. coli* at the University of Pittsburgh Medical Center (UPMC), we sequenced the genomes of 154 clinical isolates collected from patients between 2004 and 2018 (Table S1). ESBL-producing *E. coli* isolates collected between 2004 and 2016 were tested with PCR using ST131-specific primers[8], and up to ten ST131 isolates from each year were selected for whole genome sequencing. Beginning in 2016, isolates were identified as ST131 through analysis of whole genome sequence data generated previously[9]. We included isolates belonging to ST131 based on multi-locus sequence typing (MLST), as well as three isolates that belonged to ST8347 (a single locus variant of ST131) and two isolates that belonged to two additional single locus variants of ST131 that have not yet been assigned a sequence type (Fig. 1).

A recombination-filtered phylogenetic tree based on variants found in the core genome of all 154 isolates was constructed using RAxML (Fig. 1). As expected for the ST131 population[4,6,10], isolates resided on three major branches. The first branch (clade A) contained twelve isolates (7.8%), including eight with *fimH*41, three with *fimH*89, and one with a novel *fimH* sequence that was most similar to *fimH*41 (Fig. 1). These isolates were all collected in 2013 and later (Fig. S1). An additional four isolates (2.6%) collected in 2005, 2007, and 2010 encoded *fimH*22 and belonged to clade B. The third branch consisted of the remaining 138 isolates (89.6%), including one group of four isolates that encoded *fimH*5. The rest of the isolates on this branch encoded *fimH*30, indicating that the clade should be assigned as clade C (Fig. 1). QRDR mutations in *gyrA* and *parC* were detected in all 138 clade C isolates. The 86 isolates carrying two mutations described previously[4] were assigned to subclade C2. Within this clade, the four isolates encoding *fimH*5 were designated as subgroup C2a. The remaining 52 clade C isolates were

classified as subclade C1. Clade C isolates were collected throughout the study period and there was no apparent difference in collection dates of subclade C1 versus C2 isolates (Fig. S1).

**Evolution of clade C and stable maintenance of subclades C1 and C2 in the Pittsburgh area**

Prior studies have suggested that clade C emerged in approximately 1990[6,10,11]. To examine the evolution of clade C in our hospital, we performed a time-calibrated phylogenetic analysis using TreeTime (Fig. 2)[12]. The estimated substitution rate was 1.76 core genome mutations per genome per year, and the estimated root date of clade C was 1988.5. In addition, when we re-rooted the phylogenetic tree to separate subclades C1 and C2, we confirmed that the C2a subgroup was embedded within subclade C2. The estimated date of emergence of this subgroup from the subclade C2 population was approximately 2013 (Fig. 2).

We identified a roughly 40%/60% split between isolates belonging to subclades C1 versus C2. Due to the persistence of both subclades, we investigated if these subclades differed in infection sites and antimicrobial resistance (AMR) gene presence. The only differences we observed in isolate source between the two clades, however, were slightly more blood isolates belonging to subclade C2 and slightly more respiratory isolates belonging to subclade C1 (Table S1). We identified acquired AMR genes in all genomes in our dataset, and then examined the AMR gene content in subclade C1 versus C2 genomes (Table S2, Fig. S2). We found that subclade C1 isolate genomes encoded slightly more AMR genes compared to subclade C2 genomes, however the difference was not significant (mean 7.8 vs. 7.1 genes, $P$=0.178). We also observed differences in the prevalence of individual genes conferring resistance to several different antibiotic classes between the different subclades, including aminoglycosides, antifolates, macrolides, and sulfonamides (Fig. S2).

**Minimal gene enrichment in subclade C1 and C2 genomes**

We performed a pan-genome analysis for the 138 genomes in clade C using Roary[13] to identify genes that may be beneficial in clade persistence. Among the 11,587 genes in the clade C pangenome, 3,429 genes were shared among all clade C genomes, representing 70.3% of the average number of genes among genomes in this clade (Table S3). Using an 80%/20% enrichment cut-off, there were only 13 genes that were enriched among subclade C1 genomes (Table S4), and no genes were enriched among subclade C2 genomes, perhaps because this subclade was larger and more diverse than subclade C1. Nearly all the 13 genes enriched among subclade C1 genomes appeared to be plasmid-encoded and were predicted to encode hypothetical proteins (Table S4).

Within subclade C2, we identified 56 genes that were specific to the *fimH*5 allele-carrying subgroup we designated as C2a (Fig. S3, Table S5). These genes appeared to be associated with several transposable units carrying carbohydrate and lipid metabolism genes as well as cell wall and cell membrane biogenesis genes (Table S5). We also identified a group of 27 subclade C2 genomes isolated between 2007 and 2018 that resided on the same phylogenetic branch, clustered together by accessory gene content, and carried 182 group-specific genes that we designated subgroup C2b (Fig. S3, Table S6). Approximately one third of these genes were associated with prophages, and 32 genes were predicted to reside within transposons. The remaining genes with annotated functions included carbohydrate transport and metabolism genes, antibiotic and heavy metal resistance genes, toxin genes, and cell envelope-associated factors (Table S6).

**Convergent evolution in subclades C1 and C2**

We analyzed core genome non-synonymous SNPs in non-recombined genes among all isolates in each subclade to identify genes with multiple, independent SNPs in different isolates (Fig. 3, Table S7, Table S8). We focused on genes that had at least three non-synonymous SNPs among subclade C1 genomes (Fig. 3A), and at least four non-synonymous SNPs among

subclade C2 genomes (Fig. 3B), as these genes would be unlikely to accrue so many mutations due to chance alone. Among subclade C1 genomes, the hydroxyacylglutathione hydrolase gene *gloB* and the peptidoglycan D,D-transpeptidase gene *ftsI* both possessed three different non-synonymous SNPs in three different isolates, and the undecaprenyl-phosphate 4-deoxy-4-formamido-L-arabinose transferase gene *arnC* possessed four different non-synonymous SNPs in five different isolates (Fig. 3C, Table S7). Both *ftsI* and *arnC* contribute to cell wall assembly, while *gloB* is involved in methylglyoxal detoxification[14]. Among subclade C2 genomes, two genes encoding hypothetical proteins (*DVT980_3104* and *DVT980_4259*) each possessed four different non-synonymous SNPs (Fig. 3C). One of these proteins (DVT980_3104) was similar to the ribosome association toxin encoded by *ratA* and was mutated in four different isolates, while the other protein (DVT980_4259) was similar to the enterobactin siderophore exporter encoded by *entS* and was mutated in 19 isolates (Table S8). The peptidoglycan D,D-transpeptidase gene *ftsI* possessed five different non-synonymous SNPs in five different subclade C2 isolates, none of which overlapped with the three *ftsI* mutations detected in subclade C1 isolates. Two different mutations were detected at amino acid position 413 in *ftsI* (Ala413Val and Ala413Thr), strongly suggesting adaptive evolution of this gene. Finally, the yersiniabactin/pesticin outer membrane receptor gene *fyuA* possessed eight different non-synonymous SNPs in nine different C2 isolates; such a high number of independent mutations also suggests strong selection acting on this gene.

**ST131 clades carry diverse ESBL genes on both plasmids and the chromosome**

To examine the diversity of ESBL genes carried by the isolates we collected, we performed BLASTP searches against the ResFinder database[15]. A total of twelve different ESBLs were detected, including CTX-M, SHV, and TEM family enzymes (Fig. 4A, Table S9). The most common ESBL enzyme detected was CTX-M-15, which was found in 93 genomes and was dominant in subclade C2 (79/83, 95.18%). Outside of subclade C2, CTX-M-15 was also found in

nine subclade C1 genomes and in one clade A genome (Fig. 4A). The second most common ESBL enzyme detected was CTX-M-27, which was found in 32 genomes and was the most prevalent enzyme detected in subclade C1 (26/51, 50.98%) and clade A (6/12, 50%). CTX-M-27 was first detected in 2013, and was the dominant ESBL type identified in subclade C1 and in clade A in 2017 and 2018 (Table S1). The third most common enzyme we detected was CTX-M-14, which was found in nine genomes and was not associated with any specific clade or subclade (Fig. 4A). The remaining ESBL enzymes detected were CTX-M-2 (n=3), CTX-M-24 (n=3), CTX-M-1 (n=1), CTX-M-3 (n=1), SHV-12 (n=7), SHV-7 (n=1), TEM-19 (n=2), TEM-12 (n=2), and TEM-10 (n=1). One isolate (EC00670, belonging to subclade C2) was found to encode both CTX-M-14 and CTX-M-15 enzymes.

While ESBL genes are carried on MGEs, these elements can reside on plasmids or be integrated into the chromosome[1]. We assigned a putative genomic location of the ESBL enzyme in each isolate in our dataset using the MOB-RECON tool in MOB-Suite, which predicted whether ESBL-encoding contigs in each genome represented plasmid or chromosome sequences[16,17]. The majority of isolates (105/154, 68%) were predicted to carry ESBL genes on plasmids, while 46/154 (30%) were predicted to carry ESBL genes on the chromosome (Fig. 4A). The remaining isolates (3/154, 2%) were predicted to encode ESBL enzymes on both plasmids and the chromosome. Next, we used the 45 genomes that were hybrid assembled to examine the diversity and distribution of ESBL-encoding plasmids in our dataset. Among these 45 genomes we identified 35 ESBL-encoding plasmids, most of which belonged to the F family (Table S9). We then searched for each of these plasmids in all genomes in our dataset, and found that 11 plasmids were likely present in more than one isolate (Fig. S4). Four different *bla*$_{\text{CTX-M-15}}$-carrying plasmids were found among subclade C2 genomes exclusively, while six of the other seven plasmids were found in isolates belonging to multiple clades. A total of 33 isolates that had ESBL enzymes predicted to be plasmid-encoded did not match to any of the

35 resolved ESBL-encoding plasmids using the identity and coverage cut-offs we employed (detailed further in the Methods), and likely contain different plasmid sequences.

Among the 45 hybrid assembled genomes, we identified eight genomes that had ESBL genes at more than one locus (Fig. 4A). The EC00610 genome carried three separate loci encoding CTX-M-24, all of which were on the chromosome. The EC00661 genome carried three loci encoding CTX-M-15, two of which were on chromosome and one of which was on a plasmid. The DVT1260 genome also carried two chromosomal loci encoding CTX-M-15, while the EC00685 and EC00635 genomes both encoded one CTX-M-15 locus on the chromosome and another locus on a plasmid. The EC00670 genome encoded one CTX-M-14 locus and one CTX-M-15 locus, each on two different plasmids, and the DVT1003 genome carried two loci encoding TEM-10 on two different plasmids. Finally, the EC00674 genome carried two loci encoding CTX-M-27 on the same plasmid.

To assess ESBL copy number variation in the isolates we collected, we quantified the estimated ESBL gene copy number in each genome by comparing Illumina sequencing read depth of the ESBL gene with the read depth of all single copy genes in the core genome (Table S10). We found that estimated ESBL copy numbers varied from 0.39x to 40x, with a median copy number of 1.15x. Isolates with chromosomal ESBL genes had an average ESBL copy number of 1.34x and a standard deviation of 1.06x, while isolates with plasmid-encoded ESBL genes had an average ESBL copy number of 2.73x and a standard deviation of 5.28x (Figure 4B). ESBL copy numbers were significantly higher among isolates with plasmid-encoded ESBLs ($P = 0.0068$).

**ESBLs are flanked by mobile elements that vary by enzyme type**

To understand the genetic diversity of the elements carrying ESBL genes among the isolates we collected, we analyzed the genetic regions flanking the ESBL genes in each isolate in our study. Most assembled genomes allowed for examination of the genes immediately upstream

and downstream of the ESBL enzyme (Fig. 5, Fig. S5). We found that *bla*$_{CTX-M-15}$, which was present in 94% of subclade C2 isolates, very frequently resided in a conserved 3-kb region that was integrated into both plasmids and the chromosomes of different isolates (Fig. 5). We classified the *bla*$_{CTX-M-15}$-flanking regions based on similarities in their gene organization and orientation, and identified four different MGE types. The first *bla*$_{CTX-M-15}$-harboring MGE was found in isolates of clades A and C, and consisted of an IS*Ecp1* transposase and a small ORF with unknown function upstream of *bla*$_{CTX-M-15}$ (Fig. 5A). This MGE was similar to the IS*Ecp1*-*bla*$_{CTX-M-15}$-ORF477 transposition unit reported by Stoesser et al.[6]. The second MGE included the same upstream IS*Ecp1* transposase gene and small ORF with unknown function, as well as a Tn*2* transposase gene downstream of *bla*$_{CTX-M-15}$ (Fig. 5B). This MGE was similar to the putative *bla*$_{CTX-M-15}$ source element (Tn*2*-IS*Ecp1*-*bla*$_{CTX-M-15}$-ORF477-Tn*2*) reported by Stoesser et al.[6]. A third MGE was found exclusively on plasmids, and was flanked on either side by IS*26* elements (Fig. 5C). The fourth MGE was only present in subclade C2 genomes, and was found on predicted chromosomal contigs, however it appears to have integrated at different chromosomal positions in different isolates (Fig. 5D).

Apart from *bla*$_{CTX-M-15}$, a variety of different MGEs were found to carry the other ESBL genes we detected (Fig. S5). *bla*$_{CTX-M-27}$ was found on at least three different MGEs, and was associated with IS*15* and Tn*3* elements (Fig. S5A). Both *bla*$_{CTX-M-14}$ and *bla*$_{CTX-M-24}$ were found on the IS*Ecp1* MGE that also carried *bla*$_{CTX-M-15}$ (Fig. S5B, S5C). Finally, *bla*$_{SHV-12}$ was frequently found on a larger MGE that was flanked by IS*26* and contained additional carbohydrate metabolism genes (Fig. S5D).

## Discussion

In this 15-year study, we examined the genomic diversity and evolutionary dynamics of 154 ESBL-producing ST131 *E. coli* isolates from UPMC, a large healthcare system. Due to the multidrug resistance reported in ST131, numerous groups have characterized the clade structure of this pandemic lineage. Prior studies have suggested that clade C emerged around 1990[6,10,11]. Similarly, we identified the estimated root date to be midway through 1988. Our collection was dominated by isolates belonging to subclades C1 (*fim*H30-R) and C2 (*fim*H30-Rx). We identified the persistence of both clades at an approximate 40%/60% ratio, respectively. This finding suggests that these two subclades can coexist within the patient population that we sampled. We did not identify a significant difference in the number of AMR genes between the two clades, however, we did observe differences in the prevalence of individual genes conferring resistance to several different antibiotic classes. These data suggest that while subclade C1 and C2 isolates do not differ in their total AMR gene abundance, more subtle differences in the types of resistance genes they encode might contribute to their coexistence in the patient population that we sampled[18].

We sought to further investigate why the C1 and C2 subclades have stably coexisted over the last 30 years. While our data suggest that subclades C1 and C2 do not harbor clade-specific gene signatures, within subclade C2 we identified two groups that were each enriched for genes with potentially useful functions. These enriched genes may contribute to ongoing adaptation of subclade C2 in the Pittsburgh area. In addition to subclade-specifying genes, we also investigated whether distinct genes might be under positive selection in subclade C1 versus C2 genomes. We identified missense variants in *gloB* were only detected in subclade C1 genomes, suggesting that perhaps mutating this gene was only beneficial in the subclade C1 genetic background. Multiple independent mutations in *ftsI* and *arnC* were detected in both subclades, and might affect bacterial susceptibility to other cell wall-targeting antibiotics like

carbapenems[19], or membrane-targeting antibiotics like colistin[20], respectively. The *ratA*-like toxin and *entS* siderophore exporter genes were also independently mutated in multiple isolates across both subclades. These mutations might serve to decrease bacterial virulence, which frequently occurs during chronic infection and host adaptation[21]. Lastly, mutations in *fyuA* were also detected in both subclades, however they were heavily biased toward subclade C2 genomes. Prior studies have shown that *fyuA* function is critical for biofilm formation in iron-poor environments like the urinary tract[22]; mutations that alter or abrogate *fyuA* function would be predicted to decrease iron scavenging and biofilm formation. Future studies of the functional consequences of *fyuA* mutations on bacterial virulence and host-pathogen interactions may produce additional insights as to why these mutations appear to be under selection in ESBL-producing ST131 *E. coli* from our setting.

In agreement with previous reports, we identified a strong association CTX-M-15 and subclade C2 and CTX-M-27 and subclade C1[4,18,23]. The first isolate harboring CTX-M-27 in our collection was identified in 2013, coinciding with the recent emergence of CTX-M-27 documented in Europe and Asia[5,24,25]. When we predicted the location of the 154 ESBL-positive isolates, roughly a third were identified on the chromosome. While only one prior study has reported the chromosomal integration of CTX-M-14 in *E. coli* isolates from Mongolian wild birds[26], this phenomenon has been described in a previous report of *Klebsiella pneumoniae* blood isolates, where nearly a quarter showed chromosomally-encoded EBSLs[27]. This finding suggests that the integration of the ESBL enzyme onto the chromosomal might enhance stable propagation and expression.

In addition to carrying a wide variety of ESBL genes, the ST131 *E. coli* isolates we sampled also carry a large diversity of ESBL-encoding plasmids. Some of these were specific to individual ST131 subclades, while others were identified widely throughout the lineage. We identified instances where isolates carried multiple ESBLs, either on different plasmids and/or integrated onto the chromosome. These data suggest that ESBL enzymes are frequently

present at multiple loci within ST131 genomes, however these features can be difficult to resolve from Illumina draft genomes. Given that nearly 20% of our hybrid assembled genomes encoded ESBL enzymes at more than one locus, it is very likely that there are additional isolates in our dataset that also encode ESBL genes at multiple loci. The significance of this is unclear but could be due to gene dosage, plasmid instability, and/or shifting selective pressures during infection and antibiotic treatment[28,29].

Prior studies have demonstrated that copy number variation of antibiotic resistance genes like β-lactamases impacts antibiotic susceptibility and facilitates the evolution of antibiotic resistance[30,31]. Our findings of variable ESBL copy numbers among the isolates we sequenced suggests that antibiotic selection might have increased the ESBL-encoding plasmid copy number in some isolates. Alternately, plasmid instability or fitness costs could have decreased copy numbers in other isolates. These findings relate to gene abundance and not transcript or protein abundance, nonetheless we find that ESBL gene copy numbers were both higher and more variable in isolates with plasmid encoded ESBLs.

ESBLs in ST131 *E. coli* are most often carried by MGEs that are integrated into plasmids or the chromosome[32,33]. Similar to prior work, we have identified the regions flanking $bla_{\text{CTX-M-15}}$ have been found to be well conserved, even in distantly related genomes[6]. Further, through characterizing a variety of different MGE with ESBLs, our findings indicate that ESBL genes in the isolates from our medical center are likely shuttled between bacteria by MGEs that vary by enzyme type. Additionally, these elements appear to have integrated at different locations on both the plasmid and chromosome. It is notable that we observed a wide variety of different MGEs among the ST131 ESBL-producing *E. coli* sampled from a single geographic location. This suggests that as in other locations[34,35], no single ESBL enzyme or MGE type was dominant at our center during the study period.

## Conclusions

This study describes ongoing adaptation of the ST131 *E. coli* population sampled from clinical cultures of patients in a single hospital in Pittsburgh. While the vast majority of isolates we collected belonged to ST131 clade C, both subclades C1 and C2 appear to be stably maintained over time in our facility. Despite this stable maintenance, we found an abundant diversity of ESBL enzyme types and a vast array of different mobile elements carrying these enzymes on both plasmids and the chromosome. The diversity of antimicrobial resistance genes, movement of plasmids and other MGEs, and signals of adaptation we identified will be the focus of our future work in this area.

## Methods

### Sample collection

Clinical bacterial isolates were collected from patients at the University of Pittsburgh Medical Center (UPMC), an adult tertiary care hospital with over 750 beds, 150 critical care unit beds, more than 32,000 yearly inpatient admissions, and over 400 solid organ transplants per year. Bacterial isolates included in this study were collected from patients as part of routine clinical care and were collected before they otherwise would have been discarded. The study was designated by the University of Pittsburgh institutional review board as being exempt from informed consent. Isolates were collected from 2004 through 2018, and were identified as *E. coli* initially by the clinical microbiology laboratory. ST131 isolates were identified with PCR using lineage-specific primers on isolates collected between 2004 and 2016[8], or through analysis of whole genome sequences generated by the Enhanced Detection System for Healthcare-Associated Transmission (EDS-HAT) project in 2016-2018[9]. Collection of bacterial isolates was approved by the University of Pittsburgh institutional review board. ESBL phenotypes were inferred by the presence of an intact β-lactamase enzyme predicted to have

ESBL activity within the genome of each isolate. Single bacterial colonies were isolated, and were grown on blood agar plates or in Lysogeny Broth (LB) media prior to genomic DNA extraction.

**Whole-genome sequencing**

Genomic DNA was extracted from each isolate using a Qiagen DNeasy Tissue Kit according to the manufacturer's instructions (Qiagen, Germantown, MD). Illumina library construction and sequencing were conducted using an Illumina Nextera DNA Sample Prep Kit with 150-bp paired-end reads, and libraries were sequenced on the NextSeq 550 sequencing platform (Illumina, San Diego, CA) at the Microbial Genome Sequencing Center (MiGS). A total of 45 isolates were also sequenced on a MinION device (Oxford Nanopore Technologies, Oxford, United Kingdom). Long-read sequencing libraries were prepared and multiplexed using a rapid multiplex barcoding kit (catalog SQK-RBK004) and were sequenced on R9.4.1 flow cells. Base-calling on raw reads was performed using Albacore v2.3.3 or Guppy v2.3.1 (Oxford Nanopore Technologies, Oxford, UK).

Short and long reads (or short reads alone) were used as inputs for Unicycler to generate draft genomes[36]. Plasmid and chromosomal contigs were predicted with the MOB-RECON tool in MOB-Suite v3.1.7[16,17], and Prokka 1.14.5 was used for genome annotation[37]. Illumina raw reads for all isolates have been submitted to NCBI under BioProjects PRJNA475751 and PRJNA874473. Hybrid assembled genomes have been submitted to GenBank with accession numbers listed in Table S1.

**MLST, *fimH*, *gyrA/parC*, and clade C2 SNP Genotyping**

Multi-locus sequence typing (MLST) was performed with SRST2[38]. Typing of the *fimH* locus was performed by running BLASTN against the *fimH* sequence database downloaded from FimTyper[39,40]. To detect quinolone resistance-determining region (QRDR) mutations, amino acid residues 81-87 of *gyrA* and the 78-84 of *parC* were extracted and compared[41]. To detect clade

C2-specific single nucleotide polymorphisms (SNPs), targeted regions of primer sets described previously[4] were extracted from all genomes and were compared with BLASTN.

**Phylogenetic trees and the time-scaled phylogeny**

Among hybrid assembled genomes, the earliest collected isolate (DVT980) was used as a reference genome for Snippy v 4.6.0 to identify SNPs among the isolates using short read data and to generate a core SNP alignment (https://github.com/tseemann/snippy). The alignments were used as input for RAxMLHPC v 8.2.12 with [-m ASC_GTRCAT --asc-corr=lewis -V] flags to generate phylogenetic trees[42]. ClonalFrameML v1.12 was then used to filter recombinogenic regions[43]. Resulting trees were visualized with iTOL v6.3[44] or FigTree v1.4.4 (https://github.com/rambaut/figtree/). Branch bootstraps supporting the clade C phylogeny were evaluated using RaxMLHPC with 100 rapid bootstrapping replicates with [-m ASC_GTRCAT -f a --asc-corr lewis -V] flags. Estimation of evolutionary rate and a time scaled phylogeny of clade C isolates was generated with TreeTime v0.9.2[12], using a phylogenetic tree, ClonalFrameML-trimmed alignment, and the collection dates of the 138 isolates in clade C as input.

**ESBL gene detection and copy number variation**

Amino acid sequences of all protein coding genes annotated by Prokka were used as queries to run BLASTP against the ResFinder amino acid database[15,40]. Hits with 100% identity and 100% length coverage were then filtered and manually curated to only include ESBL genes. Isolates with less than perfect matches to a database entry were compared with the NCBI non-redundant protein sequences (nr) database with BLASTP. All ESBL enzymes reported are perfect protein sequence matches. To estimate the copy number of the ESBL gene(s) in each genome, Illumina raw reads were mapped to the assembled draft genome using BWA with default parameters[45]. The read depth covering each gene was then calculated via the MULTICOV function of BEDTOOLS v2.30.0, with the input BAM file generated by BWA and the BED file that includes all protein coding genes, tRNAs, and rRNAs[46]. To normalize read coverage, we used an AWK pipeline to calculate the reads per kilobase per million mapped

reads (RPKM) for each gene based on the depth list output of BEDTOOLS. A list of single copy genes shared by all genomes included in this study was extracted from the <gene_presence_absence.csv> output file of Roary v3.13.0[13]. For each genome, the median RPKM value of the single copy genes was calculated using the median() function in R. ESBL gene copy number in each genome was estimated by dividing the RPKM value of the ESBL gene(s) by the median RPKM value of single copy genes for the same genome.

**ESBL-encoding plasmid detection and analysis of flanking regions**

A list of ESBL-encoding reference plasmids was first generated from all hybrid assembled genomes and plasmid contigs identified by MOB-RECON v3.1.7[16,17]. Contigs predicted to be circular by Unicycler v0.5.0 but not recognized as plasmids were not included in the reference plasmid list. To reduce redundancy, plasmids sharing >95% nucleotide similarity (defined as the product of query coverage and nucleotide identity) and encoding the same ESBL gene were combined and only the longest plasmid was retained. The remaining reference plasmids were then queried in all genomes using BLASTN and hits that had >95% nucleotide similarity were retained. Results were then manually curated to remove hits in genomes predicted to encode ESBLs on the chromosome only and hits to reference plasmids harboring a different ESBL. Among Illumina-only genomes, if there were hits to multiple reference plasmids with the same ESBL, only the longest reference plasmid was reported. To assess ESBL flanking regions, DNA segments containing up to 15 genes upstream and downstream of each ESBL gene were visualized via the R package genoPlotR, and were manually aligned centering on the ESBL gene to visualize conservation and enable classification of ESBL-containing MGEs[47].

**Identification of subclade-specific genes and SNPs for clade C**

The 138 annotated genomes belonging to clade C, including four genomes in clade C2a, were used for pangenome analysis. The pangenome analysis tool ROARY was used to generate a gene presence and absence matrix (gene_presence_absence.csv). Genes enriched in each clade were identified as those that were present in more than 80% of isolates within the

clade and less than 20% of isolates outside the clade. The pangenome matrix was visualized using the heatmap() function in R. Genes associated with prophages and transposons were identified using PHASTER and MobileElementFinder, respectively[48-50]. Snippy was used to identify SNPs among clade C1 and C2 isolates using the DVT980 (earliest collected isolate) hybrid assembled genome as a reference. SNPs found in genomic regions identified by ClonalFrameML as putative recombinations were then masked. SNPs located in clade C core genes were annotated with gene description and locus tag of the reference genome. SNPs were then examined manually to identify genes with repeated and independent mutations within each subclade.

## Acknowledgements

We gratefully acknowledge Jane Marsh, Akansha Pradhan, and Alecia Rokes for their helpful input throughout the course of this study. This work was supported by grant R01AI127472 from the National Institutes of Health (L.H.H.), grant DAA3-19-65600-1 from the US Civilian Research & Development Foundation (D.V.T), and by the Department of Medicine at the University of Pittsburgh, School of Medicine (D.V.T.). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

## Author Contributions

SC and DVT designed the study. LHH and YD provided bacterial isolates. SC, MPG, HRN, and CLM performed experiments and generated results. SC, EGM, and DVT wrote the manuscript. All authors reviewed the manuscript and approved of its contents.

## Competing Interests

437 The authors have no relevant conflicts of interest to declare.

438

439 **Legends**

440 **Figure 1. Genetic diversity of 154 ESBL-producing ST131 *E. coli* isolates.** The maximum
441 likelihood phylogeny was constructed with RAxML from 18,734 core genome single nucleotide
442 polymorphisms (SNPs). Background shading of each isolate indicates the ST131 clade (A, B),
443 subclade (C2, C2), or subgroup (C2a). Multi-locus sequence type (ST), source, and date of
444 isolation are shown as color blocks next to each isolate. *fimH* alleles were predicted from
445 genome sequences.

446 **Figure 2. Time-calibrated phylogeny of 138 clade C isolates.** The molecular-clock phylogeny
447 was inferred from 2,656 aligned SNPs and was constructed with TreeTime. Subclades C1 and
448 C2 are indicated with green and blue branches, respectively. Subgroup C2a is shaded pink. The
449 distribution of root-to-tip distances versus isolation date of all terminal nodes in the time-scaled
450 tree is shown in the inset graph.

451 **Figure 3. Genes putatively under selection among clade C *E. coli* isolates.** Enrichment of
452 nonsynonymous (NSY) mutations among subclade (A) C1 and (B) C2 genomes. Frequency
453 distributions show the number of genes with one or more NSY mutation detected. (C) Genes
454 with at least three unique NSY mutations in subclade C1 genomes or at least four unique NSY
455 mutations in subclade C2 genomes. The number of different mutations detected in each gene
456 among the genomes in each subclade is shown.

457 **Figure 4. ESBL gene diversity, genomic location, and copy number.** (A) Distribution of
458 ESBL genes. ESBL locations (plasmid/chromosome/multiple loci) and types are shown as color
459 blocks next to the isolate names. (B) Box plot showing ESBL gene copy number in isolates
460 predicted to encode an ESBL gene on the chromosome or on a plasmid. *P*-value was calculated
461 using a two-tailed t-test.

**Figure 5. Regions flanking *bla*$_{\text{CTX-M-15}}$ among ST131 *E. coli* isolates.** (A-D) Genomic context of different *bla*$_{\text{CTX-M-15}}$-carrying MGEs is shown. Isolate names are shaded based on their phylogenetic clade assignments (clade A=purple; subclade C2=blue; subclade C1=green; subgroup C2a=pink). The genomic location of each sequence is indicated (C=chromosome, P=plasmid) and *bla*$_{\text{CTX-M-15}}$ genes are colored red. Genes were annotated with Prokka, and genes with predicted functions are labeled. Genes associated with MGEs and transposases are highlighted with black outlines, and are colored if found in more than one region. Regions that were used for MGE classification are shaded in each panel.

**Figure S1. Timelines of ST131 isolate collection.** (A) Total number of isolates collected each year. (B) Collection timelines for isolates belonging to each clade, subclade, and subgroup in the dataset.

**Figure S2. Differences in antibiotic resistance gene content between ST131 clades and subclades.** (A) Antimicrobial resistance (AMR) gene abundance in isolates belonging to different ST131 clades and subclades. Horizontal lines show median values or AMR genes per genome in each isolate. AMR genes were identified by BLASTN to the ResFinder database. (B) Frequency of individual AMR genes among isolates in each clade or subclade. Genes with notable frequency differences between groups are shown. Complete data on AMR genes is provided in Table S2.

**Figure S3. Pangenome analysis of 138 clade C ST131 *E. coli* isolates.** Phylogenetic tree on the left was generated with RAxML using a core genome, post-ClonalFrameML SNP alignment. The tree was midpoint rooted to separate subclades C1 (green shaded) and C2 (blue shaded). The heatmap on the right shows the pangenome matrix generated by Roary. Each column represents one gene group, and each row represent one genome. The presence or absence of a gene in a given genome is shown as red or yellow, respectively. Subgroups C2a and C2b are labeled below the corresponding branches on the phylogenetic tree.

**Figure S4. Distribution of ESBL-encoding plasmids among ST131 *E. coli* isolates.** The core genome phylogeny is annotated with the presence of eleven ESBL-encoding reference plasmids that were detected in more than one genome in the dataset. Plasmids DVT1294_4, DVT1284_2, EC00661_2, and EC00635_3 harbor *bla*$_{\text{CTX-M-15}}$ (red); plasmids EC00675_2, EC00763_3, and EC00637_2 harbor *bla*$_{\text{CTX-M-27}}$ (orange); plasmids DVT1252_7, DVT1006_4, and EC00617_2 harbor *bla*$_{\text{SHV-12}}$ (purple); and plasmid DVT1001_2 harbors *bla*$_{\text{CTX-M-2}}$ (brown).

**Figure S5. Regions flanking ESBL genes among ST131 *E. coli* isolates.** (A-D) Genomic context of different ESBL-carrying MGEs is shown. Isolate names are shaded based on their phylogenetic clade assignments (clade A=purple; subclade C1=green; subclade C2=blue; clade B=yellow). The genomic context of each sequence is indicated (C=chromosome, P=plasmid) and ESBL genes are colored red. Genes were annotated with Prokka, and genes with predicted functions are labeled. Genes associated with MGEs and transposases are highlighted with black outlines, and are colored if found in more than one region. Regions that were used for MGE classification are shaded in each panel.

## Data Availability

Illumina raw reads and genome assemblies for all isolates have been submitted to NCBI under BioProjects PRJNA475751 and PRJNA874473. NCBI accession numbers for genome sequence data are listed in Table S1.

## References

1 Nicolas-Chanoine, M. H., Bertrand, X. & Madec, J. Y. Escherichia coli ST131, an intriguing clonal group. *Clin Microbiol Rev* **27**, 543-574, doi:10.1128/CMR.00125-13 (2014).

2 Nicolas-Chanoine, M. H. *et al.* Intercontinental emergence of Escherichia coli clone O25:H4-ST131 producing CTX-M-15. *J Antimicrob Chemother* **61**, 273-281, doi:10.1093/jac/dkm464 (2008).

3 Adams-Sapper, S., Diep, B. A., Perdreau-Remington, F. & Riley, L. W. Clonal composition and community clustering of drug-susceptible and -resistant

Escherichia coli isolates from bloodstream infections. *Antimicrob Agents Chemother* **57**, 490-497, doi:10.1128/AAC.01025-12 (2013).

4 Price, L. B. *et al.* The epidemic of extended-spectrum-beta-lactamase-producing Escherichia coli ST131 is driven by a single highly pathogenic subclone, H30-Rx. *mBio* **4**, e00377-00313, doi:10.1128/mBio.00377-13 (2013).

5 Matsumura, Y. *et al.* Global Escherichia coli Sequence Type 131 Clade with blaCTX-M-27 Gene. *Emerg Infect Dis* **22**, 1900-1907, doi:10.3201/eid2211.160519 (2016).

6 Stoesser, N. *et al.* Evolutionary History of the Global Emergence of the Escherichia coli Epidemic Clone ST131. *MBio* **7**, e02162, doi:10.1128/mBio.02162-15 (2016).

7 Zhang, C. Z. *et al.* The Emergence of Chromosomally Located bla CTX-M-55 in Salmonella From Foodborne Animals in China. *Front Microbiol* **10**, 1268, doi:10.3389/fmicb.2019.01268 (2019).

8 Matsumura, Y. *et al.* Rapid Identification of Different Escherichia coli Sequence Type 131 Clades. *Antimicrob Agents Chemother* **61**, doi:10.1128/AAC.00179-17 (2017).

9 Sundermann, A. J. *et al.* Whole Genome Sequencing Surveillance and Machine Learning of the Electronic Health Record for Enhanced Healthcare Outbreak Detection. *Clin Infect Dis*, doi:10.1093/cid/ciab946 (2021).

10 Ludden, C. *et al.* Genomic surveillance of Escherichia coli ST131 identifies local expansion and serial replacement of subclones. *Microb Genom* **6**, doi:10.1099/mgen.0.000352 (2020).

11 Kallonen, T. *et al.* Systematic longitudinal survey of invasive Escherichia coli in England demonstrates a stable population structure only transiently disturbed by the emergence of ST131. *Genome Res*, doi:10.1101/gr.216606.116 (2017).

12 Sagulenko, P., Puller, V. & Neher, R. A. TreeTime: Maximum-likelihood phylodynamic analysis. *Virus Evol* **4**, vex042, doi:10.1093/ve/vex042 (2018).

13 Page, A. J. *et al.* Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **31**, 3691-3693, doi:10.1093/bioinformatics/btv421 (2015).

14 Reiger, M., Lassak, J. & Jung, K. Deciphering the role of the type II glyoxalase isoenzyme YcbL (GlxII-2) in Escherichia coli. *FEMS Microbiol Lett* **362**, 1-7, doi:10.1093/femsle/fnu014 (2015).

15 Bortolaia, V. *et al.* ResFinder 4.0 for predictions of phenotypes from genotypes. *J Antimicrob Chemother* **75**, 3491-3500, doi:10.1093/jac/dkaa345 (2020).

16 Robertson, J. & Nash, J. H. E. MOB-suite: software tools for clustering, reconstruction and typing of plasmids from draft assemblies. *Microb Genom* **4**, doi:10.1099/mgen.0.000206 (2018).

17 Robertson, J., Bessonov, K., Schonfeld, J. & Nash, J. H. E. Universal whole-sequence-based plasmid typing and its utility to prediction of host range and epidemiological surveillance. *Microb Genom* **6**, doi:10.1099/mgen.0.000435 (2020).

18 Mills, E. G. *et al.* A one-year genomic investigation of Escherichia coli epidemiology and nosocomial spread at a large US healthcare network. *Genome Med* **14**, 147, doi:10.1186/s13073-022-01150-7 (2022).

19 Adler, M., Anjum, M., Andersson, D. I. & Sandegren, L. Combinations of mutations in envZ, ftsI, mrdA, acrB and acrR can cause high-level carbapenem resistance in Escherichia coli. *J Antimicrob Chemother* **71**, 1188-1198, doi:10.1093/jac/dkv475 (2016).

20 Sundaramoorthy, N. S., Suresh, P., Selva Ganesan, S., GaneshPrasad, A. & Nagarajan, S. Restoring colistin sensitivity in colistin-resistant E. coli: Combinatorial use of MarR inhibitor with efflux pump inhibitor. *Sci Rep* **9**, 19845, doi:10.1038/s41598-019-56325-x (2019).

21 Gatt, Y. E. & Margalit, H. Common Adaptive Strategies Underlie Within-Host Evolution of Bacterial Pathogens. *Mol Biol Evol* **38**, 1101-1121, doi:10.1093/molbev/msaa278 (2021).

22 Hancock, V., Ferrieres, L. & Klemm, P. The ferric yersiniabactin uptake receptor FyuA is required for efficient biofilm formation by urinary tract infectious Escherichia coli in human urine. *Microbiology (Reading)* **154**, 167-175, doi:10.1099/mic.0.2007/011981-0 (2008).

23 Wilson, H. & Torok, M. E. Extended-spectrum beta-lactamase-producing and carbapenemase-producing Enterobacteriaceae. *Microb Genom* **4**, doi:10.1099/mgen.0.000197 (2018).

24 Merino, I. *et al.* Emergence of ESBL-producing Escherichia coli ST131-C1-M27 clade colonizing patients in Europe. *J Antimicrob Chemother* **73**, 2973-2980, doi:10.1093/jac/dky296 (2018).

25 Peirano, G. & Pitout, J. D. D. Extended-Spectrum beta-Lactamase-Producing Enterobacteriaceae: Update on Molecular Epidemiology and Treatment Options. *Drugs* **79**, 1529-1541, doi:10.1007/s40265-019-01180-3 (2019).

26 Guenther, S. *et al.* Chromosomally encoded ESBL genes in Escherichia coli of ST38 from Mongolian wild birds. *J Antimicrob Chemother* **72**, 1310-1313, doi:10.1093/jac/dkx006 (2017).

27 Yoon, E. J. *et al.* Beneficial Chromosomal Integration of the Genes for CTX-M Extended-Spectrum beta-Lactamase in Klebsiella pneumoniae for Stable Propagation. *mSystems* **5**, doi:10.1128/mSystems.00459-20 (2020).

28 Hong, J. S. *et al.* Molecular Characterization of Fecal Extended-Spectrum beta-Lactamase- and AmpC beta-Lactamase-Producing Escherichia coli From Healthy Companion Animals and Cohabiting Humans in South Korea. *Front Microbiol* **11**, 674, doi:10.3389/fmicb.2020.00674 (2020).

29 Dimitriu, T., Matthews, A. C. & Buckling, A. Increased copy number couples the evolution of plasmid horizontal transmission and plasmid-encoded antibiotic resistance. *Proc Natl Acad Sci U S A* **118**, doi:10.1073/pnas.2107818118 (2021).

30 San Millan, A., Escudero, J. A., Gifford, D. R., Mazel, D. & MacLean, R. C. Multicopy plasmids potentiate the evolution of antibiotic resistance in bacteria. *Nat Ecol Evol* **1**, 10, doi:10.1038/s41559-016-0010 (2016).

31 Sun, S., Berg, O. G., Roth, J. R. & Andersson, D. I. Contribution of gene amplification to evolution of increased antibiotic resistance in Salmonella typhimurium. *Genetics* **182**, 1183-1195, doi:10.1534/genetics.109.103028 (2009).

32 Harmer, C. J. & Hall, R. M. IS26-Mediated Formation of Transposons Carrying Antibiotic Resistance Genes. *mSphere* **1**, doi:10.1128/mSphere.00038-16 (2016).

33 Shawa, M. *et al.* Novel chromosomal insertions of ISEcp1-blaCTX-M-15 and diverse antimicrobial resistance genes in Zambian clinical isolates of Enterobacter cloacae and Escherichia coli. *Antimicrob Resist Infect Control* **10**, 79, doi:10.1186/s13756-021-00941-8 (2021).

34 Kanamori, H. *et al.* Genomic Analysis of Multidrug-Resistant Escherichia coli from North Carolina Community Hospitals: Ongoing Circulation of CTX-M-Producing ST131-H30Rx and ST131-H30R1 Strains. *Antimicrob Agents Chemother* **61**, doi:10.1128/AAC.00912-17 (2017).

35 Ny, S., Sandegren, L., Salemi, M. & Giske, C. G. Genome and plasmid diversity of Extended-Spectrum beta-Lactamase-producing Escherichia coli ST131 - tracking phylogenetic trajectories with Bayesian inference. *Sci Rep* **9**, 10291, doi:10.1038/s41598-019-46580-3 (2019).

36 Wick, R. R., Judd, L. M., Gorrie, C. L. & Holt, K. E. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput Biol* **13**, e1005595, doi:10.1371/journal.pcbi.1005595 (2017).

37 Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068-2069, doi:10.1093/bioinformatics/btu153 (2014).

38 Inouye, M. *et al.* SRST2: Rapid genomic surveillance for public health and hospital microbiology labs. *Genome Med* **6**, 90, doi:10.1186/s13073-014-0090-6 (2014).

39 Roer, L. *et al.* Development of a Web Tool for Escherichia coli Subtyping Based on fimH Alleles. *J Clin Microbiol* **55**, 2538-2543, doi:10.1128/JCM.00737-17 (2017).

40 Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421, doi:10.1186/1471-2105-10-421 (2009).

41 Aoike, N. *et al.* Molecular characterization of extraintestinal Escherichia coli isolates in Japan: relationship between sequence types and mutation patterns of quinolone resistance-determining regions analyzed by pyrosequencing. *J Clin Microbiol* **51**, 1692-1698, doi:10.1128/JCM.03049-12 (2013).

42 Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312-1313, doi:10.1093/bioinformatics/btu033 (2014).

43 Didelot, X. & Wilson, D. J. ClonalFrameML: efficient inference of recombination in whole bacterial genomes. *PLoS Comput Biol* **11**, e1004041, doi:10.1371/journal.pcbi.1004041 (2015).

44 Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res* **47**, W256-W259, doi:10.1093/nar/gkz239 (2019).

45 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760, doi:10.1093/bioinformatics/btp324 (2009).

46 Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842, doi:10.1093/bioinformatics/btq033 (2010).

47 Guy, L., Kultima, J. R. & Andersson, S. G. genoPlotR: comparative gene and genome visualization in R. *Bioinformatics* **26**, 2334-2335, doi:10.1093/bioinformatics/btq413 (2010).

48 Zhou, Y., Liang, Y., Lynch, K. H., Dennis, J. J. & Wishart, D. S. PHAST: a fast phage search tool. *Nucleic Acids Res* **39**, W347-352, doi:10.1093/nar/gkr485 (2011).

49 Arndt, D. *et al.* PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acids Res* **44**, W16-21, doi:10.1093/nar/gkw387 (2016).

50 Johansson, M. H. K. *et al.* Detection of mobile genetic elements associated with antibiotic resistance in Salmonella enterica using a newly developed web tool: MobileElementFinder. *J Antimicrob Chemother* **76**, 101-109, doi:10.1093/jac/dkaa390 (2021).

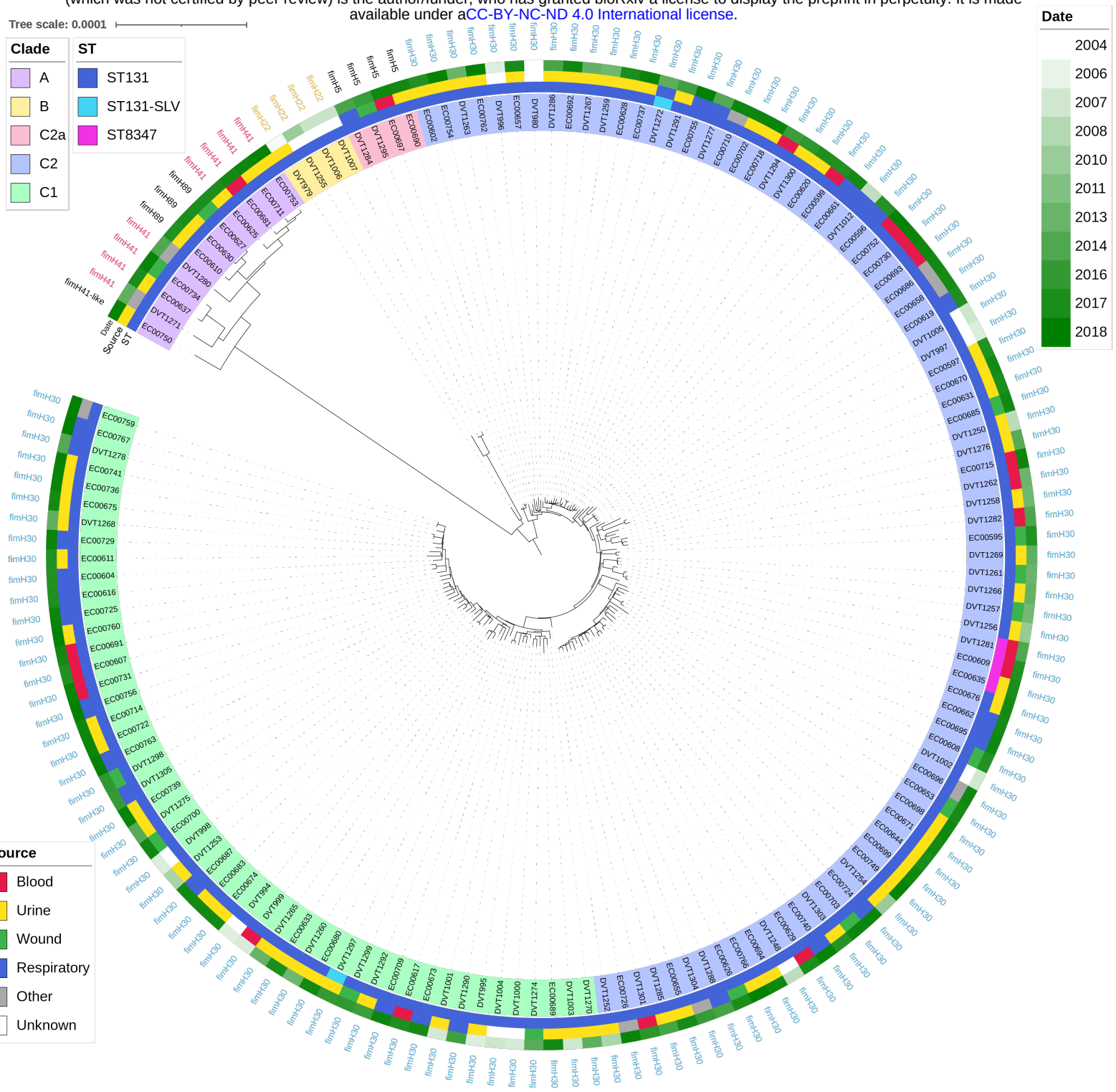



**Figure 1. Genetic diversity of 154 ESBL-producing ST131 *E. coli* isolates.** The maximum likelihood phylogeny was constructed with RAxML from 18,734 core genome single nucleotide polymorphisms (SNPs). Background shading of each isolate indicates the ST131 clade (A, B), subclade (C2, C2), or subgroup (C2a). Multi-locus sequence type (ST), source, and date of isolation are shown as color blocks next to each isolate. *fimH* alleles were predicted from genome sequences.

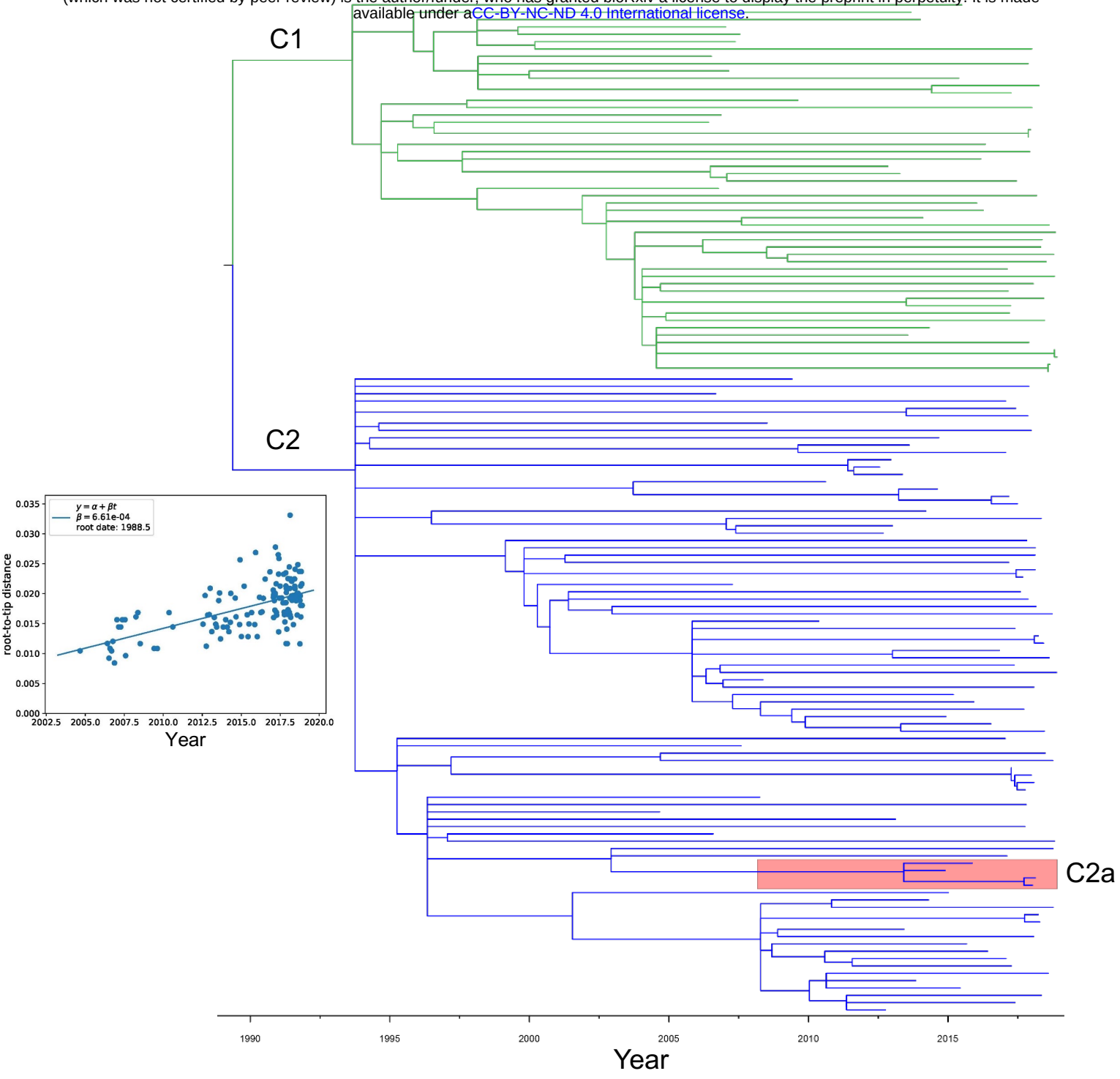



**Figure 2. Time-calibrated phylogeny of 138 clade C isolates.** The molecular-clock phylogeny was inferred from 2,656 aligned SNPs and was constructed with TreeTime. Subclades C1 and C2 are indicated with green and blue branches, respectively. Subgroup C2a is shaded pink. The distribution of root-to-tip distances versus isolation date of all terminal nodes in the time-scaled tree is shown in the inset graph.

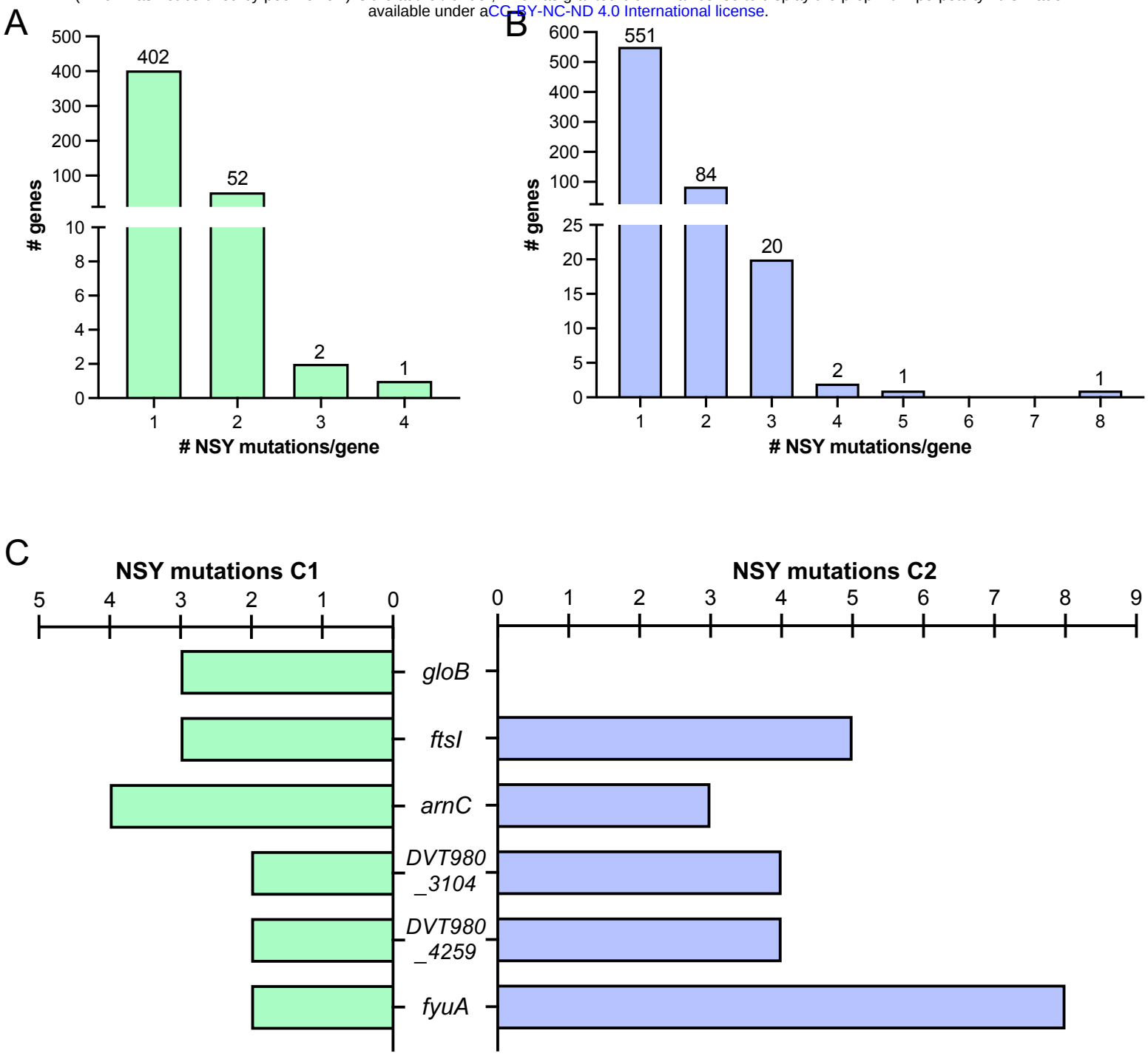



**Figure 3. Genes putatively under selection among clade C *E. coli* isolates.** Enrichment of nonsynonymous (NSY) mutations among subclade (A) C1 and (B) C2 genomes. Frequency distributions show the number of genes with one or more NSY mutation detected. (C) Genes with at least three unique NSY mutations in subclade C1 genomes or at least four unique NSY mutations in subclade C2 genomes. The number of different mutations detected in each gene among the genomes in each subclade is shown.

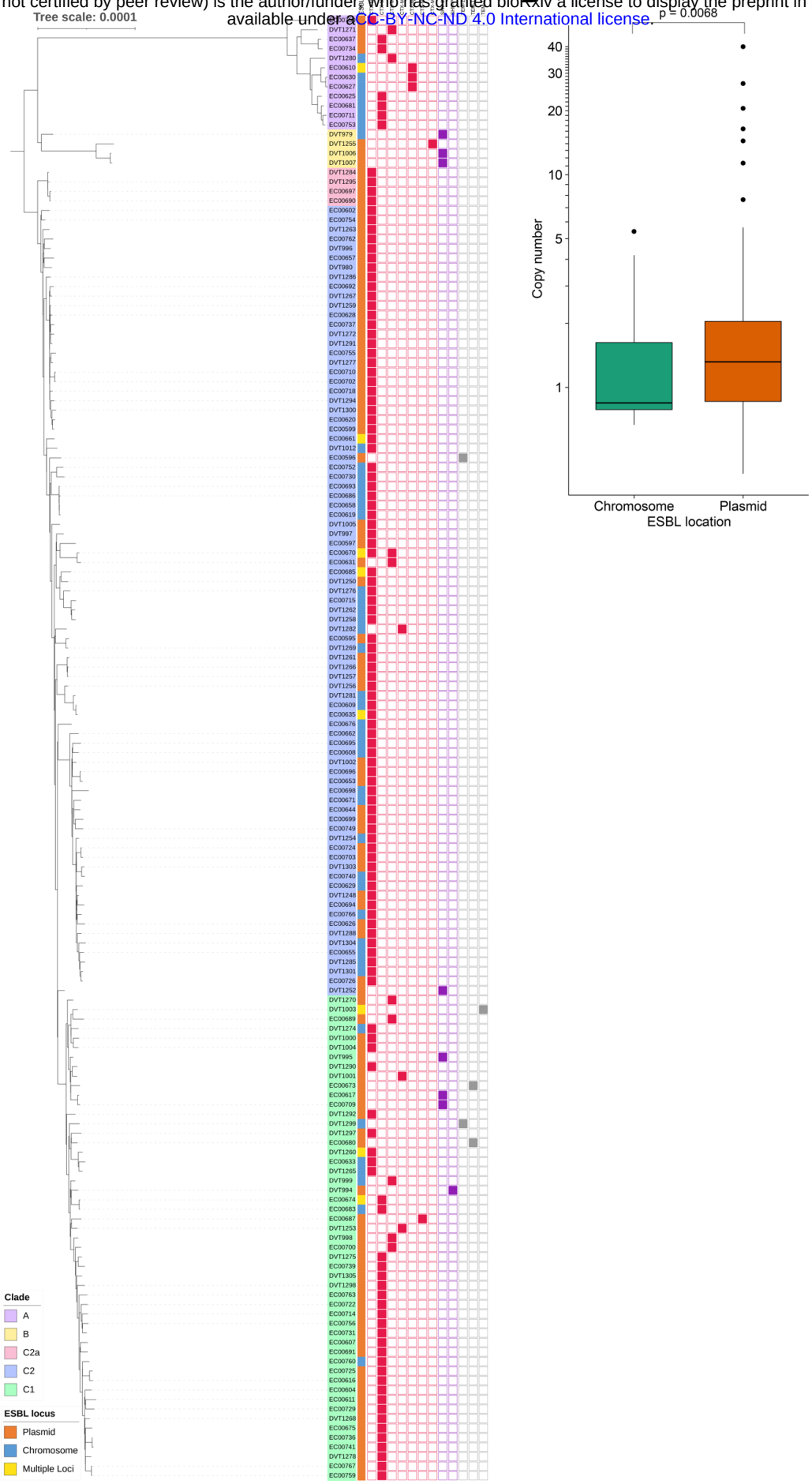
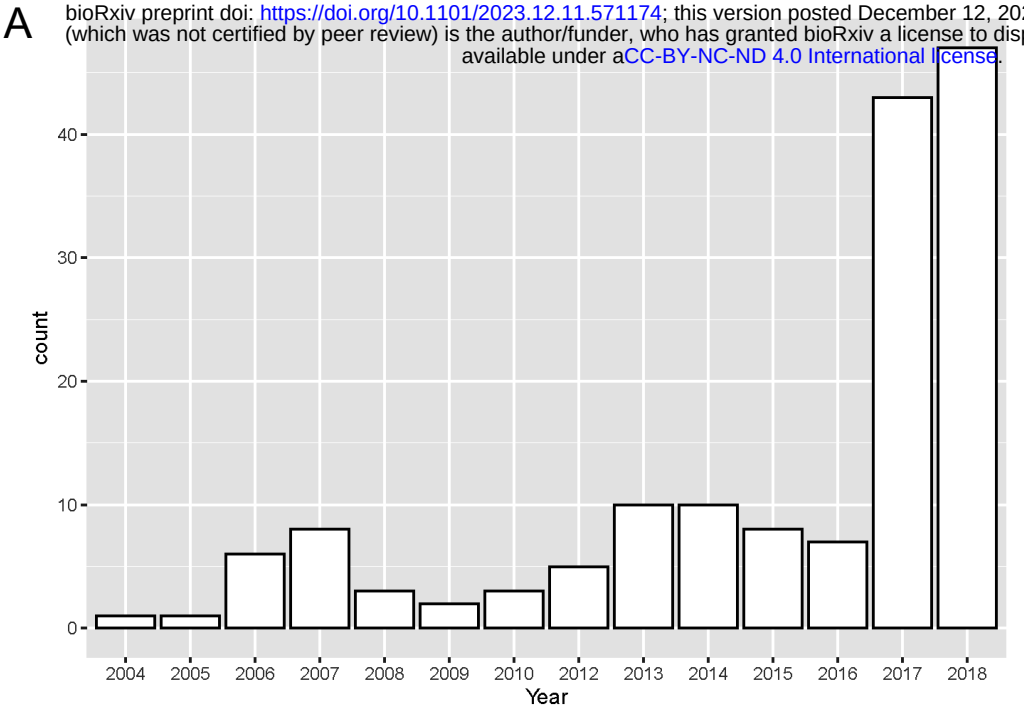



**Figure 4. ESBL gene diversity, genomic location, and copy number.** (A) Distribution of ESBL genes. ESBL locations (plasmid/chromosome/multiple loci) and types are shown as color blocks next to the isolate names. (B) Box plot showing ESBL gene copy number in isolates predicted to encode an ESBL gene on the chromosome or on a plasmid. *P*-value was calculated using a two-tailed t-test.


**Figure 5. Regions flanking *bla*$_{CTX-M-15}$ among ST131 *E. coli* isolates.** (A-D) Genomic context of different *bla*$_{CTX-M-15}$-carrying MGEs is shown. Isolate names are shaded based on their phylogenetic clade assignments (clade A=purple; subclade C2=blue; subclade C1=green; subgroup C2a=pink). The genomic location of each sequence is indicated (C=chromosome, P=plasmid) and *bla*$_{CTX-M-15}$ genes are colored red. Genes were annotated with Prokka, and genes with predicted functions are labeled. Genes associated with MGEs and transposases are highlighted with black outlines, and are colored if found in more than one region. Regions that were used for MGE classification are shaded in each panel.



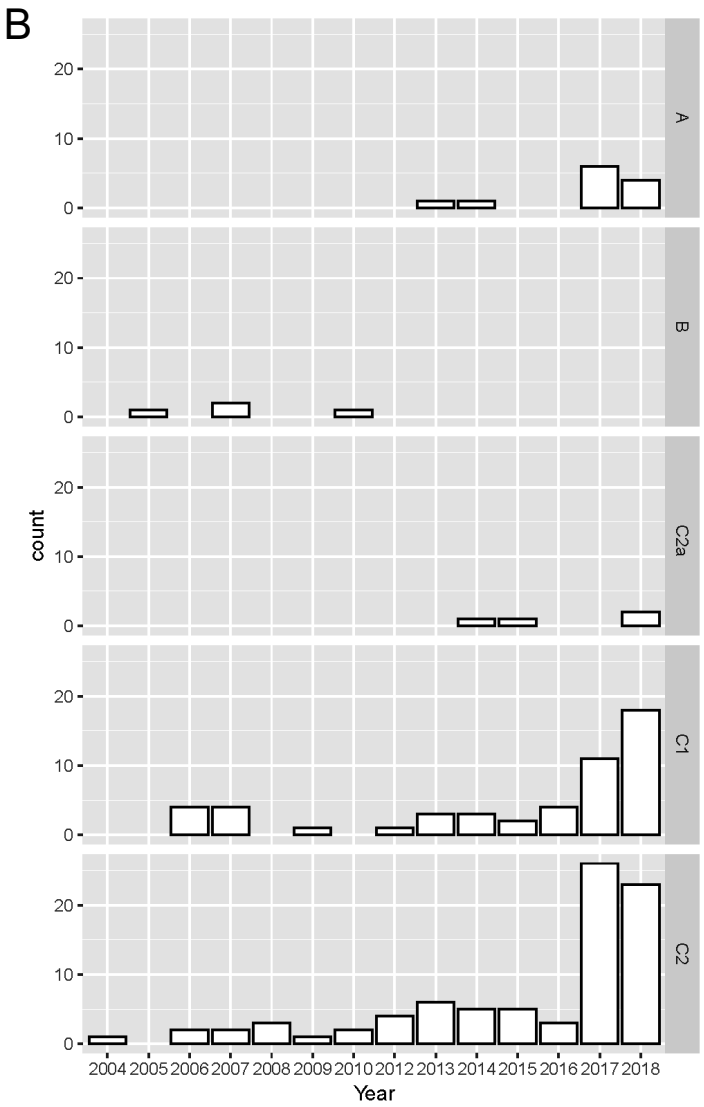




**Figure S1. Timelines of ST131 isolate collection.** (A) Total number of isolates collected each year. (B) Collection timelines for isolates belonging to each clade, subclade, and subgroup in the dataset.

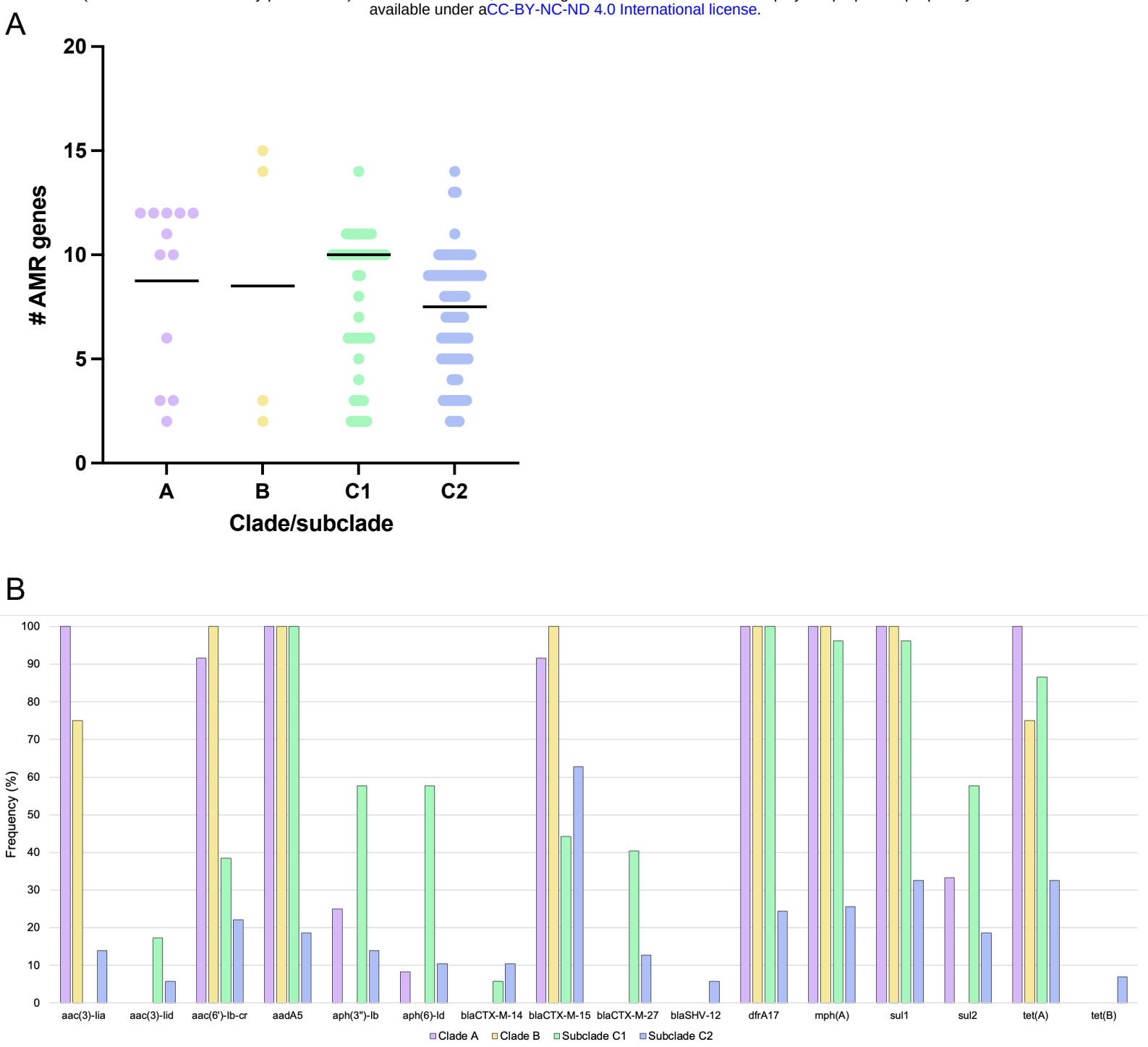



**Figure S2. Differences in antibiotic resistance gene content between ST131 clades and subclades.** (A) Antimicrobial resistance (AMR) gene abundance in isolates belonging to different ST131 clades and subclades. Horizontal lines show median values or AMR genes per genome in each isolate. AMR genes were identified by BLASTN to the ResFinder daabase. (B) Frequency of individual AMR genes among isolates in each clade or subclade. Genes with notable frequency differences between groups are shown. Complete data on AMR genes is provided in Table S2.

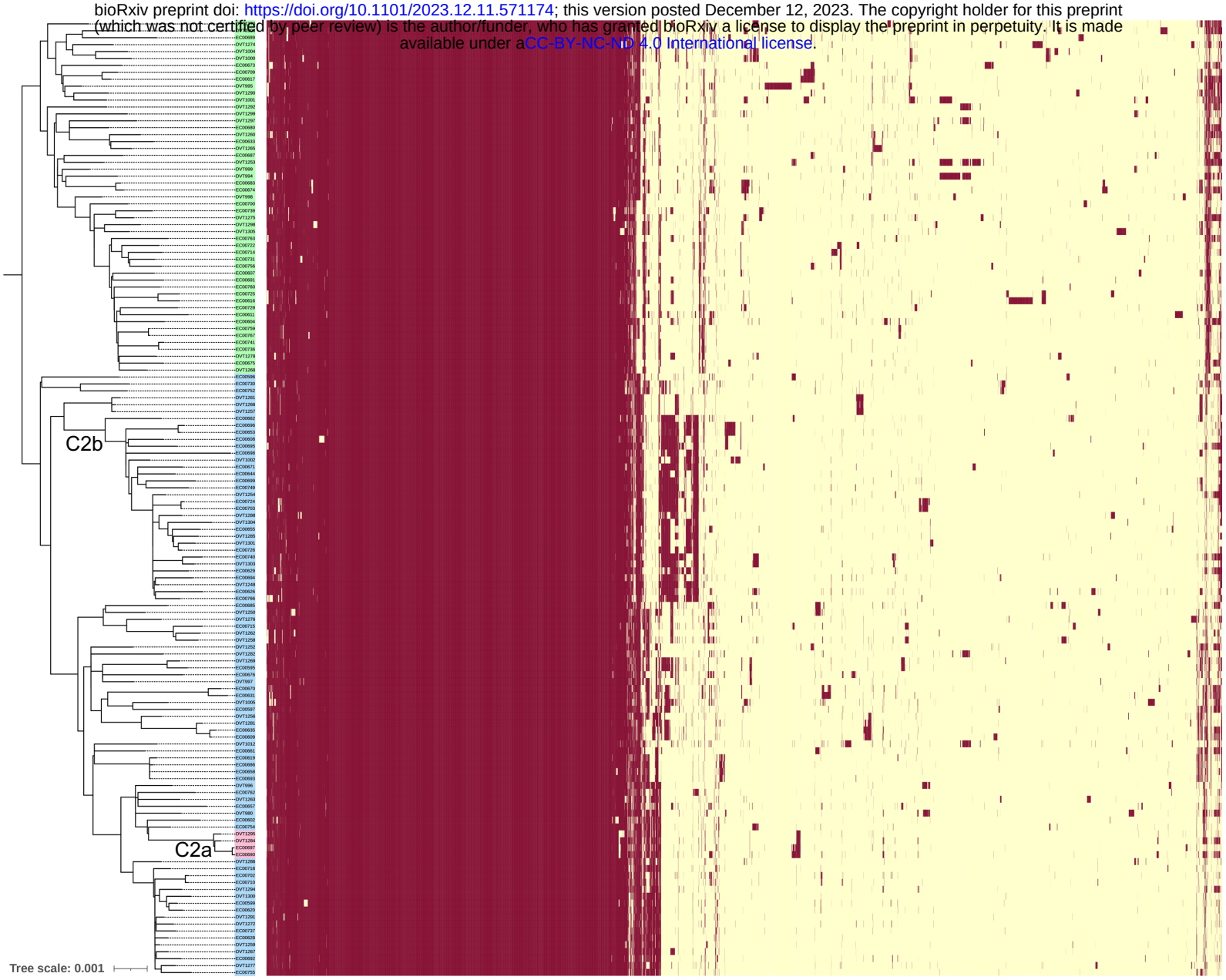



**Figure S3. Pangenome analysis of 138 clade C ST131 *E. coli* isolates.** Phylogenetic tree on the left was generated with RAxML using a core genome, post-ClonalFrameML SNP alignment. The tree was midpoint rooted to separate subclades C1 (green shaded) and C2 (blue shaded). The heatmap on the right shows the pangenome matrix generated by Roary. Each column represents one gene group, and each row represent one genome. The presence or absence of a gene in a given genome is shown as red or yellow, respectively. Subgroups C2a and C2b are labeled below the corresponding branches on the phylogenetic tree.

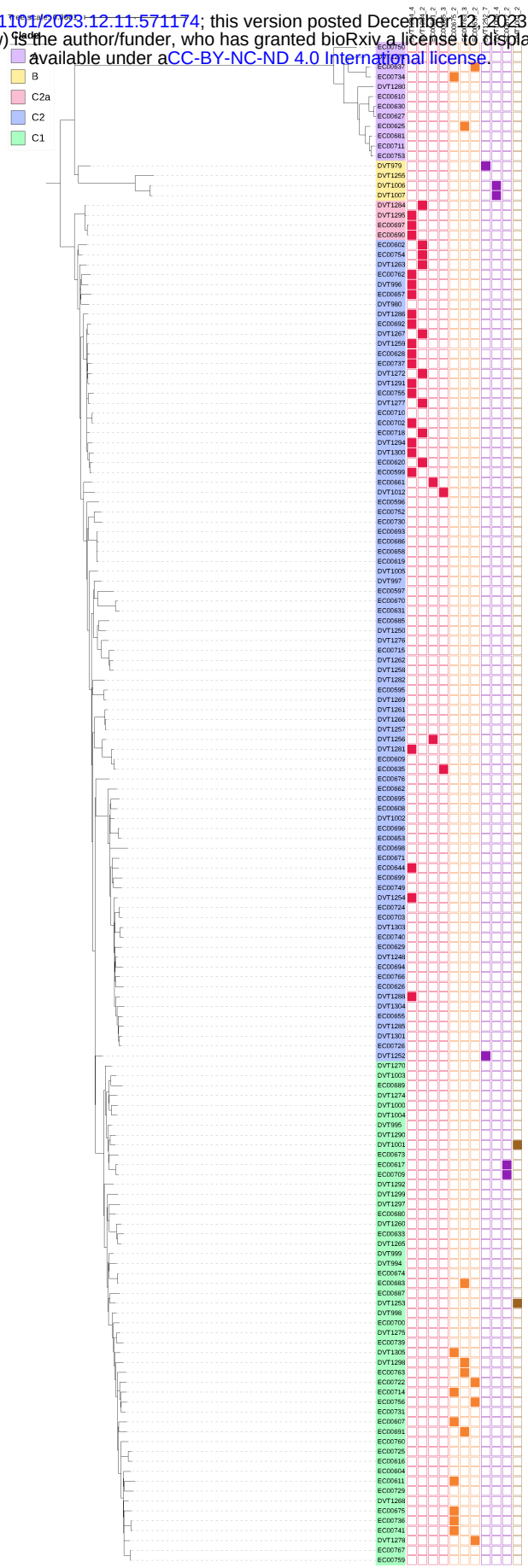



**Figure S4. Distribution of ESBL-encoding plasmids among ST131 *E. coli* isolates.** The core genome phylogeny is annotated with the presence of eleven ESBL-encoding reference plasmids that were detected in more than one genome in the dataset. Plasmids DVT1294_4, DVT1284_2, EC00661_2, and EC00635_3 harbor *bla*$_{\text{CTX-M-15}}$ (red); plasmids EC00675_2, EC00763_3, and EC00637_2 harbor *bla*$_{\text{CTX-M-27}}$ (orange); plasmids DVT1252_7, DVT1006_4, and EC00617_2 harbor *bla*$_{\text{SHV-12}}$ (purple); and plasmid DVT1001_2 harbors *bla*$_{\text{CTX-M-2}}$ (brown).

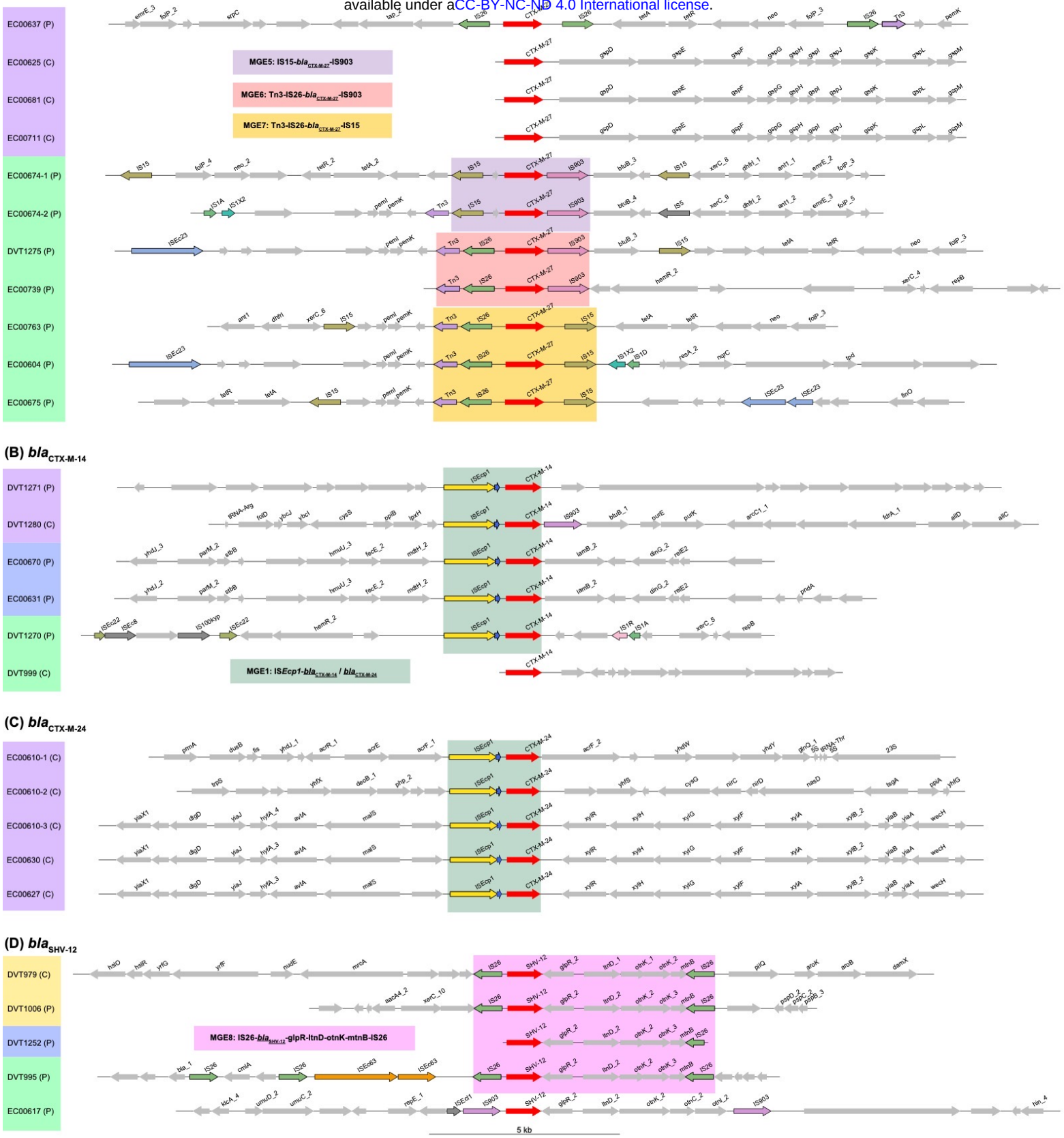



**Figure S5. Regions flanking ESBL genes among ST131 *E. coli* isolates.** (A-D) Genomic context of different ESBL-carrying MGEs is shown. Isolate names are shaded based on their phylogenetic clade assignments (clade A=purple; subclade C1=green; subclade C2=blue; clade B=yellow). The genomic context of each sequence is indicated (C=chromosome, P=plasmid) and ESBL genes are colored red. Genes were annotated with Prokka, and genes with predicted functions are labeled. Genes associated with MGEs and transposases are highlighted with black outlines, and are colored if found in more than one region. Regions that were used for MGE classification are shaded in each panel.