Genomic analysis uncovers a phenotypically diverse but genetically homogeneous Escherichia coli ST131 clone circulating in unrelated urinary tract infections

Journal:	*Journal of Antimicrobial Chemotherapy*
Manuscript ID:	Draft
Manuscript Type:	Original Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Clark, Gemma; Nottingham Trent University, Pathogen research group Paszkiewicz, Konrad; Exeter University, College of Life and Environmental Sciences Hale, James; University College Cork, Environmental research Institute Weston, Vivienne; Nottingham University Hospitals, Microbiology Constantinidou, Chrystala; University of Birmingham, School of Biosciences Penn, Charles; University of Birmingham, School of Biosciences Achtman, Mark; University College Cork, Environmental Research Institute McNally, Alan; Nottingham Trent University, Pathogen Research Group
Keywords:	Escherichia coli, ESBLs, Genomics, Molecular epidemiology

Genomic analysis uncovers a phenotypically diverse but genetically homogeneous *Escherichia coli* ST131 clone circulating in unrelated urinary tract infections

Gemma Clark[1], Konrad Paszkiewicz[2], James Hale[3], Vivienne Weston[4], Chrystala Constantinidou[5], Charles Penn[5], Mark Achtman[3], Alan McNally[1*]

[1]Pathogen research group, Nottingham Trent University, Nottingham NG11 8NS

[2]College of Life and Environmental Sciences, University of Exeter, EX4 4QD

[3]Environmental Research Institute, University college Cork, Lee Road, Cork, Ireland

[4]Nottingham University Hospitals NHS Trust, Derby road, Nottingham

[5]School of Biosciences, University of Birmingham, Birmingham B15 2TT

[*]All correspondence to: Dr Alan McNally, Pathogen Research Group, School of Science and Technology, Nottingham Trent University, Clifton Lane, Nottingham NG11 8NS.

Email: alan.mcnally@ntu.ac.uk

Phone: 0044 115 8483324

Keywords: *E. coli* ST131; genome; phenotype; phylogeny; ExPEC; ESBL

Running title: ST131 genomics

Abstract

Objectives: To determine variation at the genome level in *Escherichia coli* ST131 clinical isolates previously shown to be phenotypically diverse.

Methods: Ten ST131 isolates extensively characterised in previous studies were genome sequenced using combinations of Illumina and 454 sequencing technology. Whole genome comparisons and phylogenetic comparisons were then performed across the strain set and with other closely related ExPEC strain types

Results: *E. coli* ST131 is over-represented in a collection of clinical isolates, and there is large phenotypic variation amongst isolates. Genome sequencing of a selection of non-related clinical isolates in contrast shows almost no genomic variation between ST131 strains, and *E. coli* ST131 shows evidence of a genetically monomorphic pathogen showing similar evolutionary trend to hypervirulent *Clostridium difficile*.

Conclusions: A dominant circulating clone of *E. coli* ST131 has been identified in unrelated clinical urine samples in the UK. The clone splits into two distinct subgroups on the basis of antimicrobial resistance levels and carriage of ESBL plasmids. This provides the most comprehensive snapshot to date of the true molecular epidemiology of ST131 clinical isolates.

Introduction

Urinary tract infections (UTI) are among the most common bacterial infectious diseases in the world, with an estimated 20% of women over the age of 18 suffering from a UTI in their lifetime [1]. Of those infections among otherwise healthy women, some 80% are caused by *Escherichia coli* [1]. All *E. coli* which cause UTI are classified on the basis that that they are a pathovar of the species which cause extra-intestinal disease, and are termed Extra-Intestinal pathogenic *E. coli*, or ExPEC. This classification works on the basis that a subset of *E. coli* exist which are capable of causing infectious disease at sites other than the intestine, and also incorporates avian pathogenic *E. coli*, septicaemia *E. coli* and new born meningitis *E. coli* [2]. In addition to the disease burden of UTI, ExPEC are also of significant importance due to the levels of antimicrobial resistance observed in isolates. Epidemiological studies show resistance to front line antibiotics such as ciprofloxacin and trimethoprim in as many as 20% - 45% of isolates tested in large cohorts across Europe, North, and South America [1, 3]. Of greater concern is the observed level of extended spectrum β-lactamase (ESBL) gene carriage in ExPEC [4, 5]. ESBL render bacteria resistant to multiple antimicrobials including the cephalosporins, meaning that only carbapenems remain as a drug of choice for treatment of some ESBL producers.

Molecular epidemiological analysis of ESBL positive ExPEC isolates by multi-locus sequence typing (MLST) has uncovered the emergence of an apparently dominant sequence type of ExPEC among UTI and other extra-intestinal infections, namely ST131. The sequence type is composed of *E. coli* O25b:H4 strains, and has been implicated as the major cause of dissemination of the CTX-M-15 class of ESBL gene [6]. ST131 isolates are also unusual in that they counter the accepted dogma that bacteria exhibiting high levels of antimicrobial resistance do so at the expense of a fitness advantage which results in decreased pathogenesis [1]. ST131 strains reportedly exhibit increased pathogenesis [7] associated with high levels of virulence associated gene carriage (VAG) [8], and have been implicated in large scale disease outbreaks [9, 10], leading to the hypothesis that ST131 is a pandemic ExPEC clone [11]. Previous work by our group investigated the organisms present in polymicrobial and monomicrobial urine samples, and uncovered the presence of *E. coli* exhibiting high levels of antimicrobial resistance and a hyper-invasive phenotype in *in vitro* cell culture experiments [3]. Further characterisation of the isolates showed that ST131 was the dominant strain type within the collection, that the ST131 strains were responsible for the high levels of antimicrobial resistance observed in the collection, and that there was variation in VAG profile between strains, with no specific VAG profile associated with ST131 strains [12]. To address the dichotomy between the observations from our previous studies and the suggestion that ST131 is a pandemic clone with specific traits, we investigated a group of ExPEC ST131 strains isolated from the urinary tracts of elderly patients from a mixture of both Hospital and community settings. Mapping of phenotypes against sequence type showed wide variation in exhibited phenotypes within the ST131 cluster. An improved quality draft genome sequence for one isolate, and draft genome sequences for a further nine isolates, showed no variation in gene content between the isolates. SNP based phylogeny shows the strains are genetically homogenous and that the isolates sub-cluster according to antimicrobial resistance and ESBL plasmid carriage. In combination our data shows the circulation of a dominant ST131 clone among unrelated cases of urinary tract infection in the UK, and raises the question of ST131 being a monomorphic pathogen who's selection is being driven by antimicrobial resistance.

Materials and methods:

Strains

One hundred and fifty *E. coli* were isolated from 250 culture plates collected at random from Nottingham University Hospitals (NUH) between October 2008 and June 2009 as part of a larger study into UTI causative agents [3]. *E. coli* ST131 strains and outlier ST12 strain selected for genome sequencing analysis are listed in table 1. Antibiotic susceptibility profiles, VAG carriage and in vitro invasion phenotypes were obtained as described previously [3, 12].

Multi Locus Sequence Typing

MLST was performed using the Achtman typing scheme (http://mlst.ucc.ie/mlst/dbs/E*coli*), adhering to the protocols published on the website. Bionumerics v.6.5 was used to generate a minimum spanning tree from non-concatenated sequences of the 7 alleles.

Genome sequencing of ST131 strains

For Illumina sequencing, genomic DNA was sheared into 300bp fragments, libraries prepared using the Illumina Tru-Seq Genomic library preparation kit and multiplexed using 6bp index sequences into a single lane. They were then sequenced using paired-end 72bp reads on an Illumina GAIIx platform using SCS 2.8 software. Samples were filtered using the FASTX toolkit (v 0.13) and remapped using Bowtie 0.12.7 to the UTI89 reference genome using a minimum insert length of zero and a maximum of 600. Other parameters were left as default. Suspected PCR duplicate SNPs were called using the SAMtools (0.18) utilities. A minimum depth of 8x was required before SNP calling at a given position could begin. De novo assemblies of the genome were performed using Velvet 1.0.18 and the VelvetOptimiser script (version 2.1.7). ORFs were called using a minimum size of 102 nt and blast and PFAM scans were run of the resulting ORFs. For 454 sequencing genomic DNA for each strain was sheared into approximately 8 kb fragments. Paired-end libraries were prepared, according to the Roche/454 Sequencing 8 kb Paired End Library Preparation Method Manual. Emulsion

PCRs were performed for enrichment titration and sequencing according to the manufacturer (Roche/454 Sequencing). Titanium sequencing for each library was performed on a 454 GS-FLX apparatus. The reads were assembled and scaffolded using Newbler version 2.5.

Comparative genomics

Genome sequence annotation was manually curated using Artemis and Blast functions. Genomes were compared in a pairwise fashion using BRIG [13]. To determine levels of SNP variation in the reference genomes for strains belonging to ST95 and ST73, genome sequences were aligned using progressiveMauve, and the SNP data exported to a spreadsheet. SNPs were manually curated to remove any ambiguous calls, and to remove insertions.

Whole genome based phylogeny reconstruction

Phylogeny of ST131 in relation to the UTI89 reference genome and the outlier ST12 strain was performed by aligning genome sequences using progressiveMauve [14] and the common core genome extracted using the stripSubsetLCBs script. Bayesian phylogeny was inferred using ClonalFrame [15] from the 50% consensus of 10 runs with 10,000 iterations following a burn in phase of 10,000 iterations, with the quality of each run manually checked using Tracer. Phylogenetic trees were produced and edited using FigTree.

Results:

Phenotypic variation in *E. coli* ST131 isolated from unrelated clinical UTI cases.

As part of a wider study into the microbial population of urinary tract infections in elderly patients, 150 *E. coli* were isolated from 250 unrelated clinical urinary tract samples belonging to patients aged 70 or over across the East Midlands area of the United Kingdom [3], which contains a population of around 5 million people. During this study variation in antimicrobial resistance and epithelial cell invasion was demonstrated within the *E. coli* isolates [3]. To assess the phenotypic variation that existed within the *E. coli* ST131 population, and compare against other ST types, epithelial cell invasion (Fig 1A) and antimicrobial resistance levels (Fig1B) were overlaid against a minimum spanning tree (MST) of the ExPEC isolates. The overlaid MST show variation in phenotypes within the ST131 isolates. There is variation in levels of antimicrobial resistance within the group and in the ability to exhibit the high cell invasion phenotype described previously in this group of isolates [3].

Improved quality draft genome of an ST131 isolate uncovers common ExPEC genomic traits.

In an attempt to further characterise *E. coli* ST131, one isolate from our strain collection was chosen for high quality draft genome sequencing. Strain UTI18 was chosen as it is highly antimicrobial resistant with average invasion levels (table 1), and was sequenced using a combination of Illumina and 454 sequencing. UTI18 is equivalent to the recently published NA114 *E.. coli* ST131 genome sequence [20] in that it contains no discernible "novel" regions which would account for increased fitness or pathogenicity when compared to the available genomes of ExPEC isolates (fig 2). The pathogenicity island (PAI) which encodes *cnf*, haemolysin and the intact *pap* operon has been deleted, as has the *sfa* fimbrial operon, and there is a transposase insertion in the *fimB* gene of the Type I fimbriae operon. UTI18 does contain a fully intact High Pathogenicity Island encoding the yersiniabactin locus, and also contains two flagella encoding regions. The first region is identical to the flagella operons present in other publicly available ExPEC genome sequences, whilst the second is a truncated version of Flag-2 found in the enteroaggregative *E. coli* O42 genome sequence [16], and in ExPEC strain UMN026, as well as the publicly available *E. coli* O111 and O26 EPEC genome sequences.

Comparative analysis of regions outside the accessory virulome of ExPEC highlighted differences in metabolic pathways encoding genes between ST131 and the other publicly available ExPEC genome sequences. The *idnK* and *idnDOTR* operons, encoding for the L-idonate catabolism pathway are fully deleted in ST131. The pathway is a subsidiary pathway for Gluconate metabolism in *E. coli* and is also termed the GntII system [17]. The ancestral *asc* operon encoding a combined arbutin/salicin/cellobiose uptake and metabolism pathway is also affected by deletions of *ascF* and *ascB*, the PTS transporter enzyme and phospho-beta-glucosidase enzyme respectively which are transcribed from a single promoter [18]. Also deleted are the putative ABC transporter genes *yddA* and *yddB*, and the *yrhA* and *yrhB* genes present in a region encoding for both the GntI gluconate uptake and metabolism pathway and the GGT small peptide transporter [19].

Illumina sequencing of unrelated ST131 clinical isolates suggests circulation of a genetically homogeneous clone.

In order to confirm that the high quality draft genome sequence strain was representative of our population, a further nine ST131 strains isolated from unrelated clinical samples and displaying varied phenotypic traits (table 1) were sequenced using the Illumina GAIIx (eight isolates) or 454 (one isolate), with draft de novo assemblies produced. Stepwise BLAST comparisons using BRIG [13] were performed of the draft de novo assembled genome sequences against our improved quality UTI18 genome sequence, and against the recently announced NA114 genome sequence of an Indian ST131 isolate [20]. These comparisons showed no strain specific insertions or deletions of accessory mobile islands within our strain set, but 2 regions differing from NA114 which were annotated as fragments of plasmids (Fig 3). This heterogeneity is not observed in ST73 and ST95, where there is variation in carriage of pathogenicity islands between strains within the complex.

SNP profiling of the strains which were Illumina sequenced was performed against the publicly available UTI89 reference genome sequence as well as the genome sequence of UTI48, an ST12 isolate from our strain collection. SNP profiling shows that the ST131 strains are genetically homogeneous. There were a total of 15,060 SNPs conserved between the ST131 strains compared to UTI89, with 1,324 SNPS between the ST131 strains, 371 of which are non-synonymous. Strain UTI226 was the most divergent amongst our cohort but had only 460 strain-specific SNPs, with the remaining strains having only 10 – 60 strain specific SNPs. Such low level SNP variation is unreported in *E. coli* and rare in enterobacteriaceae in general, and is more akin to monomorphic highly pathogenic and host restricted subsets of species such as *Salmonella* Typhi [21]. To ascertain if this monomorphic observation was common across *E. coli* ST complexes the level of SNP variation was determined in ST95 and ST73 using the publicly available genome sequences of strains from those complexes (Table 2). ProgressiveMauve alignments were performed, and the extracted SNP file manually curated to remove deletions and ambiguous SNP calls. The results showed 14, 413 SNPs between the three ST95 strains, and 9, 059 SNPs between the two ST73 strains. Mapping of the ST131 specific SNPs against the UTI18 genome showed that the SNPs were not randomly distributed suggesting that recombination has played a significant role in the emergence of our ST131 clone (fig 4). The metabolic operons *glc*, *glp, ytf,* and *tre* are all ST131 SNP hotspots, as is the *fim* operon. Conversely both flagella operons and the HPI show no SNPs at all.

Whole genome alignments were performed on our ten ST131 isolates, NA114, the ST12 outlier strain UTI48, and the reference genome UTI19, and phylogeny reconstructed using Clonalframe [15]. When strain phenotypes were mapped against the resulting phylogenetic tree (Fig 5) there was a split between CTX-M-15 plasmid positive isolates and non CTX-M negative strains, which also mirrored levels of antimicrobial resistance observed in the isolates. In addition the CTX-M positive strains also had identical VAG profiles using a multiplex PCR detection method [12]. There was no correlation with invasive phenotype, community or hospital acquisition, or clinical recurrence of UTI in patients the original strain was isolated from.

Discussion

Extra-intestinal pathogenic *E. coli*, ExPEC, are an extremely diverse group of organisms classified according to disease pathology. A number of *E. coli* genotypes, as defined by Multi-locus sequence typing classification, are capable of causing extra-intestinal infection [22], and genome sequencing combined with comparative genomics of ExPEC isolates has shown no classical genetic blueprint for an *E. coli* to become a successful ExPEC strain [23-25]. Recently *E. coli* ST131 has emerged as the most frequent ST isolated from human clinical cases of ExPEC infection, leading to it being tagged as an emerging pandemic *E. coli* [11, 26]. In particular ST131 ExPEC have garnered interest for their role in the rapid spread of the CTX-M15 Extended spectrum β-lactamase determinant, conferring multiple drug resistance to extra-intestinal infectious agents [9, 12, 27]. This emergence of a dominant ExPEC strain type is in contrast to the hypothesis that there is no set genomic blueprint for a successful ExPEC strain.

Previous work by our group showed variation in phenotypic characteristics among ExPEC isolated from elderly patients [3]. Molecular epidemiology on this group of strains uncovered a large proportion of ST131 isolates within the population exhibiting variation in virulence associated gene carriage [12]. In this study we further investigated this apparent variation in phenotypes of ST131 by mapping phenotypic traits against a minimum spanning tree of our ExPEC population. Our data corroborates the current ST131 literature reporting significant increases of isolation of the organism from extra-intestinal infections with ST131 the most common ST isolated in our ExPEC population. Our data also shows variation in phenotypes observed within our ST131 population, correlating with our earlier observation of variation in virulence associated gene carriage within the cohort [12]. Most reports of ST131 populations have focussed on the likelihood of an emerging clone, and focus on the ST131 isolates carrying CTX-M variants, however our previous work [12] combined with data presented here, show that clinical ST131 isolates are phenotypically heterogeneous, and that this lies beyond simple variation in carriage of the CTX-M encoding plasmids.

In order to investigate if this phenotypic variation was mirrored in genotypic variation we genome sequenced ten ST131 strains isolated from unrelated clinical episodes in elderly patients living in a catchment area of approximately 5 million people. The strains were chosen to represent the wide spectrum in phenotypic and virulence gene carriage profiles observed in our population. In addition our data was compared to the recently announced NA114 genome, an ST131 strain isolated in India [20]. The striking observation from our data is the lack of variation across the genomes of the ST131 strains isolated. Previous ExPEC genome studies have shown heterogeneity in genome architecture and content among strains, including between strains of the same sequence type as exemplified by UTI89, APEC01, and S88 which are all ST95, and ABU83972 with CFT073 which are both ST73 [23, 28]. In contrast all ten of our ST131 isolates show characteristics of being genetically monomorphic, with no variation in accessory genome content beyond carriage of antimicrobial resistance genes and associated plasmids. This would suggest the ST131 circulating in our population is not participating in accessory genome flux and that is a stable clone. Similarly there were no obvious discriminatory genomic signatures such as novel or unusual pathogenicity islands or virulence associated genes, although the absence of the *sfa* and *pap* fimbrial operons and deletion in the *fimB* gene from all isolates merits further study for biological relevance. Previous work by our group highlighted that both *sfa* and *pap* operons were statistically less frequently found in ExPEC strains exhibiting an increased virulence phenotype [12]. The absence of P fimbriae in our clinical ST131 sequenced isolates, and the insertion in *fimB* raises questions on the true virulent nature of our ST131 isolates. The relevance of these mutations and the true virulence of our ST131 strains is the focus of current work in our group.

The genetically monomorphic nature of ST131 was further confirmed when phylogenetic analyses were performed based on whole genome data. SNP analysis of the ten ST131 genome sequences showed low level polymorphism of 1324 SNPs between strains, (typically 10-60 strain specific SNPs with one strain containing 386) in contrast to the 14, 413 SNPs between ST95 genome sequenced strains and the 9, 059 between ST73 genome sequenced strains. Indeed the levels of variation between our ST131 strains are similar to those observed in intra-strain variation during human bladder passage using ABU83972, where some 29 SNPs occurred accompanied by one large deletion and four smaller deletions [29]. Such low levels of variation are only seen in monomorphic, highly niche restricted and pathogenic subsets of species such as *Salmonella* Typhi and hypervirulent *C. difficile* O27 where inter-strain SNP variation levels of 1,964 [21] and 1,874 [30] SNPs respectively have been reported. Both these organisms are subtypes of their respective species which have independently evolved into highly-pathogenic variants, and in the case of *S.* Typhi accompanied by gene loss and niche restriction. The inclusion of the Indian NA114 isolate in the middle of our phylogenetic tree raises the possibility that ST131 is a globally disseminated monophyletic clone which is evolving into subclades on the basis of antimicrobial resistance.

Together the data from our study provides evidence of the circulation of a genetically monomorphic *E. coli* ST131 clone as a dominant strain isolated from unrelated clinical cases. To our knowledge this is the first time such a phenomenon has been reported for a sequence type of *E. coli*, where most studies focus on pathotypes encompassing diverse sequence types.

In order to determine the emergence of ST131 from a common environment to dominant human pathogen a full genome level investigation of a contemporaneous strain set separated geographically, temporally and by source reservoir is required, in conjunction with comparative studies of closely related strain types and more distant ExPEC relatives. This would allow detailed Bayesian analysis of clonal expansion of the ST131 with accurate dating, and provide clues as to the triggers for the evolution of pathogenic lineages of *E. coli,* particularly the role of antimicrobial resistance and ESBL carriage in driving evolutionary selection of ST131. Such informative clues will be of great value not just in understanding the emergence of ST131, but also how new dominant pathogenic variants of *E. coli*, such as the recent O104 epidemic, arise.

Acknowledgements:

We would like to thank Nicholas Thomson and Thomas Connor of the Wellcome Trust Sanger Institute for advice and guidance during genome and phylogenetic analysis, and appraisal of the manuscript, and Xavier Didelot for advice with the stripSubsetLCBs algorithm.

Funding:

This work was supported by a Dowager Countess Eleanor Peel Trust grant awarded to A. M. and grant 05/FE1/B882 from the Science Foundation Ireland to M. A.

Transparency Declaration:

The authors declare no competing or financial interests in this work

References

1. Foxman B. The epidemiology of urinary tract infection. *Nat Rev Urol.* 2010; **7:** 653-660.

2. Russo TA, Johnson JR. Proposal for a new inclusive designation for extraintestinal pathogenic isolates of *escherichia coli*: ExPEC. *J Infect Dis.* 2000; **181:** 1753-4.

3. Croxall G, Weston V, Joseph S et al. Increased human pathogenic potential of *escherichia coli* from polymicrobial urinary tract infections in comparison to isolates from monomicrobial culture samples. *J Med Microbiol.* 2011; **60:**102-109.

4. Livermore DM, Hawkey PM. CTX-M: Changing the face of ESBLs in the UK. *J Antimicrob Chemother.* 2005; **56:** 451-4.

5. Pitout JDD, Nordmann P, Laupland KB et al. Emergence of enterobacteriaceae producing extended-spectrum β-lactamases (ESBLs) in the community. *J Antimicrob Chemother.* 2005; **56:** 52-9.

6. Peirano G, Pitout JDD. Molecular epidemiology of *escherichia coli* producing CTX-M β-lactamases: The worldwide emergence of clone ST131 O25:H4. *Int J Antimicrob Agents.* 2010; **35:** 316-21.

7. Clermont O, Lavollay M, Vimont S et al. The CTX-M-15-producing *Escherichia coli* diffusing clone belongs to a highly virulent B2 phylogenetic subgroup. *J Antimicrob Chemother.* 2008; **61:** 1024-8.

8. Coelho A, Mora A, Mamani R, et al. Spread of *Escherichia coli* O25b:H4-B2-ST131 producing CTX-M-15 and SHV-12 with high virulence gene content in Barcelona (Spain). *J Antimicrob Chemother.* 2011; **66:** 517-26.

9. Johnson JR, Johnston B, Clabots C et al. *Escherichia coli* Type ST131 as the major cause of serious multidrug-resistant *E. coli* infections in the united states. *Clin Infect Dis.* 2010; **51:** 286-294.

10. Lau SH, Kaufmann ME, Livermore DM et al. UK epidemic *Escherichia coli* strains A–E, with CTX-M-15 b-lactamase, all belong to the international O25:H4-ST131 clone. *J Antimicrob Chemother.* 2008; **62:** 1241-4.

11. Rogers BA, Sidjabat HE, Paterson DL. *Escherichia coli* O25b-ST131: A pandemic, multiresistant, community-associated strain. *J Antimicrob Chemother.* 2011; **66:** 1-14.

12. Croxall G, Hale J, Weston V, et al. Molecular epidemiology of Extra-Intestinal Pathogenic *E. coli* isolates from a regional cohort of elderly patients highlights prevalence of ST131 strains containing increased antimicrobial resistance in both community and hospital care settings. *J Antimicrob Chemother.* In Press.

13. Alikhan NF, Petty NK, Ben Zakour NL, Beatson SA. BLAST ring image generator (BRIG): Simple prokaryote genome comparisons. *BMC Genomics.* 2011; **12:** 402.

14. Darling AE, Mau B, Perna NT. progressiveMauve: Multiple genome alignment with gene gain, loss and rearrangement. *PLoS One.* 2010; **5:** e11147.

15. Didelot X, Falush D. Inference of bacterial microevolution using multilocus sequence data. *Genetics.* 2007; **175:** 1251-66.

16. Chaudhuri RR, Sebaihia M, Hobman JL, et al. Complete genome sequence and comparative metabolic profiling of the prototypical enteroaggregative *Escherichia coli* strain 042. *PLoS One.* 2010; **5:** e8801.

17. Bausch C, Ramsey M, Conway T. Transcriptional organization and regulation of the L-idonic acid pathway (GntII system) in *Escherichia coli. J Bacteriol.* 2004; **186:** 1388-97.

18. Hall BG, Xu L. Nucleotide sequence, function, activation, and evolution of the cryptic asc operon ofE*escherichia coli* K12. *Mol Biol Evol.* 1992; **9:** 688-706.

19. Chen SL, Hung CS, Xu J et al. Identification of genes subject to positive selection in uropathogenic strains of *Escherichia coli*: A comparative genomics approach. *Proc Nat Acad Sci US A.* 2006; **103:** 5977-82.

20. Avasthi TS, Kumar N, Baddam R, N, Jadhav S, Ahmed N. Genome of multidrug-resistant uropathogenic *Escherichia coli* strain NA114 from India. *J Bacteriol.* 2011; **193:** 4272-3.

21. Holt KE, Parkhill J, Mazzoni CJ, et al. High-throughput sequencing provides insights into genome variation and evolution in *Salmonella* Typhi. *Nat Genet.* 2009; **40:** 987-993.

22. Wirth T, Falush D, Lan R et al. Sex and virulence in *Escherichia coli*: An evolutionary perspective. *Mol Microbiol.* 2006; **60:** 1136-51.

23. Rasko DA, Rosovitz MJ, Myers GSA et al. The pangenome structure of *Escherichia coli*: Comparative genomic analysis of *E. coli* commensal and pathogenic isolates. *J Bacteriol.* 2008; **190:** 6881-93.

24. Lloyd AL, Rasko DA, Mobley HLT. Defining genomic islands and uropathogen-specific genes in uropathogenic *Escherichia coli. J Bacteriol.* 2007; **189:** 3532-46.

25. Bielaszewska M, Dobrindt U, Gartner J et al. Aspects of genome plasticity in pathogenic *escherichia coli. Internat J Med Microbiol.* 2007; **297:** 625-39.

26. Kim J, Hong SG, Bae IK et al. Emergence of *Escherichia coli* sequence type ST131 carrying both the blaGES-5 and blaCTX-M-15 genes. *Antimicrob Agents Chemother.* 2011; **55:** 2974-5.

27. Zong Z, Yu R. *bla*CTX-M-carrying *Escherichia coli* of the O25b ST131 clonal group have emerged in china. *Diagn Microbiol Infect Dis.* 2011; **69:** 228,228-231.

28. Lukjancenko O, Wassenaar TM, Ussery DW. Comparison of 61 sequenced *Escherichia coli* genomes. *Microb Ecol.* 2010; **60:** 708 - 720.

29. Zdziarski J, Brzuszkiewicz E, Wullt B et al. Host imprints on bacterial Genomes—Rapid, divergent evolution in individual patients. *PLoS Pathog.* e1001078.

30. He M, Sebaihia M, Lawley TD, et al. Evolutionary dynamics of *Clostridium difficile* over short and long time scales. *Proc Natl Acad Sci U S A.* 2010; **107:** 7527-7532.

31. Welch RA, Burland V, Plunkett G et al. Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. *Proc Nat Acad Sci USA.* 2002; **99:** 17020-4.

32. Johnson TJ, Kariyawasam S, Wannemuehler Y, et al. The genome sequence of avian pathogenic *Escherichia coli* strain O1:K1:H7 shares strong similarities with human extraintestinal pathogenic *E. coli* genomes. *J Bacteriol.* 2007; **189:** 3228-36.

33. Schneider G, Dobrindt U, Brüggemann H, et al. The pathogenicity island-associated K15 capsule determinant exhibits a novel genetic structure and correlates with virulence in uropathogenic *Escherichia coli* strain 536. *Infect Immun.* 2004; **72:** 5993-6001.

34. Levine MM, Bergquist EJ, Nalin DR, et al. *Escherichia coli* strains that cause diarrhoea but do not produce heat-labile or heat-stable enterotoxins and are non-invasive. *Lancet.* 1978; **1:** 1119-22.

Table 1. Strains sequenced as part of this project

Strain	*ST*	*Patient source*	*Antibiotic resistance*											*CTX-M*	*Invasion (cfu/ml)*	*VAG profile*
			AMP	*PIP/ TAZ*	*RAD*	*CTX*	*CAZ*	*MEM*	*GEN*	*AMC*	*TMP*	*CIP*	*NIT*			
UTI18	*131*	*Community*	*R*	*S*	*R*	*R*	*R*	*S*	*R*	*S*	*R*	*R*	*R*	*CTX-M-15*	*1.20E+03*	*PAI, fimH, fyuA, iutA, traT, kpsMT II, K5*
UTI188	*131*	*Community*	*R*	*S*	*S*	*S*	*S*	*S*	*S*	*S*	*R*	*R*	*S*	*-*	*3.22E+03*	*papC, papG allele II, papG II, III, PAI, papA, fimH, traT,*
UTI226	*131*	*Hospital*	*S*	*S*	*S*	*S*	*S*	*S*	*S*	*S*	*S*	*S*	*S*	*-*	*9.44E+03*	*PAI, fimH, ibeA, fyuA, traT, kpsMT II, K5*
UTI306	*131*	*Community*	*R*	*R*	*R*	*R*	*R*	*S*	*R*	*R*	*R*	*S*	*R*	*CTX-M-15*	*7.22E+04*	*PAI, papA, fyuA, iutA, traT, kpsMT II, K5*
UTI32	*131*	*Hospital*	*R*	*S*	*R*	*R*	*R*	*S*	*S*	*S*	*R*	*R*	*S*	*CTX-M-15*	*4.17E+04*	*papC, papG allele II, papG II, III, PAI, papA, fimH, afa/draBC, fyuA, iutA, traT*
UTI423	*131*	*Community*	*R*	*S*	*R*	*S*	*S*	*S*	*S*	*S*	*R*	*R*	*R*	*-*	*1.01E+05*	*PAI, fimH, fyuA, iutA, traT, kpsMT II, K5*
UTI524	*131*	*Community*	*R*	*S*	*R*	*R*	*R*	*S*	*S*	*S*	*R*	*R*	*R*	*CTX-M-15*	*7.20E+01*	*PAI, fimH, fyuA, iutA, traT, kpsMT II, K5*
UTI570	*131*	*Community*	*R*	*S*	*S*	*S*	*S*	*S*	*S*	*S*	*S*	*R*	*S*	*-*	*7.83E+05*	*PAI, fimH, fyuA, iutA*
UTI587	*131*	*Community*	*R*	*S*	*R*	*R*	*R*	*S*	*R*	*S*	*R*	*R*	*S*	*CTX-M-15*	*1.34E+05*	*PAI, fimH, fyuA, iutA, traT, kpsMT II, K5*
UTI62	*131*	*Community*	*R*	*S*	*R*	*R*	*R*	*S*	*S*	*S*	*R*	*R*	*R*	*CTX-M-15*	*1.05E+02*	*PAI, fimH, fyuA, iutA, traT, kpsMT II, K5*
UTI48	*12*	*Community*	*R*	*S*	*S*	*S*	*S*	*S*	*S*	*S*	*R*	*R*	*R*	*-*	*1.09E+03*	*PAI, fimH, fyuA, kpsMT II, K5*

The ST131 strains selected for sequencing represent the variation within the ST131 study population with regards to antibiotic resistance, CTX-M-15 possession, ability to invade epithelial cells and virulence associated gene (VAG) possession. Antibiotic abbreviations; AMP - Ampicillin (32 µg/ml), RAD – Cephradine (32 µg/ml), CTX – Cefotaxime (1 µg/ml), CAZ – Ceftazidime (1 µg/ml), PIP-TAZ – Piperacillin/Tazobactam (85 µg/ml), TMP – Trimethoprim (2 µg/ml), CIP – Ciprofloxacin (4 µg/ml), GEN – Gentamicin (2 µg/ml), AUG – Augmentin (32 µg/ml), NIT – Nitrofurantoin (32 µg/ml), MEM – Meropenem (2 µg/ml). VAG abbreviations; *papC*, *papG*, *papA* – regions within the pap operon which codes for P pili, *afa/draBC* – DR adhesins, PAI – CFT073 pathogenicity island marker, *fimH* – mannose specific adhesion subunit of type 1 fimbriae, *fyuA* – yersiniabactin, *iutA* – aerobactin, *traT* – serum resistance, *kpsMT* II – group II capsule synthesis, K5 –K5 capsule synthesis.

Table 2. Publicly available reference genomes used in this study

Strain	ST	Strain History	Reference
UTI89	95	Uncomplicated cystitis	[19]
CFT073	73	Acute pyelonephritis	[31]
ABU83972	73	Asymptomatic bacteriuria	[29]
Apec01	95	Poultry collibacilosis	[32]
E. coli 536	92	Acute pyelonephritis	[33]
E. coli HS	46	Human commensal	[34]
IAI39	62	Urinary tract infection	http://www.genoscope.cns.fr/spip/-Escherichia-fergusonii-coli-.html
S88	95	Neonatal meningitis	http://www.genoscope.cns.fr/spip/-Escherichia-fergusonii-coli-.html
UMN026	597	Urinary tract infection	http://www.genoscope.cns.fr/spip/-Escherichia-fergusonii-coli-.html

Figure legends:

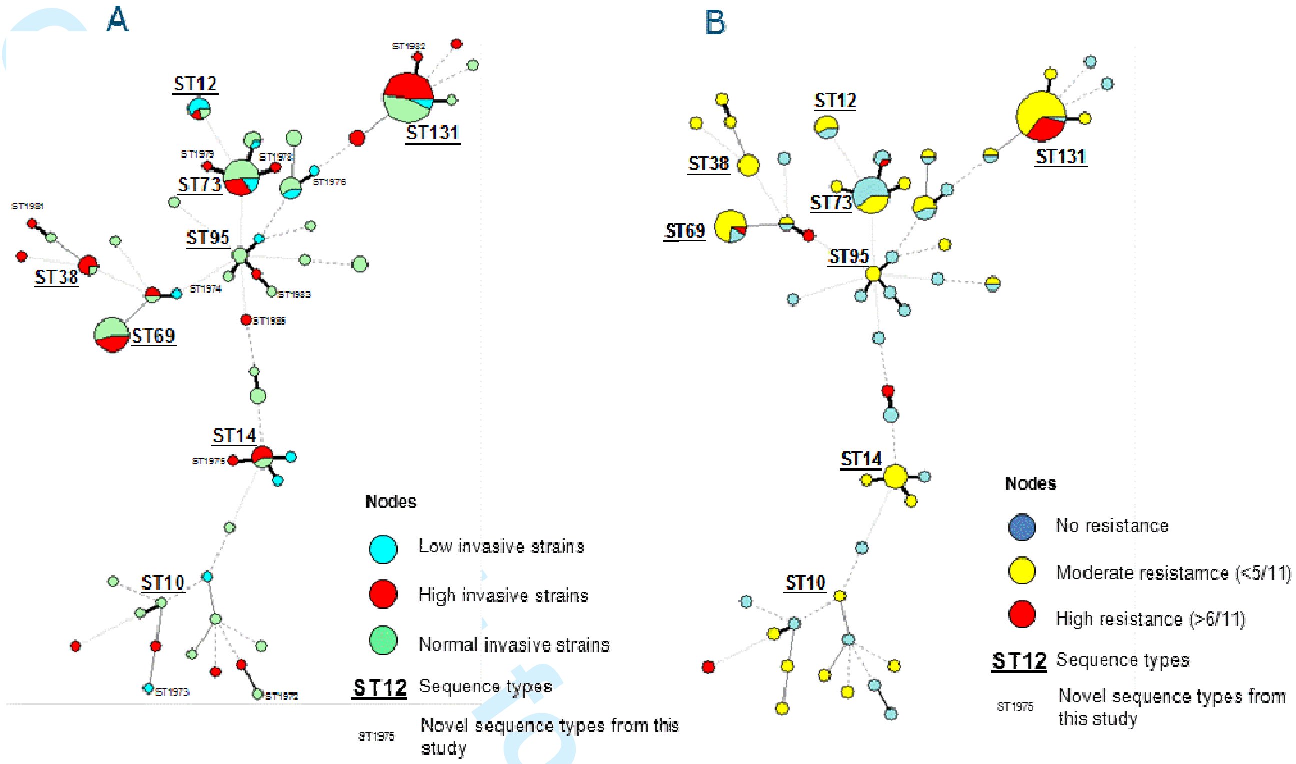
Figure 1. Minimum spanning trees of ExPEC isolated from our previous studies, with phenotypes (A – in vitro epithelial cell invasion; B – antimicrobial resistance) overlaid.

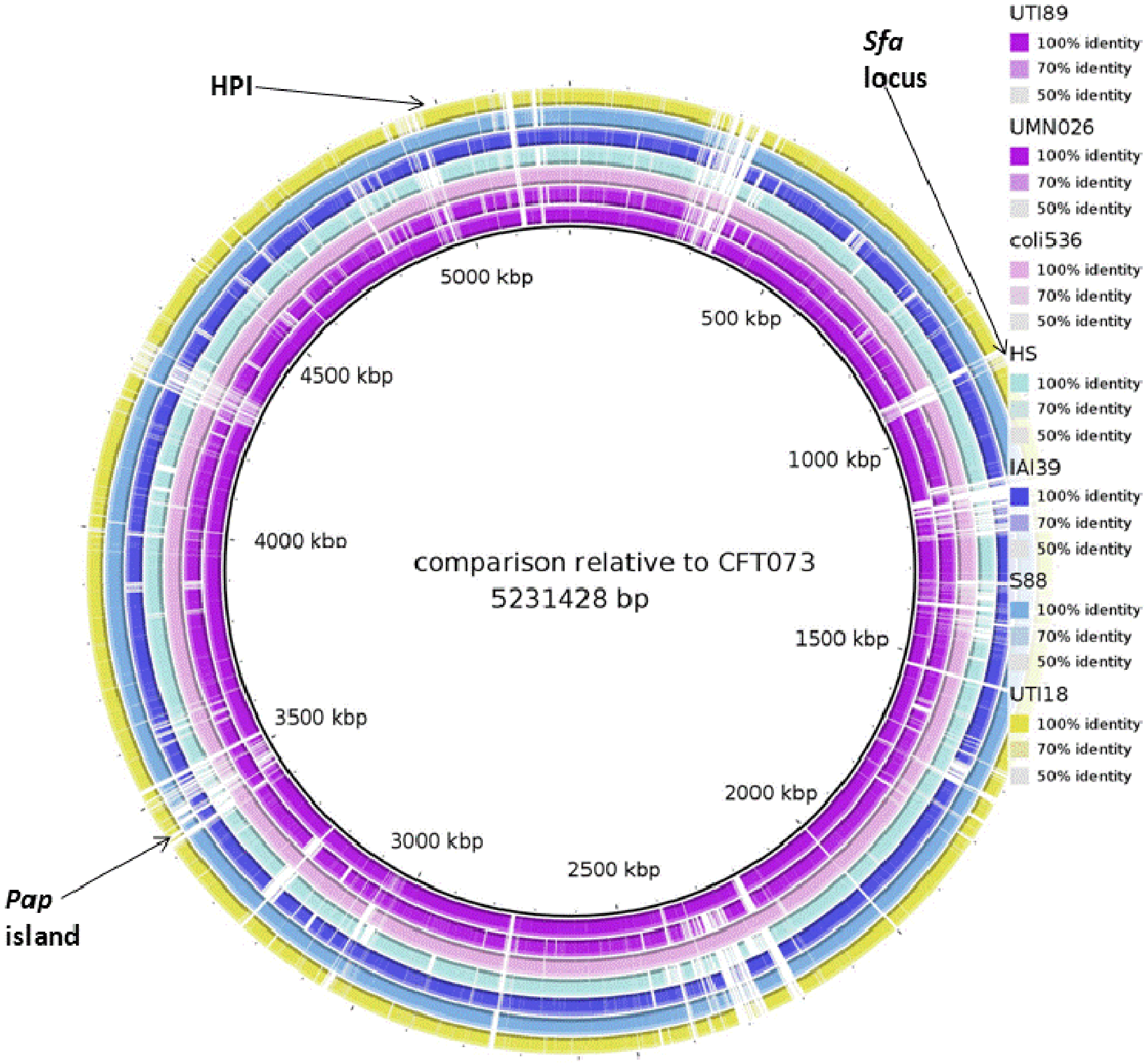
Figure 2. BRIG alignment of *E. coli* ST131 UTI18 genome with publicly available ExPEC reference genomes. The location of the *sfa* and *pap* islands deletions are annotated, as is the location of the intact HPI. The comparisons are made relative to *E. coli* CFT073.

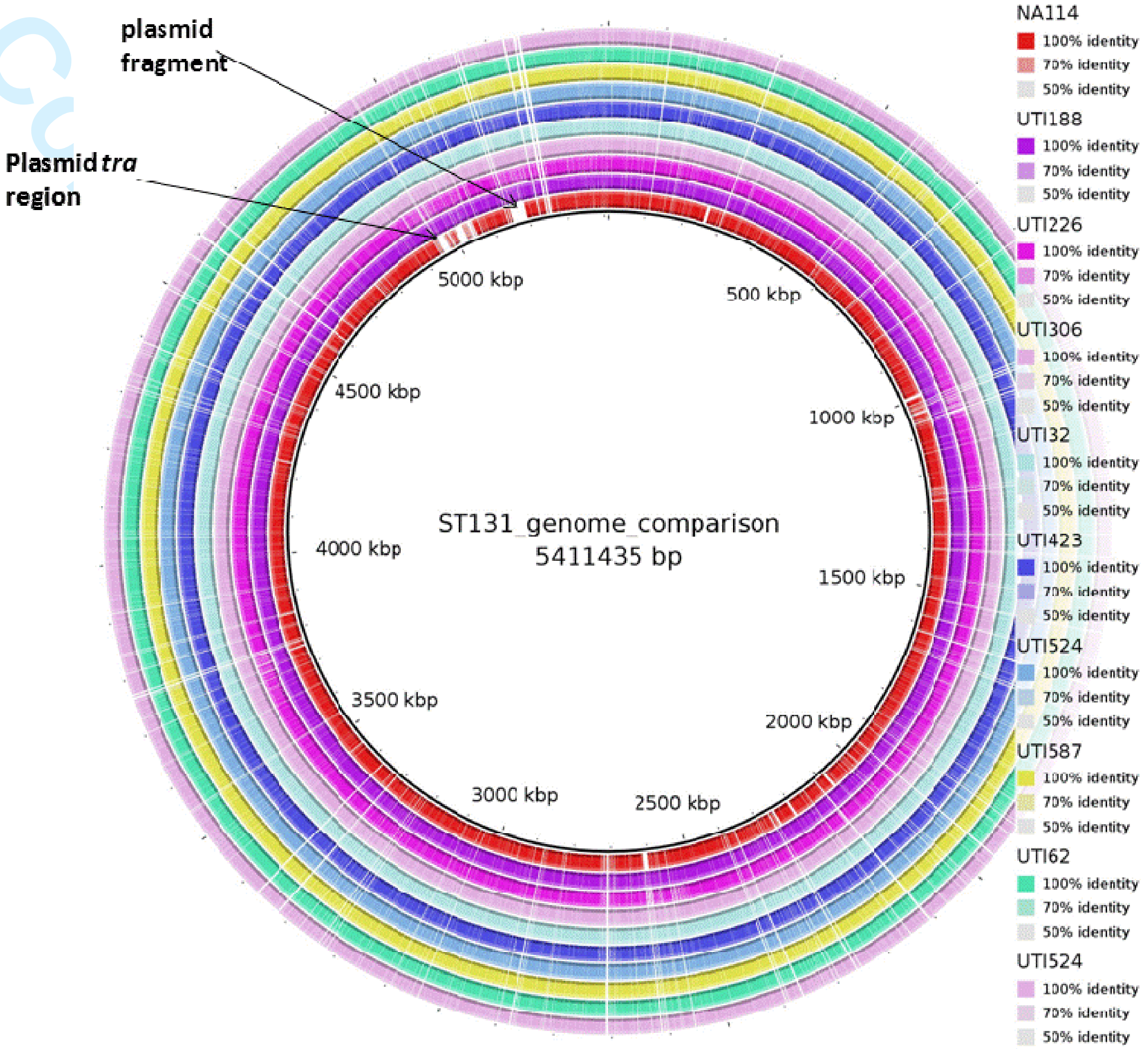
Figure 3. BRIG alignment of the nine ST131 genomes sequenced using Illumina GAIIx. The comparisons are made relative to the Indian ST131 strain NA114, which is missing plasmid DNA fragments found in our ST131 isolates annotated on the circular diagram.

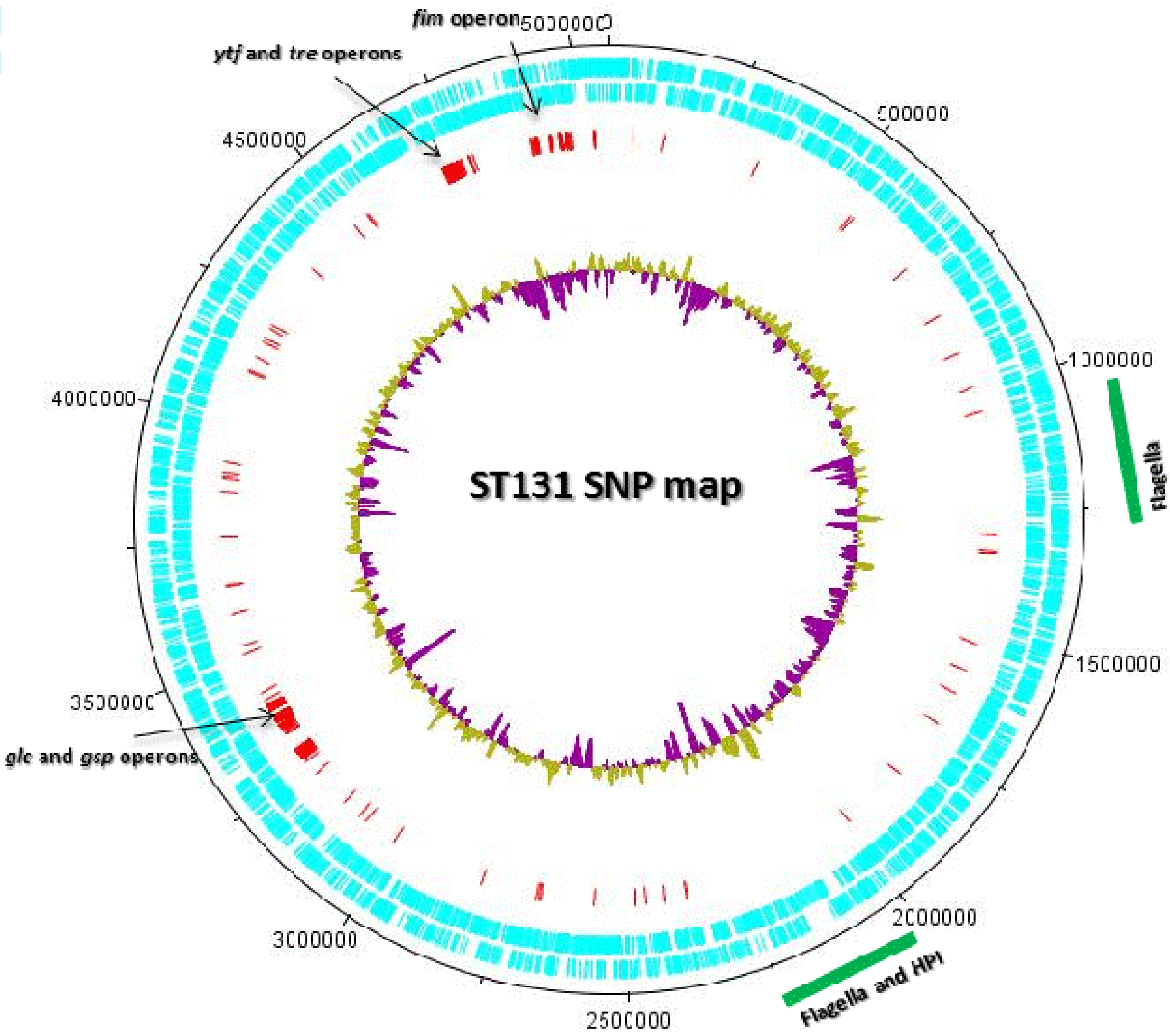
Figure 4. Circular diagram showing the location of ST131 specific SNPs relative to the UTI18 genome. The innermost ring is GC content. The two outermost rings are CDS found on the coding and complementary strand. Red marks illustrate the position of ST131 SNPs. The SNP hotspot regions are annotated with arrows. The two regions completely free of SNPs are marked by rectangles outside of the circular diagram

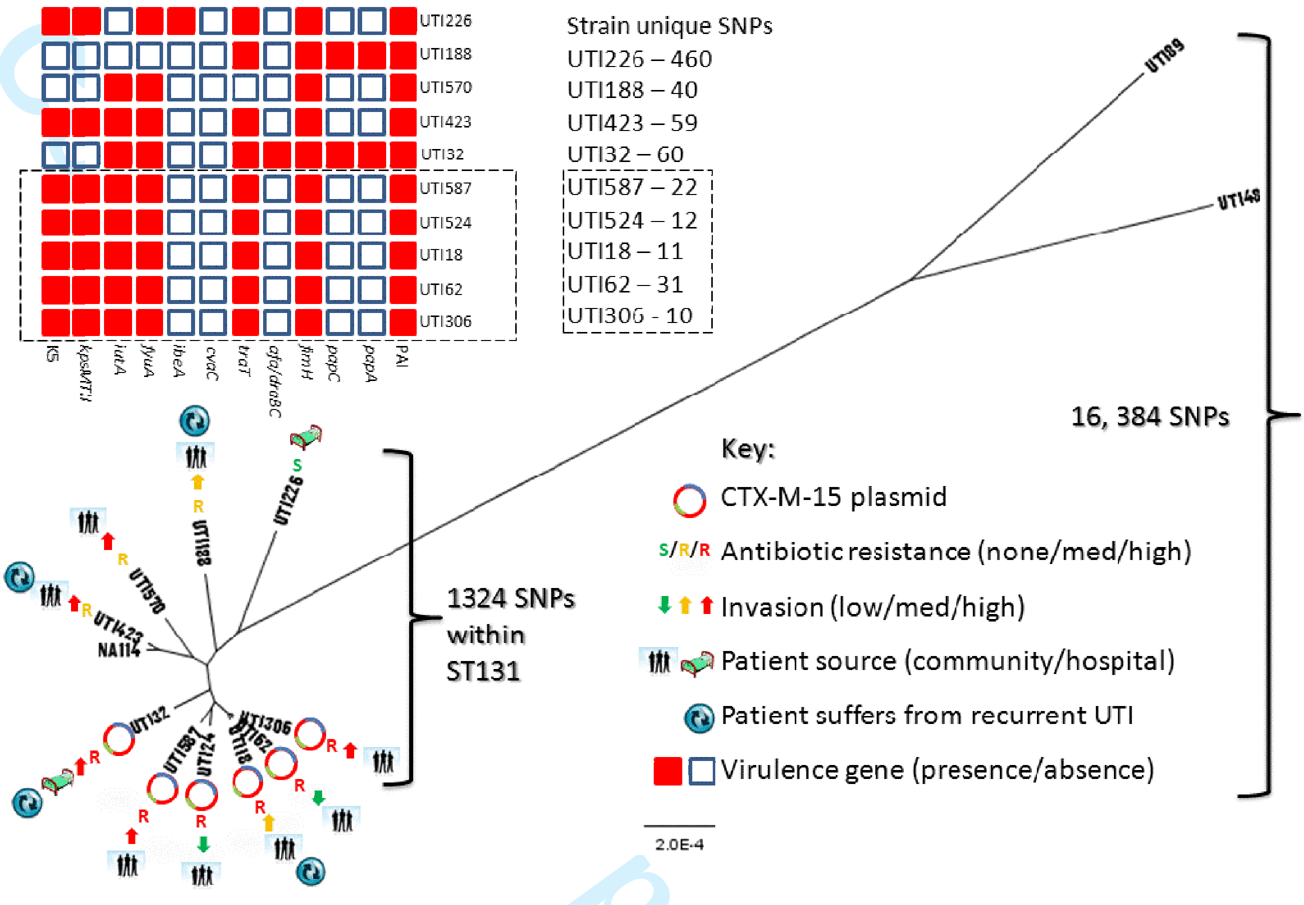
Figure 5. Phylogenetic tree of the ten ST131 isolates sequenced in this study relative to the outlier ST12 strain UTI48, and the reference strain used to assemble sequences and call SNPs, UTI89. The number of discriminatory SNPs are numerically presented. The Virulence associated gene carriage profile of the isolates is also presented by presence (red block) or absence (white block) of genes as determined by PCR in a previous study [17]. Strain characteristics are mapped on to the tree according to the key.


A
ST12
ST131
ST1982
ST1979
ST1978
ST73
ST1976
ST1981
ST95
ST38
ST1974
ST1983
ST1985
ST69
ST14
ST1975
ST10
ST1973
ST1972
Nodes
Low invasive strains
High invasive strains
Normal invasive strains
ST12 Sequence types
ST1975 Novel sequence types from this study
B
ST12
ST131
ST38
ST73
ST69
ST95
ST14
ST10
Nodes
No resistance
Moderate resistamce (<5/11)
High resistance (>6/11)
ST12 Sequence types
ST1975 Novel sequence types from this study


HPI
Sfa
locus
Pap
island
comparison relative to CFT073
5231428 bp
500 kbp
1000 kbp
1500 kbp
2000 kbp
2500 kbp
3000 kbp
3500 kbp
4000 kbp
4500 kbp
5000 kbp
UTI89
100% identity
70% identity
50% identity
UMN026
100% identity
70% identity
50% identity
coli536
100% identity
70% identity
50% identity
HS
100% identity
70% identity
50% identity
IAI39
100% identity
70% identity
50% identity
S88
100% identity
70% identity
50% identity
UTI18
100% identity
70% identity
50% identity


plasmid fragment
Plasmid *tra* region
5000 kbp
500 kbp
4500 kbp
1000 kbp
ST131_genome_comparison
5411435 bp
4000 kbp
1500 kbp
3500 kbp
2000 kbp
3000 kbp
2500 kbp
NA114
100% identity
70% identity
50% identity
UTI188
100% identity
70% identity
50% identity
UTI226
100% identity
70% identity
50% identity
UTI306
100% identity
70% identity
50% identity
UTI32
100% identity
70% identity
50% identity
UTI423
100% identity
70% identity
50% identity
UTI524
100% identity
70% identity
50% identity
UTI587
100% identity
70% identity
50% identity
UTI62
100% identity
70% identity
50% identity
UTI524
100% identity
70% identity
50% identity


fim operon
5000000
ytf and tre operons
500000
4500000
1000000
4000000
ST131 SNP map
Flagella
1500000
3500000
glc and gsp operons
2000000
3000000
2500000
Flagella and HPI



	K5	kpsMTII	iutA	fyuA	ibeA	cvaC	traT	afa/draBC	fimH	papC	papA	PAI
UTI226	🟥	🟥	□	🟥	🟥	□	🟥	□	🟥	□	□	🟥
UTI188	□	□	□	□	□	□	🟥	□	🟥	🟥	🟥	🟥
UTI570	□	□	🟥	🟥	□	□	□	□	🟥	□	□	🟥
UTI423	🟥	🟥	🟥	🟥	□	□	🟥	□	🟥	□	□	🟥
UTI32	□	□	🟥	🟥	□	□	🟥	🟥	🟥	🟥	🟥	🟥
UTI587	🟥	🟥	🟥	🟥	□	□	🟥	□	🟥	□	□	🟥
UTI524	🟥	🟥	🟥	🟥	□	□	🟥	□	🟥	□	□	🟥
UTI18	🟥	🟥	🟥	🟥	□	□	🟥	□	🟥	□	□	🟥
UTI62	🟥	🟥	🟥	🟥	□	□	🟥	□	🟥	□	□	🟥
UTI306	🟥	🟥	🟥	🟥	□	□	🟥	□	🟥	□	□	🟥

Strain unique SNPs
UTI226 – 460
UTI188 – 40
UTI423 – 59
UTI32 – 60
UTI587 – 22
UTI524 – 12
UTI18 – 11
UTI62 – 31
UTI306 - 10

Key:
○ CTX-M-15 plasmid
S/R/R Antibiotic resistance (none/med/high)
⬇⬆⬆ Invasion (low/med/high)
Patient source (community/hospital)
Patient suffers from recurrent UTI
🟥□ Virulence gene (presence/absence)