Defining nosocomial transmission of *Escherichia coli* and antimicrobial resistance genes: a genomic surveillance study

Running title: Detecting nosocomial *Escherichia coli* transmission

Catherine Ludden Ph.D.*, Francesc Coll Ph.D., Theodore Gouliouris Ph.D,Olivier Restif Ph.D, Beth Blane B.Sc., Grace A. Blackwell Ph.D., Narender Kumar Ph.D., Plamena Naydenova B.Sc., Charles Crawley F.R.C.Path, Nicholas M. Brown F.R.C.Path, Julian Parkhill Ph.D, Sharon J. Peacock F.R.C.Path

Affiliations:

From the Department of Infection Biology, Faculty of Infectious & Tropical Diseases, London School of Hygiene & Tropical Medicine, London, UK (C.L. and F.C.); Department of Medicine, University of Cambridge, Cambridge, UK (C.L, T.G., B.B., N.K., P.N. and S.J.P.); Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK (T.G., C.C., N.M.B, S.J.P); Public Health England, London UK, (N.M.B, S.J.P); Department of Veterinary Medicine, University of Cambridge, Cambridge, UK (O.R. and J.P.); EMBL-EBI, Wellcome Trust Sanger Institute, Hinxton, UK (G.A.B).

*To whom correspondence should be addressed: cl636@medschl.cam.ac.uk

Keywords: *Escherichia coli*, transmission, genomics, epidemiology, nosocomial

ABSTRACT

Background: E. coli is a leading cause of bloodstream infections. Developing interventions to reduce this burden requires an understanding of the frequency of nosocomial transmission, but available evidence is limited. This study aimed to detect and characterise transmission of *E. coli* and associated plasmids in a hospitalised cohort.

Methods: Genomic surveillance of *E. coli* was conducted in a prospective observational cohort study of hospitalised adult patients over 6 months in Cambridge, England. Stool samples were collected from study participants on admission, weekly and discharge. We sequenced multiple *E. coli* colonies (median=5) from each stool. A genetic threshold to infer *E. coli* transmission was defined by maximum within-host SNP diversity and the probability of drawing observed pairs of between-patient isolates at different SNP thresholds.

Findings: We obtained and cultured 376 stools from 149 patients, of which 152 stools from 97 patients grew *E. coli*. We identified extensive diversity in the bacterial population (90 sequence types, STs), and mixed *E. coli* ST carriage in almost half of patients (26%, 13% and 6% patients carried 2, 3 or ≥4 STs, respectively). Using a 17 SNP cut-off we identified 10 clusters (defined as ≥2 cases) involving 20 patients. The largest cluster contained 7 patients, while 4 patients were linked to multiple clusters. Half of cases in the 10 clusters also had a strong epidemiological link to another patient in the cluster. A minority of all patients (17/149, 11%) carried extended-spectrum beta-lactamase (ESBL)-producing *E. coli*, the most common of which was $bla_{CTX\text{-}M\text{-}15}$ (12/17, 71%). Long-read sequencing revealed that $bla_{CTX\text{-}M\text{-}15}$ was often integrated into the chromosome, with little evidence for plasmid-mediated transmission. Seven patients developed *E. coli* bloodstream infection, four with identical strains in those in stool; two of these had documented nosocomial acquisition.

Interpretation: We provide evidence of bacterial transmission and endogenous infections during routine care by integrating genomic and epidemiological data and through determination of a genetic similarity cut-off informed by within-host diversity in the population studied. Our findings challenge single colony-based investigations, and the paradigm of plasmid spread in this setting.

Funding: UK Department of Health, Wellcome Trust, UK National Institute for Health Research

RESEARCH IN CONEXT

Evidence before this study

We searched PubMed for studies published up to March 2020 using the terms "*Escherichia coli*", "whole genome sequencing", "transmission" AND "hospital". We excluded reviews and kept articles where whole-genome sequencing had been applied to study *E. coli* transmission in human populations in a hospital setting (15 out of 75). Twelve of the fifteen studies were focused on carbapenem or colistin resistance and were not further evaluated. Of the three remaining studies, one focused on the national epidemiology of a single clone (ST410) in Denmark and was based on 127 whole-genome-sequenced isolates. Five possible regional outbreaks were identified using ≤10 SNPs. In a second study performed in Denmark, whole genome multi locus sequence typing (wgMLST) was used to distinguish between epidemiologically related and unrelated isolates of extended-spectrum beta-lactamase (ESBL) producing *E. coli*. Isolates obtained from the same patient, belonging to the same wgMLST, and cultured within a time window of 30 days were defined as epidemiologically related. In a third study, transmission of *E. coli* among haematology and oncology patients was performed in German hospitals using core genome MLST and closely related isolates were defined as a maximum of 10 allele differences

Added value of this study

Our findings capture what happens during routine care, beyond much of the current bacterial genomics literature which largely focuses on outbreak investigations. This study shows that surveillance/outbreak investigations based on single colonies are likely to underestimate transmission events and the diversity of antimicrobial susceptibility profiles present in a sample. Our study also adds to the existing evidence on suitable methods to determine transmission events. We established a genome-based SNP threshold to infer *E. coli* transmission in the population studied by comparing SNP distances of isolates from the same host and combining this with epidemiological data. We identified transmission clusters involving predominately patients with non-ESBL *E. coli*, which were likely to be missed by other investigations focused on antimicrobial-resistant *E. coli*. Using long-read sequencing, we were able to accurately study

the transmission of antimicrobial resistance genes conferring resistance to cephalosporin drugs (extended-spectrum beta-lactamases) and plasmids. Whilst *E. coli* from patients carried the same genes conferring resistance, they were rarely carried on the same plasmids as those found in other patient samples. This would not have been identified using short-read sequencing. By comparing *E. coli* isolates from blood and stool from individual patients we identified indistinguishable isolates from both, suggesting endogenous infection.

Implications of all the available evidence

Our study highlights polyclonal *E. coli* colonisation, the pathogenesis of extraintestinal *E. coli* infection (endogenous vs. exogenous) and the clinical relevance of *E. coli* transmission in the hospital setting. Our findings challenge the paradigm of plasmid spread, at least for *E. coli* in this setting. Interventions to reduce *E. coli* bacteraemia should aim to prevent endogenous infections as this was observed as a major source of infections.

INTRODUCTION

Escherichia coli is a leading cause of bloodstream and urinary tract infections, a proportion of which are healthcare-associated[1]. Rates of *E. coli* bloodstream infections have undergone a marked increase in numerous countries, including in England where the incidence increased from 60.4 per 100,000 population (32,309 reported cases) in 2012-2013 to 77.7 per 100,000 population (43,209 reported cases) in 2018-2019[1]. This rate has increased from 76.6 per 100,000 population to 125.1 per 100,000 population at Cambridge University Hospitals in the same time period[2]. This problem is compounded by a global increase in the frequency of *E. coli* infections caused by strains that are resistant to numerous antibiotics, which are associated with excess morbidity, mortality, longer hospital stays and higher healthcare costs[3-6].

Interventions to support a reduction in healthcare-associated *E. coli* bloodstream infection require an understanding of the frequency of nosocomial transmission, but available evidence is limited. Previous studies that addressed this using bacterial sequencing, an essential methodology that provides the necessary genetic resolution, have been conducted on either small patient cohorts or solely extended-spectrum beta-lactamase (ESBL)-producing *E. coli* or specific STs, which is likely to under-represent transmission of *E. coli* overall (that is, including both ESBL and non-ESBL *E. coli*)[7-9]. Furthermore, transmission studies require an understanding of the frequency of mixed strain *E. coli* carriage, and within-host diversity of the same lineage.

Here, we report the findings from genomic surveillance of *E. coli* in a cohort of adult hospitalised patients over 6 months, performed to understand within-host diversity, and transmission of *E. coli* and associated plasmids encoding antimicrobial resistance genes.

METHODS

Study design and participants

We evaluated *E. coli* acquisition and transmission during six months of a prospective observational study of consecutive patients admitted to two adult haematology wards at the Cambridge University Hospitals NHS Foundation Trust (CUH) in England (13 May to 13 Nov 2015). All in-patients on the two haematology wards at Addenbrooke's Hospital, aged 16 years and over, who were being treated for hematologic malignancies were eligible to be included in the study. Patients under the age of 16 years and patients not being treated for hematologic malignancies were excluded. Patients were enrolled following informed written consent. This patient cohort was previously studied to investigate the transmission of *Klebsiella pneumoniae* and *Enterococcus faecium*[10,11]. The study protocol was approved by the National Research Ethics Service (ref: 14/EE1123 and 12/EE/0439) and the Cambridge University Hospitals NHS Foundation Trust Research and Development Department (ref: A093285 and A092685).

Procedures

Hospital admission and bed movement data were extracted electronically using the hospital bed tracking system. Admission to the same bay, room or ward at the same time or within 7 days was classified as a strong epidemiological link; admission in the same ward separated by more than 7 days or to the study hospital but to different wards (regardless of dates of admission) was classified as a weak epidemiological link; and no epidemiological link was reported if neither of these occurred. After patients were enrolled into the study, stool samples were provided by participants on admission and then every week until in-patient discharge and cultured for *E. coli* for the purpose of this prospective study. Stool samples were enriched in Tryptic Soy Broth (Sigma, Dorset, UK) prior to culture and directly cultured onto Brilliance UTI Chromagar (Oxoid, Basingstoke, UK) to detect all *E. coli,* and onto Brilliance™ ESBL agar (Oxoid, Basingstoke, UK) to detect ESBL-producing *E. coli*. Up to 15 *E. coli* (10 putative ESBL and 5 non-ESBL) colonies cultured from each stool sample were selected for sequencing (See appendix p 2). For stools that grew less than this, all of the available *E. coli* colonies were sequenced.

Blood cultures were collected to identify endogenous infection in participants and to further understand the genetic diversity of *E. coli* causing bloodstream infections in the haematology population. Hospital acquired and healthcare associated infections were based on definitions by Friedman *et al* [12]. To determine if patients acquired *E. coli* after admission, we identified instances where patients changed from stool culture-negative to culture-positive, and where existing *E. coli* carriers appeared to acquire a new ST during admission. Evidence of acquisition within and between hospital admissions was investigated using hospital admissions data. During the 6-month study, blood cultures were obtained from the study patients. In addition, for blood cultures positive for *E. coli* retrieved from patients residing in the haematology wards in the 12 months before (May 2014 –May 2015) and six-months after the study (November 2015-May 2016), one colony was obtained for sequencing from the culture in the freezer archive. See appendix p 2 for additional details on culture protocols, selection of colonies and antimicrobial susceptibility testing. The number of invasive infections per 1,000 admissions was determined based on the number of admissions of recruited patients to haematology wards.

Sequencing and bioinformatic analyses

DNA was extracted, libraries prepared and sequenced on an Illumina HiSeq2000 with 125-cycle paired-end reads. Following quality control, genomes were assembled using SPAdes 3.11.0, mapped against the *E. coli* reference strain (GenBank: LT632320) using SMALT v0.7.4 (http://www.sanger.ac.uk/science/tools/smalt-0)[13]. The core genome was derived using Roary version 1.7.1 using the 'don't split paralogs' option[14]. Whole-genome alignments were created by calling nucleotide alleles along the LT632320 reference genome and pairwise SNP distances in core-genome alignments using pairsnp (https://github.com/gtonkinhill/pairsnp) (See appendix pp 3-4 for a detailed description). The core-genome coordinates are publicly available (https://doi.org/10.6084/m9.figshare.13227746.v1). The SNP distances cannot be compared to whole-genome SNP differences, but should be comparable to distances reported using the same reference genome (*E. coli* LT632320) and coordinates used in this study. The genomes of multiple *E. coli* isolated from the same patient were used to ascertain *E. coli* within-host diversity for all participants and

subsequently determine an appropriate threshold to define transmission of *E. coli* STs between patients. The analysis was limited to instances where different patients shared the same ST. The upper limit for a SNP cut-off was provisionally established from the maximum within-host diversity (the number of core genome differences in isolates of the same ST from the same patient), which defines the upper limit of transferable diversity from one person to another.

Detection of antimicrobial resistance and mobile elements

A detailed description of the methods applied to detected antimicrobial resistance genes in all isolates and the rationale for selecting isolates for long-read sequencing to investigate plasmid sharing between patients is provided in appendix p 5. In brief, *E. coli* genomes from all enrolled participants were screened for acquired genes encoding antibiotic resistance using Antibiotic Resistance Identification By Assembly (ARIBA)[15]. Chromosomal mechanisms of fluoroquinolone resistance were identified by screening isolates for the presence of associated amino acid changes in the quinolone resistance-determining regions of *gyrA* and *parC* alleles [16,17]. To investigate if plasmids encoding ESBL genes were shared between patients during the study, one $bla_{\text{CTX-M-15}}$ and $bla_{\text{CTX-M-14}}$ isolate from each ST per positive sample were selected for long-read sequencing. *in silico* PCR was used to perform plasmid incompatibility group/replicon typing[18]. Geneious (version 11.1) was used for manual annotation and visualisation of complete plasmid sequences. ISFinder (https://isfinder.biotoul.fr) and BLAST was used to identify insertion sequences and transposon fragments. Blast comparisons, visualised in ACT were used for plasmid comparisons (see Appendix p 5).

Statistical analysis

The significance of differences on the number of positive or negative cultures between patients who received antimicrobials in the previous 30 days or not was assessed with a two-tailed Fisher exact test using the fisher.test function from R package stats (v3.6.3). We used a two-tailed Mann-Whitney test to assess the difference in the number of sequenced colonies per stool sample between samples with one ST and samples with multiple STs. This test was performed using the wilcox.test function from R

package stats (v3.6.3). Plots were created using ggplot2 version 3.3.1. To further validate the SNP threshold, we used a statistical approach that compared a range of cut-off values (appendix p 5-6)

Role of the funding source

The funder of the study played no role in study design, data collection, data analysis, data interpretation, or writing of the report.

RESULTS

Study design, patients and samples

We recruited 174 of the 338 adult patients (51%) admitted during the 6-month study period. Details of their characteristics at the time of enrolment are provided in Supplementary Table 1[11] (appendix p 9). Of the 174 patients, the majority(149, 86%) were able to provide one or more stool samples, with a total of 376 stool samples and a median of 3 per case (IQR 2-5). 101 patients provided two or more samples. This subset of 149 patients formed the basis for all further analyses. These 149 participants had a median age of 61 years (IQR 49-69, range 19-94), 281 admissions in total, a median of 1 admission (IQR 1-2), and stayed a median of 16 days (IQR 7 to 27 days), as described previously [10]. 97 of the 149 participants (65%) had at least one stool positive for *E. coli*, with a total of 152 positive stool samples identified. The majority of positive participants (80/97, 83%) carried non-ESBL *E. coli* only, 5/97 (5%) carried ESBL-producing *E. coli* only, and 12/97 (12%) carried both (Figure 1).

114 (77%) of 149 participants received antimicrobials in the previous 30 days and/or on enrolment. including 47 (87%) of 52 patients with negative *E. coli* stool culture, and 67 (69%) of 97 patients with a positive culture (p= 0.00036) (appendix p 6).

E. coli diversity and putative acquisition

We picked a median of five *E. coli* colonies (IQR 5-5, range 1-15, hereafter termed isolates) from each of the 152 primary stool culture plates from the 97 *E. coli* positive patients. This gave an overall total

of 970 isolates (686 non-ESBL, 284 ESBL *E. coli*), which underwent whole genome sequencing. From this we identified 90 different STs (Supplementary Table 2 and Supplementary Figure 1A, appendix p 9 and appendix p 13). The most frequent STs identified in stools were ST131, ST10 and ST69, which were isolated from 14, 9 and 8 patients (Supplementary Figure 1B, appendix p 13), respectively, and accounted for 232/970 (24%) isolates. Seventeen patients had stool samples positive for ESBL *E. coli*, with variation in the presence of genes encoding ESBL between different STs (Supplementary Table 2, appendix p 9).

To quantify the amount of within-host *E. coli* diversity, we determined the number of different *E. coli* STs identified from each patient using data on 149/152 stool samples from 94/97 patients (excluding three stools/patients from which only a single *E. coli* colony was isolated). Around half of patients (52/94, 55%) were positive for a single ST. Of the remainder, 26% (24/94), 13% (12/94) and 6% (6/94) patients carried two, three, or 4 or more STs, respectively, with a maximum of 8 STs found in a single patient (Figure 2). On a per stool analysis (unit of analysis is individual stool samples), 70% (104/149) of stools contained a single ST, and 23% (35/149) and 7% (10/149) contained two or more STs, respectively, with a maximum of 5 STs recovered from a single stool. Out of the 149 stool samples with multiple isolates sequenced, 104 (69.8%) contained isolates of the same ST, and 45 samples (30.2%) contained more than one ST. There was no statistical difference in the number of colonies picked from samples containing a single ST (median of 5 colonies (IQR 5-5)) and samples with multiple STs (median 5, IQR 5-14) ($p=0.09$, confidence interval of the difference between the medians = -2.82×10^{-5} - 7.15×10^{-7}).

We then identified STs that were isolated from stools obtained from two or more patients, which revealed that 27 STs were carried by at least two patients. This led us to question whether this represented coincidental carriage of the same ST, or transmission from one patient to another. Acquisition analysis was possible for 71/101 patients who provided at least 2 stools during the study and had a stool sample positive for *E. coli*. This demonstrated almost half of patients (30/71, 42%) had putative acquisition of one or more *E. coli* STs through a total of 50 acquisition events (Supplementary

Table 3, appendix p 9). Of the 17 patients that tested positive for ESBL-*E. coli*, 13 (76%) were positive for ESBL-*E. coli* on their first stool sample while the other 4 patients (24%) tested positive on follow up sampling indicating putative acquisition of ESBL-*E. coli* during hospitalisation.

Determining a SNP threshold to infer E. coli transmission

Having demonstrated the possibility of *E. coli* acquisition following hospital admission, we sought to use the sequence data to define a cut-off of genetic similarity between two genomes that was consistent with *E. coli* transmission in the population studied, as measured by the number of single nucleotide polymorphisms (SNPs) in the core genome. A core genome pairwise comparison of isolates from the same patient/same ST demonstrated a maximum diversity of 17 SNPs (6.8 SNPs per million bases) (Figure 3) with the exception of 3 patients that carried isolates which belonged to distinct clades of the same ST (>300 SNPs different, see Supplementary Table 4, appendix p 9). The results from the Poisson distribution indicated an upper limit of 25 SNPs (see Supplementary Figure 2, appendix p 14 for details). Having defined two putative but different cut-offs of 17 and 25 SNPs, we used epidemiological information to select the final proposed cut-off. We found that patient-pairs with a strong epidemiological link (same bay, room or ward at the same time or within 7 days) carried isolates that were up to 17 SNPs different, while patient pairs carrying isolates 17 to 25 SNPs apart did not have strong epidemiological links. In light of this, we selected a 17 SNP cut-off, appreciating that this is likely to be more specific but less sensitive than 25 SNPs.

Genetic and epidemiological links support putative acquisition and transmission of E. coli

We then applied the 17 SNP cut-off to all 970 *E. coli* isolates, reflecting a strictly genomic investigation of putative transmission. This identified 10 clusters (defined as containing 2 or more cases) involving 20 patients, 4 of whom were involved in multiple clusters (Table 1 and Supplementary Figure 3, appendix p 16). Strong epidemiological links were found between patients in 7/10 clusters (Supplementary Figure 3, appendix p 15). The two largest clusters contained 7 and 4 patients respectively, associated with two different STs (ST7095 and ST635, see Supplementary Figure 4 and

Supplementary Figure 5 [appendix pp 16-19] for phylogenetic trees and timelines for the two lineages). These STs appeared to have been acquired following admission in 6 and 2 patients respectively, further supporting hospital acquisition. The remaining 8 clusters each contained 2 patients and were associated with 8 different STs (ST69, ST131, ST443, ST648, ST1193, ST1196, ST6151 and ST7094).

Implications of E. coli carriage and transmission

A serious consequence of *E. coli* carriage is the development of bloodstream infection. This occurred in 9/174 patients during the 6-month study (5%), equating to around 32 invasive infections per 1,000 admissions (n=174 patients, 281 admissions). Characteristics of these 9 cases are shown in Supplementary Table 5, appendix pp 10-11. All 9 cases had bloodstream infection onset associated with healthcare contact (hospital acquired (n=4) or healthcare-associated (n=5)). The majority (7/9 cases) were infected by non-ESBL *E. coli*. The other 2 patients were infected by ESBL *E. coli*. Seven of the 9 patients had at least one positive stool cultured. The other two patients did not provide a stool sample. Four of the seven patients provided a stool sample before infection onset. We sequenced 100 colonies from 12 stools from the 7 patients (median 15 colonies per patient, range 5-30). The same ST was identified in both the blood and stool samples in 4 cases (ST131 (2 cases), ST95 and ST1193 (1 case each). Pairwise core genome comparison of these stool and disease-associated *E. coli* genomes demonstrated that the blood and stool isolates were very highly related (0 SNPs different).

Over a longer time-frame (May 2014-May 2016), we identified 36 additional positive blood cultures from the same two study wards (from 25 patients) with at least one *E. coli* isolate available for sequencing. The *E. coli* isolates belonged to 18 STs, with 9 (25%) of 36 isolates being ST131 and 12 (33%) producing ESBL *E. coli*

Analysis of putatively transmissible antimicrobial resistance determinants

34 (23%) of 149 patients had *E. coli* resistant to ciprofloxacin in stool isolates and mechanisms of resistance were identified (Supplementary Table 6). Identified types of ESBL genes and the STs that

carried each type are shown in Supplementary Table 2 appendix p 9 and further described in appendix p 7. We selected 31 ESBL *E coli* isolates (21 stools and 10 blood cultures) for long-read sequencing. Plasmids carrying $bla_{CTX\text{-}M\text{-}15}$ shared only segments (mostly over regions carrying antibiotic resistance genes) of high identity (potentially shared mobile genetic elements) or the isolates carried identical plasmids but were themselves only 25 SNP different and the patients that carried them had a weak epidemiological link. See appendix p 8 for a more detailed comparison of the plasmids.

$bla_{CTX\text{-}M\text{-}14}$ was plasmid-bourne (all IncB/O/K/Z) in all 5 $bla_{CTX\text{-}M\text{-}14}$ positive isolates (4 STs) from two patients (C062 and C047, see Table 2). $bla_{CTX\text{-}M\text{-}14}$ positive plasmids from patient C062 were identical (>99% identity over >99% coverage), including plasmids from two different STs, consistent with within-host plasmid sharing between STs. However, the $bla_{CTX\text{-}M\text{-}14}$ plasmids from C047 showed great diversity and were different to those found in C062. Representative $bla_{CTX\text{-}M\text{-}14}$ -carrying plasmids and plasmid comparisons are shown in Supplementary Table 6, Supplementary figure 6, and Supplementary figure 7).

DISCUSSION

Here, we extensively examined within-host diversity by serial sampling of 94 patients. This demonstrated that almost half of all patients carried more than 1 ST and over 70% of ESBL-positive patients were also positive for non-ESBL *E. coli*, indicating that surveillance/outbreak investigations based on single colonies or focussed on ESBL-producing isolates[19,20] are likely to underestimate transmission events and the diversity of antimicrobial susceptibility profiles present in a sample. A previous study of 127 genomes from eight children, seven of whom were ESBL positive, identified a median of four STs per child (range 1-10)[7]. Analysing seven ESBL-producing *E. coli* genomes from three stool samples from a single cystic fibrosis patient identified up to 3 *E. coli* STs per sample[9].

Diversity was also identified within specific STs. A maximum of 17 SNPs was detected per ST in each patient, similar to that previously reported (12 SNPs) for ST131 isolated from nursing residents [8]. To

date, few studies have investigated within-host diversity of *E. coli* using sequencing, and those that have were limited in size and/or are restricted by the inclusion of only ESBL-positive strains.

Based on genomic data, we identified that almost a third of patients appeared to acquire one or more *E. coli* STs through a total of 50 acquisition events. Three (6%) of the 50 acquisition events were due to ESBL-producing *E. coli* and in total 34 unique STs were acquired. A major strength of our study was the development of a SNP cut-off to support *E. coli* transmission in the population studied. Using a cut-off 17 SNPs we found evidence for transmission that was generally restricted to small patient clusters. In addition, we highlight the importance of investigating the transmission of non ESBL-*E. coli* as 8/10 transmission clusters identified in this study were non ESBL-*E. coli*, including the two largest clusters.

The number of *E. coli* bloodstream infections are continuing to increase annually but resistance to third-generation cephalosporins only accounts for around 14% of such infections in the UK, leading us to include both ESBL and non-ESBL *E. coli*[21]. By examining all *E. coli* positive blood cultures from the two haematology wards over a two year period we identified a diverse collection of invasive strains (19 STs) that were predominately non-ESBL producers. These results are consistent with that observed in a national survey of bloodstream infections from 2001-2012 in England where <15% of invasive isolates on an annual basis were non-susceptible to third-generation cephalosporins compared to 17% ESBL-positive identified in this study[22]. The results are in concordance with previous publications that reported ST131 as one of most frequently recovered lineages from bloodstream infections in the UK and the predominant ESBL *E. coli* lineage[22-24]. All patients with a bloodstream infection during the 6-month study had a genetically distinct strain when compared to isolates from other patients recruited to the study, but 4/7 patients had highly similar strains in their blood and stool samples, suggesting an endogenous source for the infection.

We also revealed the complexity of investigating the transmission of ESBL genes ($bla_{CTX\text{-}M\text{-}15}$ and $bla_{CTX\text{-}M\text{-}14}$). Previous studies have shown that characterisation of large plasmids (>50 kbp) from short-read genome sequence data is challenging due to the presence of repeated sequences[25]. All ESBL plasmids were fully characterised here using long-read sequencing, which provided confidence in our conclusions on plasmid structure, genetic context of ESBL genes and transmission. We found that $bla_{CTX\text{-}M\text{-}15}$ was commonly integrated into the chromosome, unlike previous studies which showed $bla_{CTX\text{-}M\text{-}15}$ to be plasmid-encoded[26]. Our data shows that antimicrobial susceptibility data and plasmid replicon typing is not sufficient to identify plasmid transmission and long-read sequencing is required to fully understand the dissemination of antimicrobial resistance genes.

Our study has several limitations. We sampled less than 50% of the patients admitted to the two haematology wards, and we did not sample the environment or healthcare workers. This would be predicted to lead to under-estimates of epidemiological links and could explain the lack of links between patients carrying highly related isolates, the lack of genetic links in putative acquisition events and the inability to identify the source of 3 putative exogenous infections. In addition, we did not sequence the full diversity of *E. coli* in stool samples (median of 5 *E. coli* colonies from each stool). This can lead to some STs being misclassified as acquired but instead may have been present at low abundance in previous samples. We observed that stool samples contained multiple STs, but and we cannot exclude that these did not contain additional STs. Future studies could sequence directly from plate sweeps to capture greater diversity within individuals. We established a SNP cut-off to infer *E. coli* transmission in this cohort of hospitalised patients. A limitation of this approach is that directionality of transmission cannot be inferred. It is also essential to combine epidemiological with genomic data to confirm definite transmission, but this cut-off restricts the number of patients requiring detailed epidemiological follow-up. In addition, the dataset and methodology described in this study are of great value to establish a SNP threshold, but more datasets from other settings would be needed to conclude a “universal” SNP cut-off.

In conclusion, the findings from our study have important implications for carriage, acquisition and transmission analyses of *E. coli*. Our study highlights polyclonal *E. coli* colonisation, the value of characterising multiple isolates per sample and the clinical relevance of *E. coli* transmission in the hospital setting. Using the diversity of the same strains from the same host from multi-pick data we defined a cut-off of clonality that led to the identification of limited nosocomial transmission of *E. coli* strains driving carriage and bloodstream infections in the hospitalised patients. Using long-read sequencing we identified diverse mechanisms of $bla_{\text{CTX-M-15}}$ and $bla_{\text{CTX-M-14}}$ carriage with no evidence of plasmid sharing between patients. High diversity was observed in bacteraemia isolates, but we identified patients with indistinguishable isolates from stool and blood suggesting an endogenous infection. Interventions to reduce the number of *E coli* bacteraemia should focus on preventing endogenous infections.

AUTHOR CONTRIBUTIONS

C.L., T.G. and C.C. were responsible for collecting samples, clinical and epidemiological data. O.R. performed statistical analysis. C.L., T.G., B.B. and P.N. isolated and identified *E. coli*. C.L., B.B. and P.N. undertook susceptibility testing and B.B and P.N. extracted genomic DNA. N.M.B. provided access to *E. coli* cultures in the routine diagnostic microbiology laboratory, and provided expert opinion relating to infection control. C.L. undertook the bioinformatic analyses with contributions from F.C and N.K. C.L and F.C. performed the epidemiological analyses. G.B annotated plasmids and created plasmid visualisations. T.G. and S. J. P. wrote the case record forms, obtained ethical and research and development approvals for the study. J.P. supervised the genomic sequencing. C.L. and S.J.P. wrote the manuscript. S.J.P. supervised and managed the study. All authors had access to the data and read, contributed and approved the final manuscript.

DISCLOSURE DECLARATION

J.P is a paid consultant of Next Gen Diagnostics.

All other authors declare no competing interests

DATA ACCESS

The sequence data generated in this study have been submitted to the NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject/) under accession number PRJEB19918 and PRJEB21499 and individual accession numbers for illumina and PacBio data are listed in Supplementary Table 2 and Supplementary Table 7, appendix p 9 and appendix p 12 respectively.

ACKNOWLEDGMENTS

We gratefully acknowledge the contribution of the nurses and healthcare workers on both wards at CUH for assistance with sample collection and the support of the ward matrons. We thank the EPIC and Clinical Informatics teams for providing patient movement data. We are grateful to the library construction, sequencing and Pathogen Informatics teams at the Wellcome Trust Sanger Institute for assistance with Illumina sequencing. This publication presents independent research supported by the Health Innovation Challenge Fund (WT098600, HICF-T5-342), a parallel funding partnership between the Department of Health and Wellcome Trust. The views expressed in this publication are those of the author(s) and not necessarily those of the Department of Health or Wellcome Trust. C.L. is a Wellcome Trust Sir Henry Postdoctoral Fellow (110243/Z/15/Z). T.G. is a Wellcome Trust Research Training Fellow (103387/Z/13/Z). F.C. is a Wellcome Trust Sir Henry Postdoctoral Fellow (201344/Z/16/Z). S.J.P. is a National Institute for Health Research (NIHR) Senior Investigator.

REFERENCES

1. Public Health England. Annual epidemiological commentary: Gram-negative bacteraemia, MRSA bacteraemia, MSSA bacteraemia and C. difficile infections, up to and including financial year April 2018 to March 2019. Available from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/843

870/Annual_epidemiological_commentary_April_2018-March_2019.pdf. Last acessed: 30 November 2020.

2. Public Health England. *Escherichia coli* (*E. coli)* bacteraemia: financial year counts and rates by acute trust and CCG, up to financial year 2018 to 2019. Available from: https://www.gov.uk/government/statistics/escherichia-coli-e-coli-bacteraemia-annual-data. Last accessed: 30 November 2020.

3. Schwaber MJ, Carmeli Y. Mortality and delay in effective therapy associated with extended-spectrum β-lactamase production in *Enterobacteriaceae* bacteraemia: a systematic review and meta-analysis. *Journal of Antimicrobial Chemotherapy* 2007; **60**(5): 913-20.

4. Tumbarello M, Sanguinetti M, Montuori E, et al. Predictors of mortality in patients with bloodstream infections caused by extended-spectrum-β-lactamase-producing *Enterobacteriaceae*: Importance of inadequate initial antimicrobial treatment. *Antimicrobial Agents and Chemotherapy* 2007; **51**(6): 1987-94.

5. Rottier WC, Ammerlaan HSM, Bonten MJM. Effects of confounders and intermediates on the association of bacteraemia caused by extended-spectrum β-lactamase-producing *Enterobacteriaceae* and patient outcome: a meta-analysis. *Journal of Antimicrobial Chemotherapy* 2012; **67**(6): 1311-20.

6. Melzer M, Petersen I. Mortality following bacteraemic infection caused by extended spectrum beta-lactamase (ESBL) producing *E. coli* compared to non-ESBL producing *E. coli*. *Journal of Infection* 2007; **55**(3): 254-9.

7. Stoesser N, Sheppard AE, Moore CE, et al. Extensive within-host diversity in fecally carried extended-spectrum-beta-lactamase-producing *Escherichia coli* isolates: Implications for transmission analyses. *Journal of Clinical Microbiology* 2015; **53**(7): 2122-31.

8. Brodrick HJ, Raven KE, Kallonen T, et al. Longitudinal genomic surveillance of multidrug-resistant *Escherichia coli* carriage in a long-term care facility in the United Kingdom. *Genome Medicine* 2017; **9**(1): 70.

9. Knudsen PK, Gammelsrud KW, Alfsnes K, et al. Transfer of a bla(CTX-M-1)-carrying plasmid between different *Escherichia coli* strains within the human gut explored by whole genome sequencing analyses. *Scientific Reports* 2018; **8**: 280.

10. Ludden C, Moradigaravand D, Jamrozy D, et al. A One Health study of the genetic relatedness of *Klebsiella pneumoniae* and their mobile elements in the East of England. *Clinical Infectious Diseases* 2019; **70**(2): 219-26.

11. Gouliouris T, Coll F, Ludden C, et al. Quantifying acquisition and transmission of *Enterococcus faecium* using genomic surveillance. *Nature Microbiology* 2020.

12. Friedman ND KK, Stout JE, McGarry SA, Trivette SL, Briggs JP, Lamm W, Clark C, MacFarquhar J, Walton AL, Reller LB, Sexton DJ. . Health Care–Associated Bloodstream Infections in Adults: A Reason To Change the Accepted Definition of Community-Acquired Infections. *Annals of Internal Medicine* 2002; **137**(10): 791-7.

13. Harris SR, Feil EJ, Holden MT, et al. Evolution of MRSA during hospital transmission and intercontinental spread. *Science* 2010; **327**.

14. Page AJ, Cummins CA, Hunt M, et al. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* 2015; **31**(22): 3691-3.

15. Hunt M, Mather AE, Sánchez-Busó L, et al. ARIBA: rapid antimicrobial resistance genotyping directly from sequencing reads. *Microbial genomics* 2017; **3**(10): e000131-e.

16. Komp Lindgren P, Karlsson A, Hughes D. Mutation rate and evolution of fluoroquinolone resistance in *Escherichia coli* isolates from patients with urinary tract infections. *Antimicrobial agents and chemotherapy* 2003; **47**: 3222-32.

17. Johnson JR, Tchesnokova V, Johnston B, et al. Abrupt emergence of a single dominant multidrug-resistant strain of *Escherichia coli*. *The Journal of Infectious Diseases* 2013; **207**(6): 919-28.

18. Carattoli A, Bertini A, Villa L, Falbo V, Hopkins KL, Threlfall EJ. Identification of plasmids by PCR-based replicon typing. *Journal of Microbiological Methods* 2005; **63**(3): 219-28.

19. Adams-Haduch JM, Rivera JI, Shutt KA, et al. Community-associated extended-spectrum β-lactamase–producing *Escherichia coli* infection in the United States. *Clinical Infectious Diseases* 2012; **56**(5): 641-8.

20. Olesen B, Hansen DS, Nilsson F, et al. Prevalence and characteristics of the epidemic multiresistant *Escherichia col*i ST131 clonal group among extended-spectrum beta-lactamase-producing *E. coli* isolates in Copenhagen, Denmark. *Journal of Clinical Microbiology* 2013; **51**(6): 1779-85.

21. PHE. English Surveillance Programme for Antimicrobial Utilisation and Resistance (ESPAUR) Report 2018-2019. Published November 2019. Available from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/843129/English_Surveillance_Programme_for_Antimicrobial_Utilisation_and_Resistance_2019.pdf. Last accessed: 30 November 2020.

22. Kallonen T, Brodrick HJ, Harris SR, 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 Research* 2017; **27**(8): 1437-49.

23. Ciesielczuk H, Jenkins C, Chattaway M, et al. Trends in ExPEC serogroups in the UK and their significance. *Eur J Clin Microbiol Infect Dis* 2016; **35**(10): 1661-6.

24. Day MJ, Hopkins KL, Wareham DW, et al. Extended-spectrum beta-lactamase-producing *Escherichia coli* in human-derived and foodchain-derived samples from England, Wales, and Scotland: an epidemiological surveillance and typing study. *The Lancet Infectious Diseases* 2019; **19**(12): 1325-35.

25. Arredondo-Alonso S, Willems RJ, van Schaik W, Schürch AC. On the (im)possibility of reconstructing plasmids from whole-genome short-read sequencing data. *Microbial Genomics* 2017; **3**(10).

26. Mshana SE, Imirzalioglu C, Hossain H, Hain T, Domann E, Chakraborty T. Conjugative IncFI plasmids carrying CTX-M-15 among *Escherichia coli* ESBL producing isolates at a University hospital in Germany. *BMC infectious diseases* 2009; **9**: 97-.

Figure legends

Figure 1: Description of study participants and *E. coli* culture positivity

Figure 2. Number of STs observed per patient (n=92)

Figure 3. Histogram of maximum pairwise SNP differences within STs observed from the same patient when at least two isolates of the same ST were identified.

The colour of the bar denotes the time span between isolates.

Table 1. Ten patient clusters based on genomic analysis of *E. coli* isolate from stool

Cluster	Patient ID	ST	*Acquired ST	SNP distance
1	C011	ST7095	Yes	1st case detected
1	C016	ST7095	Yes	2-6
1	C095	ST7095	Yes	2-3
1	C098	ST7095	Yes	0-2
1	C100	ST7095	Yes	5-7
1	C104	ST7095	Yes	2-4
1	D058	ST7095	No	1-3
2	D013	ST635	No	1st case detected
2	C100	ST635	Yes	0
2	D038	ST635	No	3
2	D045	ST635	Yes	1-2
3	C031	ST1193	No	1st case detected
3	C043	ST1193	Yes	0-2
4	C023	ST1196	No	1st case detected
4	C035	ST1196	No	0-7
5	C022	ST131	No	1st case detected
5	C027	ST131	No	0
6	C043	ST6151	No	1st case detected
6	C031	ST6151	Yes	0-2
7	C031	ST648	No	1st case detected
7	C043	ST648	Yes	0-1
8	C096	ST69	No	1st case detected
8	C100	ST69	Yes	0-1
9	C059	ST7094	No	1st case detected
9	D058	ST7094	Yes	0-1
10	C005	ST443	No	1st case detected
10	D030	ST443	No	8-11

*Patients were previously negative for *E. coli* or acquired a new ST. Where shown, the SNP distance range refers to the minimum-maximum SNPs between the isolate from that case and others in the cluster.

554 **Table 2. Plasmids encoding *bla*$_{CTX-M-15}$ or *bla*$_{CTX-M-14}$ based on PacBio sequencing**

PacBio Plasmid accession	Sample ID	Patient ID	Sample type	ST	Plasmid size (bp)	Inc Group	Phenotypic Resistance	Antimicrobial resistance genes on plasmid
LR595882	3546	B005*	Blood	648	152153	IcFIA, IncFIB, IncFII	Cxm, Czm, CoAmox, Cip, Gen, Pip/Taz	CTX-M-15; TEM-1; *aac(3)-IIa*; *dfrA17*; *sul1*; *tetB*; *mphA*; *aadA5*; *strAB*; *ermB*
LR595874	3547	B005*	Blood	648	152153	IcFIA, IncFIB, IncFII	Cxm, Czm, CoAmox, Cip, Gen, Pip/Taz	CTX-M-15; TEM-1; *aac(3)-IIa*; *dfrA17*; *sul1*; *tetB*; *mphA*; *aadA5*; *strAB*; *ermB*
LR595875	3580	B006*	Blood	131	111743	IncFIB	Cxm, Czm, CoAmox, Gen	CTX-M-15
LR595876	3550	C042*	Blood	2006	170000	IncFIA, IncFIB, IncFII	Cxm, Czm, CoAmox, Cip, Gen	CTX-M-15; OXA-1; *aac(3)-IIa*; *aac6_prime-Ib-cr*; *dfrA17*; *sul1*; *tetB*; *mphA*; *aadA5*
LR595878	3271	C025	Faeces	1723	111381	IncFIB	Cxm, Czm, Cip	CTX-M-15
LR595886	2898	C065	Faeces	131	164328	IncFIA, IncFII, IncN	Cxm, Czm, Cip	CTX-M-15; OXA-1; *aac6_prime-Ib-cr;* *dfrA17*; *sul1*; *mphA*; *aadA5*; *tetA*
LR595884	2981	C071	Faeces	131	61991	IncFIA, IncFIB	Cxm, Czm, Amk, CoAmox, Cip, Gen	CTX-M-15; OXA-1; *aac(3)-IIa*; *aac6_prime-Ib-cr*; *dfrA17*; *sul1*; *mphA*; *aadA5*; *tetA*(x2); *strAB*
LR595879	3060	C071	Faeces	131	69882	IncFIA, IncFIB	Cxm, Czm, Amk, CoAmox, Cip, Gen, Pip/Taz	CTX-M-15; OXA-1(x2); *aac(3)-IIa*; *aac6_prime-Ib-cr*; *dfrA17*; *sul1*; *mphA*; *aadA5*; *tetA*(x3); *strAB*
LR595890	2766	D038	Faeces	1723	111381	IncFIB	Cxm, Czm, Cip	CTX-M-15
LR595881	3125	D050	Faeces	7097	81285	IncFIB	Cxm, Czm	CTX-M-15; *qnrS1*; *dfrA14*; *sul2*; TEM-1
LR595877	2656	C047	Blood	156	111594	IncB/O/K/Z	Cxm, Gen	CTX-M-14; *aac(3)-IIa*; *dfrA17*; *sul1*; *mphA*; *aadA5*
LR595889	2604	C047	Blood	428	94296	IncB/O/K/Z	Cxm	CTX-M-14
LR595871	2656	C047	Blood	428	94061	IncB/O/K/Z	Cxm	CTX-M-14
LR595888	2887	C062	Faeces	3877	96306	IncB/O/K/Z	Cxm	CTX-M-14
LR595880	2978	C062	Blood	131	96305	IncB/O/K/Z	Cxm, CoAmox, Cip, Gen	CTX-M-14
LR595872	3877	C062	Faeces	3877	96306	IncB/O/K/Z	Cxm, CoAmox	CTX-M-14

555

Antimicrobial non-susceptibility detected by VITEK2 are listed in the phenotypic resistance column; Cxm, Cefotaxime; Czm, Ceftazidime; CoAmox, Co-Amoxiclav; Cip, Ciprofloxacin; Gen, Gentamicin; Pip/Taz, Piperacillin tazobactam. *refers to blood samples taken before and after the 6-month study