Systematic longitudinal survey of invasive *Escherichia coli* in England demonstrates a stable population structure only transiently disturbed by the emergence of ST131

Running title: Genome sequencing of invasive *E. coli* population

Teemu Kallonen[1], Hayley J. Brodrick[2], Simon R. Harris[1], Jukka Corander[1,3,4], Nicholas M. Brown[5,6], Veronique Martin[7], Sharon J. Peacock[1,2,6,8], Julian Parkhill[1]

[1]Wellcome Trust Sanger Institute, Hinxton, Cambridge, United Kingdom

[2] Department of Medicine, University of Cambridge, Cambridge, United Kingdom

[3]Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland

[4]Department of Biostatistics, University of Oslo, Oslo, Norway

[5]Public Health England, Clinical Microbiology and Public Health Laboratory, Addenbrooke's Hospital, Cambridge, United Kingdom

[6]Cambridge University Hospitals NHS Foundation Trust, Cambridge, United Kingdom

[7]British Society of Antimicrobial Chemotherapy, Birmingham, United Kingdom

[8]London School of Hygiene and Tropical Medicine, London, United Kingdom

Corresponding author: Teemu Kallonen, Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, CB10 1SA, United Kingdom. Tel: +441223495373 E-mail: tk9@sanger.ac.uk

Keywords: *Escherichia coli*, ESBL, ST131, genome, sequence, phylogeny

ABSTRACT

Escherichia coli associated with urinary tract infections and bacteremia have been intensively investigated, including recent work focusing on the virulent, globally disseminated, multi-drug resistant lineage ST131. To contextualise ST131 within the broader *E. coli* population associated with disease, we used genomics to analyse a systematic 11-year hospital-based survey of *E. coli* associated with bacteremia using isolates collected from across England by the British Society for Antimicrobial Chemotherapy, and from the Cambridge University Hospitals NHS Foundation Trust. Population dynamics analysis of the most successful lineages identified the emergence of ST131 and ST69 and their establishment as two of the five most common lineages along with ST73, ST95 and ST12. The most frequently identified lineage was ST73. Compared to ST131, ST73 was susceptible to most antibiotics, indicating that multi-drug resistance was not the dominant reason for prevalence of *E. coli* lineages in this population. Temporal phylogenetic analysis of the emergence of ST69 and ST131 identified differences in the dynamics of emergence, and showed that expansion of ST131 in this population was not driven by sequential emergence of increasingly resistant sub-clades. We showed that over time, the *E. coli* population was only transiently disturbed by the introduction of new lineages before a new equilibrium was rapidly achieved. Together, these findings suggest that the frequency of *E. coli* lineages in invasive disease is driven by negative frequency-dependant selection occurring outside of the hospital, most probably in the commensal niche, and that drug resistance is not a primary determinant of success in this niche.

INTRODUCTION

Escherichia coli is a common commensal of the gastrointestinal tract of humans and other vertebrates and can be isolated from soil and water. *E. coli* is also the leading cause of bloodstream infection in England, elsewhere in Europe and the United States (US) (Elixhauser et al. 2011; de Kraker et al. 2013; Gerver et al. 2015). Annual rates increased in England by 80% between 2003 and 2011 (from 16,542 to 29,777), which led to the introduction of mandatory surveillance from 2011. This documented a 10% increase between 2012/13 and 2014/15, from 32,309 to 35,676 cases (Gerver et al. 2015). The most common underlying causes for bloodstream infection in a national collection of the British Society for Antimicrobial Chemotherapy (BSAC) bacteremia resistance surveillance program during 2001–2010 related to urinary tract infection (UTI), gastrointestinal and hepatobiliary infections (Day et al. 2016).

Previous genetic studies of *E. coli* lineages associated with UTI and/or bacteremia in England and the US have reported that the most prevalent multilocus sequence types (MLST) are sequence types (ST) ST73, ST131, ST95 and ST69 (Gibreel et al. 2012; Adams-Sapper et al. 2013; Alhashash et al. 2013; Banerjee et al. 2013; Horner et al. 2014). ST131 has received particular attention following its apparent emergence in the 2000s due to its rapid global dissemination and frequent multi-drug resistant (MDR) phenotype (Nicolas-Chanoine et al. 2014). This has led to ST131 being well characterised by publications that propose biological explanations for its emergence and spread (Price et al. 2013; Petty et al. 2014; Salipante et al. 2015; Ben Zakour et al. 2016; Stoesser et al. 2016). Other common STs are less well characterized despite their association with disease, in part because they are less often defined as MDR, and because ST131 is an important player in the broader global problem of increasing antibiotic resistance in Gram-negative bacteria, with clinical isolates that are

resistant to aminoglycosides, fluoroquinolones, extended-spectrum β -lactamases, carbapenems and colistin beginning to emerge (Chen et al. 2014; Zhang et al. 2014; Liu et al. 2016; Skov and Monnet 2016).

Many of the published whole genome sequencing (WGS) studies on *E. coli* have largely concentrated on ST131, with fewer focused on other extra-intestinal pathogenic *E. coli* (ExPEC). Studies have characterised ST131 in detail and highlighted genetic events leading to the success of this lineage. Two studies investigating the origin of enteropathogenic *E. coli* (EPEC) and atypical enteropathogenic *E. coli* (aEPEC) and the association of genetic factors with clinical disease severity illustrated the power of WGS by showing that aEPEC and EPEC emerged several times in different lineages (Hazen et al. 2016; Ingle et al. 2016), and a further study analysed a global collection of 362 enterotoxigenic *E. coli* (ETEC) (von Mentzer et al. 2014). Smaller studies of local epidemics have concentrated on other pathotypes and single STs.

Here, we used WGS to analyse the genetic diversity of a large collection of *E. coli* isolates associated with bloodstream infection over more than a decade, using nested systematic surveys of England and the Cambridge area. These were not selected based on ST or other bacterial characteristic. We investigated trends in population structure and mechanisms of antibiotic resistance and captured the introduction of ST131 and ST69, which enabled us to study the dynamics of emergence and its effect on the wider *E. coli* population.

RESULTS

Study design and bacterial isolates

We conducted a retrospective study in which we analysed WGS data for 1509 *E. coli* isolates drawn from national BSAC (n=1094) and local (n=415) collections. The BSAC collection consisted of isolates submitted to a bacteremia resistance surveillance program (www.bsacsurv.org) between 2001-2011 by 11 hospitals across England. From each hospital, the first ten isolates (when available) for each year were included into the study. The local collection was sourced from the diagnostic laboratory at the Cambridge University Hospitals NHS Foundation Trust (CUH), Cambridge. Using the laboratory database, we selected every third isolate associated with bacteremia that had been stored in the -80°C freezer archive between 2006 and 2012.

Phylogeny and pan-genome

The 1509 *E. coli* isolates were resolved into 228 STs. The most frequent STs were ST73 (17.3%), ST131 (14.4%), ST95 (10.6%), ST69 (5.5%) and ST12 (4.6%), which accounted for more than half of the collection. The distribution of STs between the BSAC and CUH collections was comparable. Details of all STs are provided in Supplemental Table S1. The population structure of the collection based on core genome single nucleotide polymorphisms (SNPs) was defined using Bayesian Analysis of Population Structure (BAPS), which provides an independent method of assessing the population structure based on the data in the collection, not based on previous definitions. This correlated well with ST (Figure 1), and we therefore used STs to allow for direct comparisons between our data and previous studies. However, there were inconsistencies with phylogroups, which have been linked to the source of isolation and virulence (Picard et al. 1999), and have been previously used to describe the *E. coli* population structure (Lecointre et al. 1998). Most isolates (n=1018, 67%) were assigned to phylogroup B2. The remainder were distributed between phylogroups F (n=151, 10%), A

(n=130, 9%), D (139, 9%), B1 (n=69, 5%), and E (n=2, <1%) (Figure 1). Four of the five most common STs resided in phylogroup B2 (ST73, ST131, ST95 and ST12), while ST69 belonged to phylogroup D. A comparison of ST, BAPS clusters, phylogroup and a maximum likelihood (ML) tree based on core genome SNPs is shown in Figure 1. The phylogeny showed five large clades, which generally correspond to phylogroups. However, comparison between phylogroup and core genome-based phylogeny showed that phylogroups F and D were mixed rather than monophyletic groups (Figure 1). This is consistent with the PCR data from Clermont et al. (2013), as well as the presence of an A genotype within the B1 group (Clermont et al. 2013). The ML tree was dominated by phylogroup B2, which showed large clonal expansions. These were mostly absent from groups A and B1, which were in turn dominated by isolates on long branches.

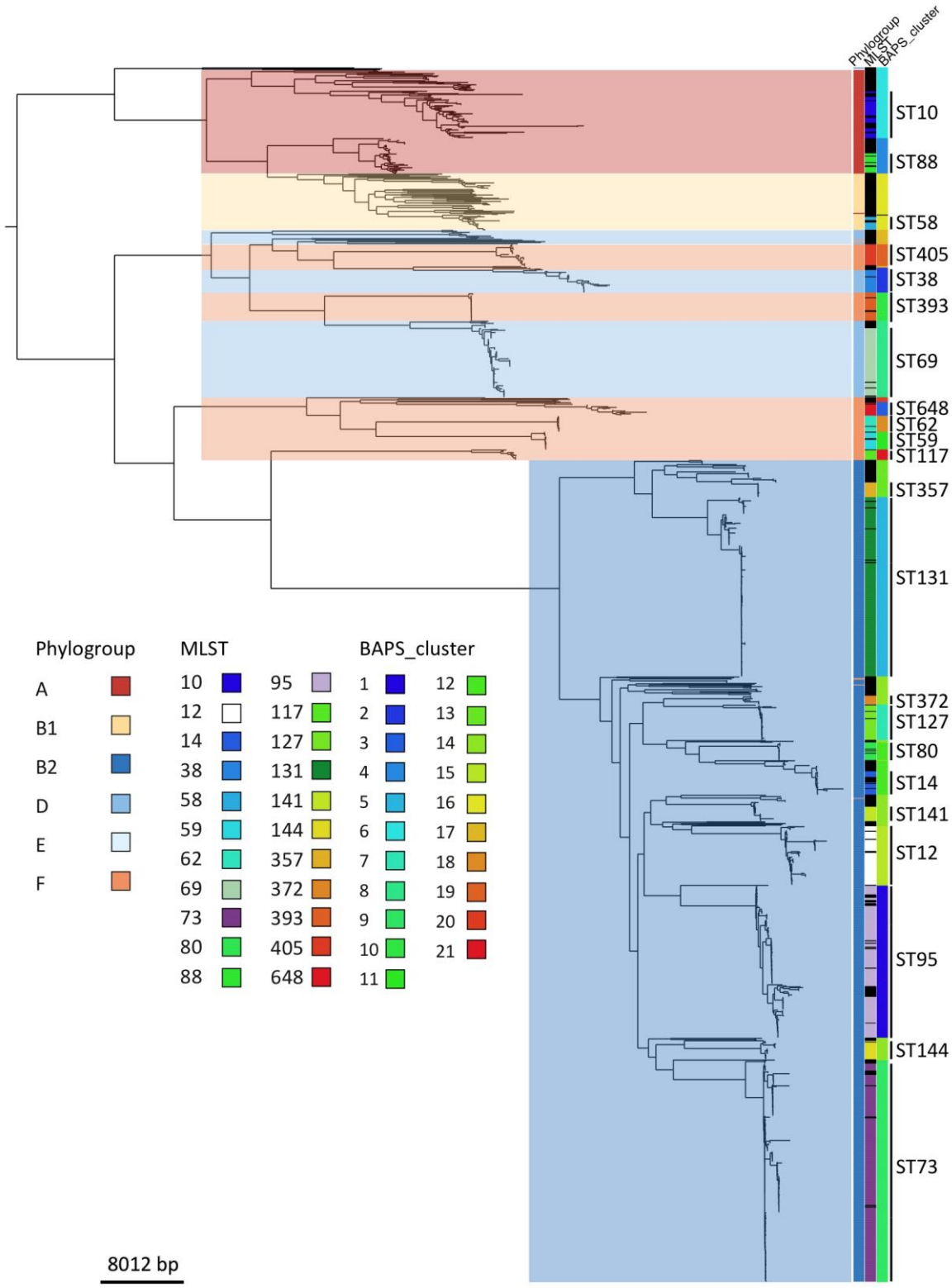


Figure 1. Maximum-likelihood core genome phylogeny of *E. coli* associated with bacteremia in England. The columns on the right show, from left to right, phylogroup, STs containing more than ten isolates, and hierBAPS clusters. Phylogroups are also presented by background shading and STs labelled on the right. Black represents ST designation not shown due to

these having less than ten isolates. The root has been placed according to previous understanding of *Enterobacteriaceae* phylogeny.

Analysis of the pan-genome demonstrated an open pan-genome containing 69,274 genes and no sign of reaching a plateau (Supplemental Figure S1). Using a strict definition of core genome, only 885 genes were present in all 1509 isolates although this rose to 1744 genes using a cut-off of presence in 99% of isolates. The vast majority of genes (62,753 of 69,274, 91%) were present in less than 15% of the isolates.

The population structure of *E. coli* associated with bacteremia

Two STs appeared in the collection for the first time during the timeframe of the study, with ST69 first detected in 2002 and ST131 in 2003. The proportion of STs in each year of the collection is shown in Figure 2. Within a short period after the emergence and spread of ST69 and ST131 the population established a new equilibrium, whereby the proportion of the major STs remain relatively unchanged. The proportion of ST73, ST95 and ST12 before and after the emergence of ST131 were on average 24% versus 17%, 8% versus 11%, and 7% versus 4%, respectively. The proportion of the remaining STs fell from 59% before the emergence of ST131 to 42% after, but was stable from then until the end of the sampling period.

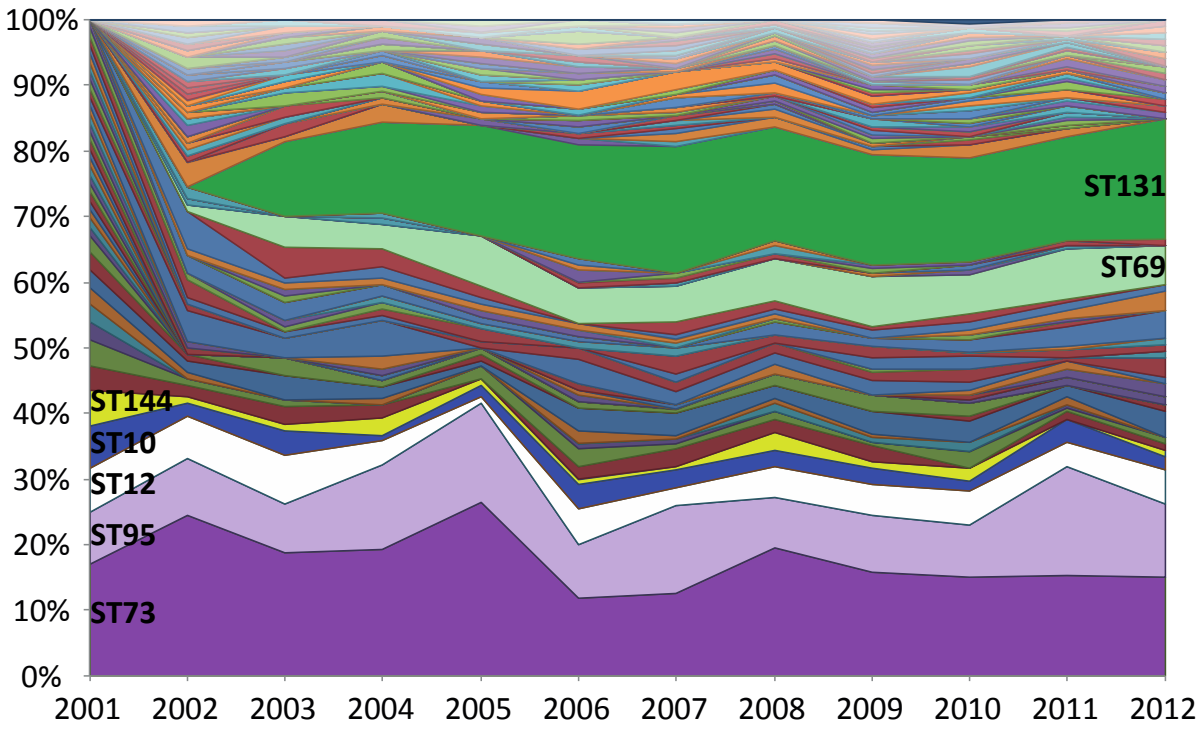
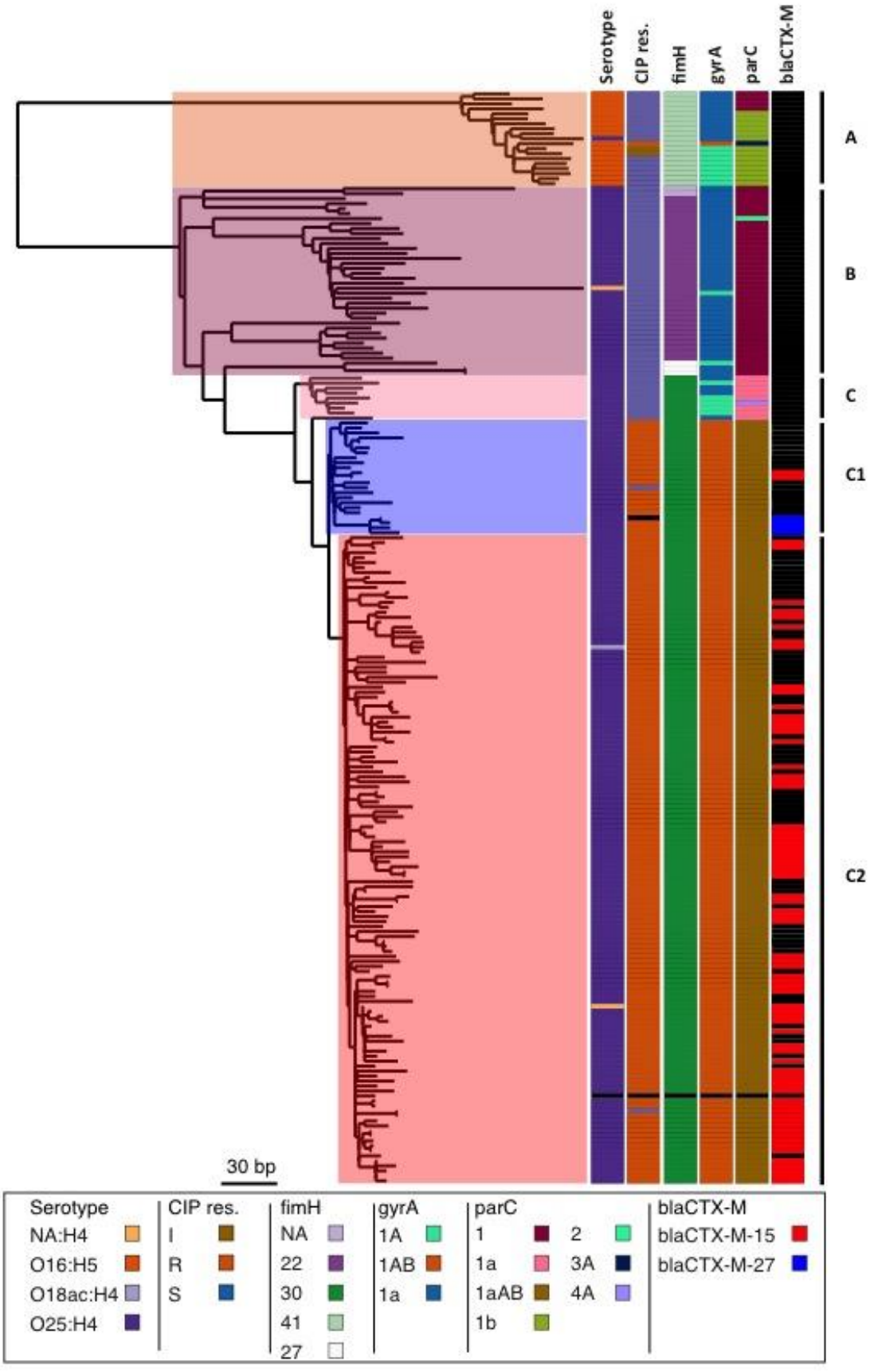


Figure 2. Proportions of STs during the 11 year sampling framework. The percentage of each ST has been plotted by year ordered by the frequency at the start of the study (most common at the bottom). The emergence of ST131 and ST69 can be observed in 2003 and 2002, respectively.

Genetic characterisation of ST131

Three major clades have been identified for ST131 (Price et al. 2013; Petty et al. 2014): clade A corresponds to serotype O16:H5, clade B is serotype O25:H4 and is negative for the *fimH*30 allele, and clade C (H30) is serotype O25:H4 and positive for *fimH*30 and sub-divided by the acquisition of fluoroquinolone resistance in clade C1 (H30-R in Price et al). This has been further divided into clade C2 (H30-Rx in Price et al), described previously as having acquired $bla_{CTX\text{-}M\text{-}15}$ encoding extended spectrum beta-lactamase, followed by expansion of this clade (Price et al. 2013; Olesen et al. 2014). Our 218 ST131 isolates were assigned to these lineages using previously-described lineage-defining variation. This demonstrated that 197 (90%) were serotype O25:H4, and 18 (9%) isolates at the base of the lineage were serotype O16:H5 and *fimH41*. For two isolates the serotype could not be

explicitly defined *in silico* and one was defined as O18ac:H4. One O25:H4 isolate was within the O16:H5 positive clade A. The C1 clade was defined based on a comparative phylogenetic analysis with the Price et al. 2013 isolates (Supplemental Figure S2) and *in silico* PCR to detect H30-Rx (C2) specific SNPs. Of the 161 ST131 isolates in lineage C, 129 belonged to the C2 clade (Figure 3). The assignment of isolates to clades was confirmed by investigating six previously reported clade specific SNPs for B, C, C1 and C2 (Ben Zakour et al. 2016). This confirmed our assignment of isolates to clades and revealed that the three *fimH*27 isolates in the B clade most likely belong to the B0 clade defined by Ben Zakour et al. (2016).


Serotype
CIP res.
fimH
gyrA
parC
blaCTX-M
A
B
C
C1
C2
30 bp
Serotype
NA:H4
O16:H5
O18ac:H4
O25:H4
CIP res.
I
R
S
fimH
NA
22
30
41
27
gyrA
1A
1AB
1a
parC
1
1a
1aAB
1b
2
3A
4A
blaCTX-M
blaCTX-M-15
blaCTX-M-27

184

Figure 3. ST131 maximum-likelihood phylogenetic tree based on SNPs called against the reference EC958. Columns to the right of the tree show the *in silico* predicted serotype (O16-H5 or O25-H4), phenotypic resistance to ciprofloxacin (CIP res.), SNP based definition of *fimH*, *gyrA* and *parC* genotypes and the presence of $bla_{\text{CTX-M}}$ and the type. NA:H4 in the serotype indicates that we were unable to assign a definite O type for the isolate. It has not been counted as a new serotype. Clades assigned based on the markers and clade specific SNPs are shown on the right. The only isolate with missing data (black) is the reference strain EC958. The tree is mid-point rooted.

We then mapped $bla_{\text{CTX-M-15}}$ and fluoroquinolone resistance across the ST131 collection. $bla_{\text{CTX-M-15}}$ was present in both C1 and C2 clades, but was not detected in clades A or B (Figure 3). A parsimony reconstruction of the presence of $bla_{\text{CTX-M-15}}$ within a EC958 reference genome-based ML phylogeny using both acctran and deltran methods indicated at least 28 introductions and/or losses of the $bla_{\text{CTX-M-15}}$ gene in this data set. This indicates that $bla_{\text{CTX-M-15}}$ has been acquired and lost repeatedly in the C1 and C2 clades (Supplemental Figure S3). The majority of the *fimH*30 positive C isolates were ciprofloxacin (fluoroquinolone) resistant, with a small number of exceptions. A cluster of eight fluoroquinolone-susceptible isolates resided close to the root of the C clade, together with two sporadic isolates in the C1 clade. Altogether, 75 of ST131 isolates were $bla_{\text{CTX-M-15}}$ positive. The C1 clade contained only 23 isolates but of these, two were $bla_{\text{CTX-M-15}}$ positive and another four had acquired $bla_{\text{CTX-M-27}}$. Of the 129 clade C2 isolates 73 were $bla_{\text{CTX-M-15}}$ positive. Isolates in lineage A and B were mostly susceptible to ciprofloxacin, the exceptions being one isolate in clade A that was resistant to ciprofloxacin and two clade A isolates with intermediate resistance. The resistant isolate had the *gyrA* mutation associated with fluoroquinolone resistance (*gyrA*1AB) (Johnson et al. 2013).

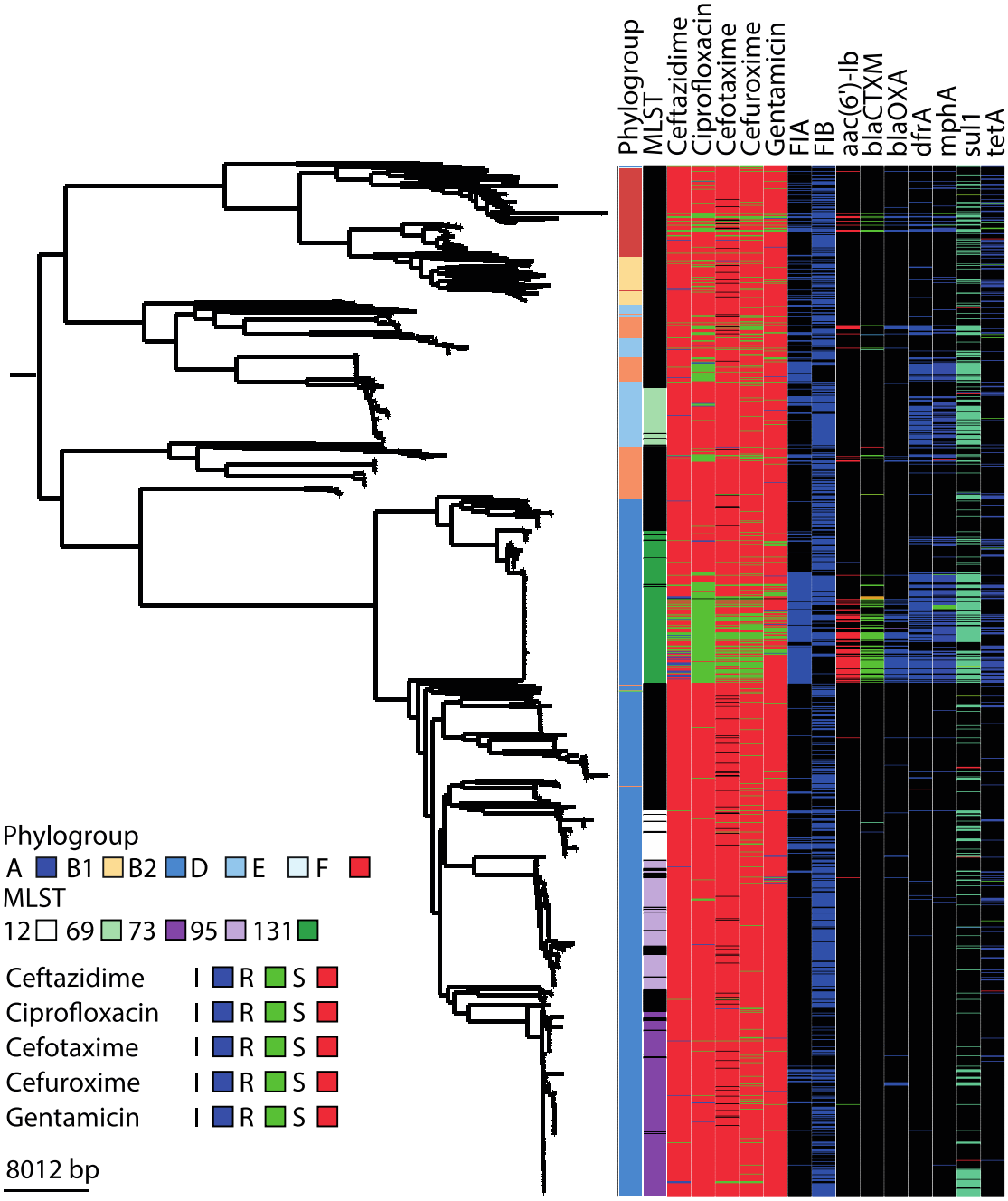
ST131 *espC* island

We analysed the presence of 3511 virulence genes in the whole collection and observed that just one gene was almost specific to the ST131. This was present in ST131 and the closest lineages in the B2 phylogroup. This was more common in ST131 (N=216, 99.08%) compared with other STs (p-value<2.2e-16) The gene was annotated as *espC* (a member of the Serine Protease Autotransporters of *Enterobactericeae*, SPATE, family). The gene is contained in a genetic island reported previously as ROD3 in ST131 strain EC958 (Totsika et al. 2011), but is not identical to the first description of an *espC* pathogenicity island which was originally reported in EPEC (Stein et al. 1996; Mellies et al. 2001; Schmidt and Hensel 2004). The sequence identity/similarity of the EPEC *espC* and EC958 *espC* was 68% for DNA and 69/73% for protein. The ST131 *espC* island has genes coding for *fimD*, *focC* (*fimC*), *tsh*, *cfaD* (*regA*), and *espC*, along with two poorly characterised proteins, one with similarity to fimbrial adhesins and one to DNA binding proteins. The island is bordered by *yjdJ* genes. The locus where the island is inserted is conserved at the *yjdIJKO* gene region, between the *dcu*S and *lysU* genes in *E. coli* reference strain K-12 MG1655 (NC_000913.2). Analysis of mapping coverage to the reference EC958 showed that the island was present in all ST131 isolates (N=218), closely related clades to ST131 in the B2 phylogeny (N=45) and four isolates in phylogroup D and six in phylogroups B1/A. A region of the island is missing from 11% of the ST131 isolates, 19 of which belong to the clade A and were missing a common region. The same region was missing in the clades close to ST131 in the phylogeny (Supplemental Figure S4). The *espC* allele present outside of the ST131 clades B and C was often different than the *espC* in the rest of ST131 (Supplemental Figure S5).

ST73 and ST131 have different strategies to achieve prevalence

ST73 and ST131 represented the predominant STs in the collection, but are known to have contrasting antibiotic resistance profiles. Consistent with this, our ST131 isolates were

predominantly MDR and ST73 was largely susceptible (Figure 4 and Supplemental Figure S6). This was reflected by the presence of numerous antibiotic resistance genes in ST131 compared with ST73, which was accounted for at least in part by different plasmid profiles. ST131 was the main lineage in the collection to carry an incFIA plasmid(s) (Figure 4 and Supplemental Figure S6). This contained *aac6'*-lb-cr, *bla*$_{\text{CTX-M-15}}$ and *bla*$_{\text{OXA1}}$ and indicates that this plasmid is mostly responsible for the multi-drug resistant ST131 phenotype (Supplemental Figure S6). A more widely disseminated plasmid in ST131 was also present and carried incFIB often in addition to incFIA. This plasmid encodes *bla*$_{\text{TEM-1}}$, *dfrA, mphA, sul1* and *tetA*. Due to the limitations of Illumina short-read technology, it is not possible to further delineate the structure of the genetic element encoding these genes and therefore the presence of genes in specific plasmids is determined by association alone and has a level of uncertainty. The difference in susceptibility profiles of ST73 and ST131, which remain at stable proportions of the population throughout the study after the introduction of ST131, suggests that resistance may not be the primary determinant of successful establishment and maintenance in the reservoir niche.



254

255 Figure 4. Multi-drug resistance plasmids present in ST131. Phylogeny of the whole collection
256 with columns to the right representing phylogroup, five most frequent STs, phenotypic
257 antibiotic resistance data linked to the plasmid (ceftazidime, ciprofloxacin, cefotaxime,
258 cefuroxime and gentamicin), the presence of incFIA and incFIB as well as antibiotic
259 resistance genes carried by the plasmid (*aac(6')-Ib*, $bla_{\text{CTX-M}}$, bla_{OXA}, *dfr*A, *mph*A, *sul*1 and
260 *tet*A), black=missing, colour=present. The phylogenetic tree is the same as in Figure 1.

261

Comparing the phylogeny of the two lineages showed that ST131 was mainly dominated by clade C isolates, which were very closely related, but the ST73 tree has several more divergent clades. The average pairwise SNP distance between all the ST131 isolates was 156 SNPs (median=74 SNPs) (Figure 3). By contrast, the ST73 phylogeny comprised at least 8 clades with isolates that were much less closely related (average pairwise SNP distance in ST73 is 335 SNPs, median=332 SNPs) (Figure 5). This is underlined by the observation that ST73 isolates were assigned to 9 serotypes *in silico*, and the different serotypes were in phylogenetically separate lineages. It seems likely that a change in serotype has occurred at least 7 times in ST73 (Figure 5). By contrast, within ST131 only three serotypes were identified. O16:H5 was present in clade A and O25:H4 present in the rest of the phylogeny represented by clades B and C. One isolate in the C2 clade was O18ac:H4. A comparison of the presence of virulence genes in ST73 and ST131 revealed differences in the presence of the UPEC/ExPEC virulence genes between the two (Supplemental Figure S5), suggesting that again, there is not a single configuration that is best for success in this niche represented by MDR ST131, but rather also susceptible, but fit and virulent, STs can become prevalent. For example, most ST131 isolates lacked gene clusters *hlyABCD* (hemolysin) and *iroBCDN* (salmochelin) but carried genes for aerobactin (*iucABCD*, *iutA*), hemin uptake (*chuASTUWXY*;), and yersiniabactin (*fyuA*, *irp1*, *irp2*, *ybtAEPQSTUX*) (Supplemental Figure S7). It seems that ST131 can use only aerobactin, yersiniobactin and the *chuASTUWXY* for iron acquisition, but for example ST73 has the potential to use all these and others.

An interesting similarity between the two most prevalent lineages ST131 and ST73 is that both are positive for *sat* (ST131 89% and ST73 87%), a gene encoding a secreted autotransporter toxin that is toxic for kidney and bladder cell lines (Guyer et al. 2000) and is less widely present in rest of the population (27%) (p-values < 2.2e-16). Another similarity is

the absence of *aec* genes in ST73 and ST131 (except *aec7* which is present in ST73), which encode genes associated with type 6 secretion, although they were widely present in the rest of the phylogeny (Supplemental Figure S5).

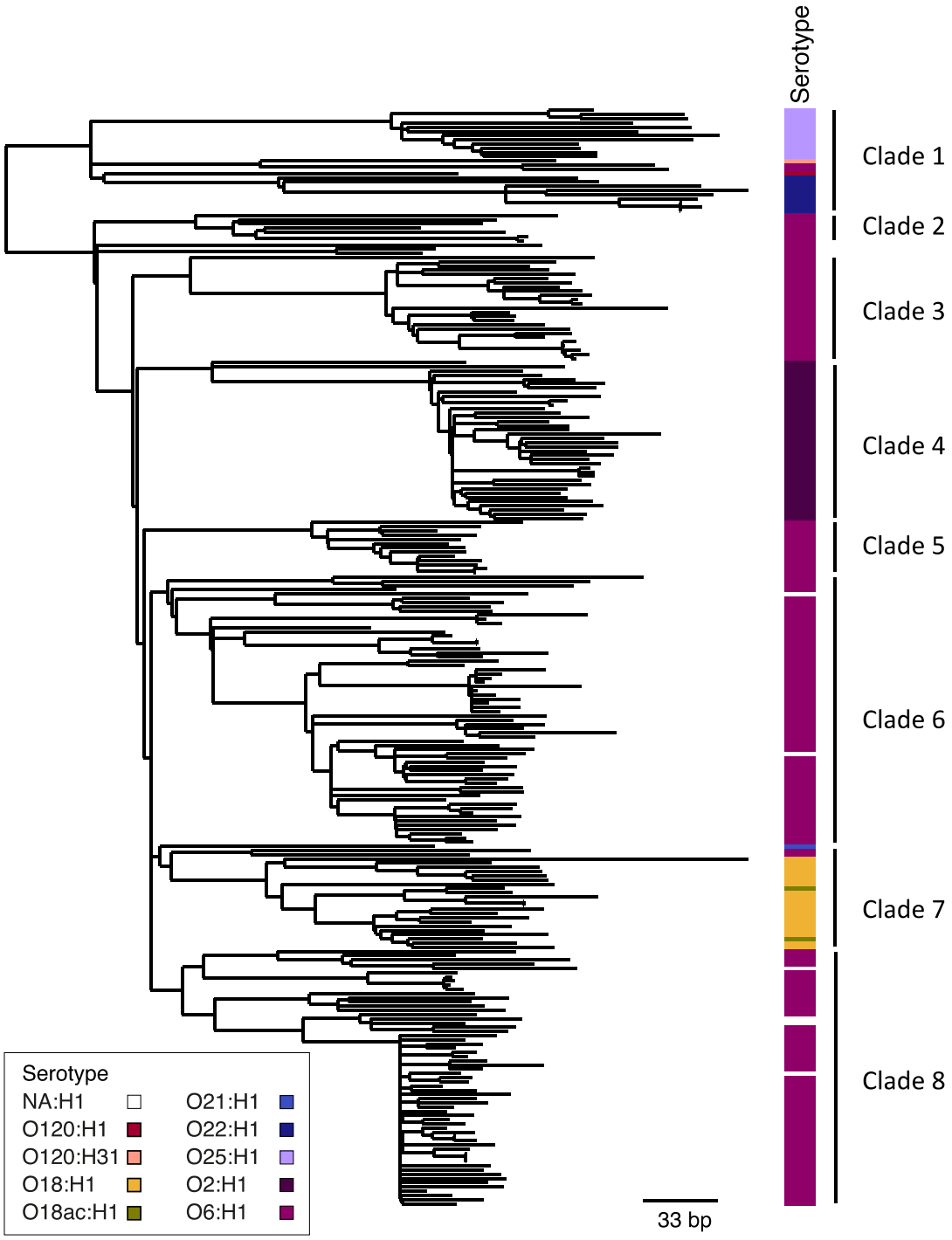


Figure 5. ST73 maximum-likelihood phylogenetic tree based on SNPs called against the reference CFT073. The *in silico* predicted serotype is shown to the right of the tree. For serotype NA:H1 the O type could no be assigned. This is not counted as a new serotype. The clades are labelled on the right. The tree is mid-point rooted.

Two strategies for emergence – ST131 and ST69

Two independent lineages (ST69 and ST131) became established in our study population during the study period. To investigate the structure and history of this emergence we constructed temporally resolved phylogenies using Bayesian evolutionary analysis by sampling trees (BEAST). Based on the BEAST analysis the MRCA (most recent common ancestor) for ST69 was dated to 1956 (95% highest posterior density (HPD) interval, 1935–1971). This separated the major clade in the ST69 tree that subsequently divided into two large lineages in 1977 (95% HPD interval, 1965–1986) (Supplemental Figure S8). Using the Bayesian skyline model we could estimate the effective population size in the past. The analysis showed three increases in the population size. The first increase beginning in the late 1970s and the second in the 1990s were smaller than the last rapid increase that occurred relatively close to the year 2000 (Supplemental Figure S9). If the confidence interval is taken into account, we hypothesise that we observed the last increase in population size during this study.

The BEAST analysis estimated that the MRCA of the clades with the two different serotypes in ST131 was around 1874 (95% HPD interval, 1697–1951) (Figure 6 and Supplemental Figure S10). The C clade diverged from the rest of the phylogeny around 1960 (95% HPD interval, 1899–1985) and the fluoroquinolone resistant C1 clade is estimated to have diverged from the rest of the ST131 lineages around 1982 (95% HPD interval, 1948–1995). We

repeated this analysis with Least-Squares Dating (LSD) which gave a date for the MRCA of the complete ST131 collection of 1828 (confidence interval 1672-1891), 1934 (1871-1960) for divergence of clade C from A and B clades, 1979 (1953-1991) for divergence of the fluoroquinolone resistant clade from the rest, and 1986 (1965-1993) for the divergence of C2 clade from C1. The majority of the ST131 nodes, including the fluoroquinolone susceptible lineages in the tree had a divergence time of 30 years or less. This is indicative of the fact that the whole ST131 lineage, rather than just a single clade within it, has increased in prevalence after its observed emergence in England in the 2000s. This is also apparent from an analysis of isolation dates against the phylogeny of ST131 in our collection (Supplemental Figure S11). It can be seen that nearly all of the major clades of the tree contained isolates from every year, indicating that the whole population was present in each year throughout the 11 years in this study, not just the clade C2. The skyline plot showed a sharp increase in the population size around the year 2000 (Supplemental Figure S9). This most likely correlates with the emergence of the ST131 lineage that we observed in our data set.

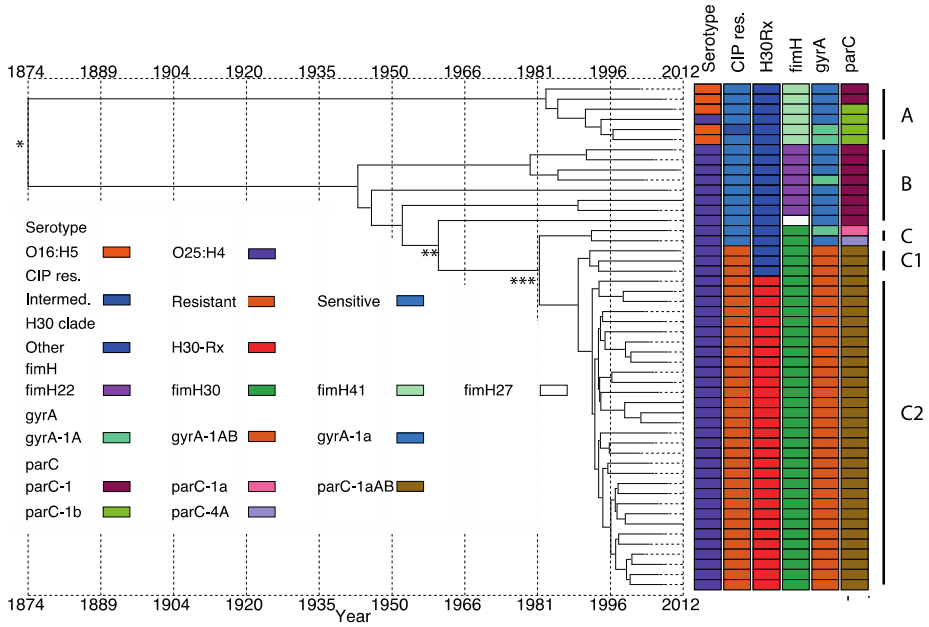


Figure 6. Temporal analysis on ST131 using BEAST. Figure shows the serotype, resistance to Ciprofloxacin (CIP res.), assignment to clade C2 (H30-Rx), and the gene alleles of *fimH*, *gyrA* and *parC*. * MRCA, ** emergence of clade C, *** emergence of CIP resistant clade C1.

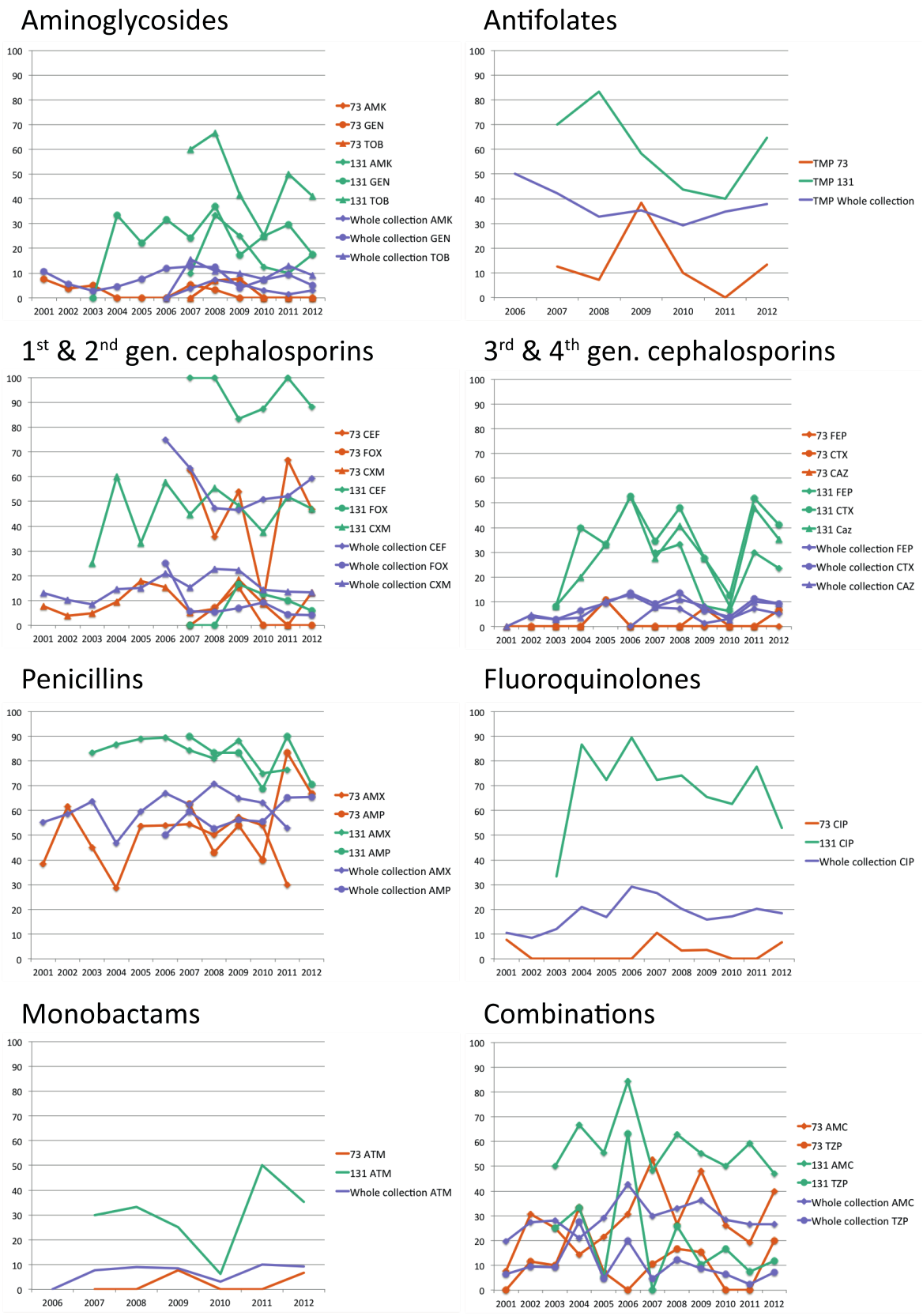
Distribution of virulence factors

We analysed the repertoire of virulence factors, focusing on genes present or absent in the most prevalent sequence types in our collection, ST73 and ST131. ST131 is known to cause urinary tract infections, yet this clade has only a partial *pap* (P fimbriae or pyelonephritis associated pilus) gene operon (Clark et al. 2012): 82% of ST131 had only *papABIX* or fewer genes from the operon and not e.g. the tip adhesin *papG*. In more detail, 90% of ST131 isolates had *papA*, but only 17% had the tip adhesion *papG* (Supplemental Figure S7). Most of the other clades in the B2 phylogroup contained the intact operon, which was also present in other phylogroups. In addition, the most closely related clades to ST131 were also missing most of this operon. Genes encoding enterotoxins *set1AB*, autotransporter *pic*, F1C fimbrial genes *focAGH*, and the autotransporter *upaH* (Supplemental Figure S5) were found in ST73 but were rare in other STs.

Mapping against the reference strain EC958 also enabled us to analyse the presence of the reported ST131 genomic island ROD3 (Totsika et al. 2011) and the type 6 secretion system (T6SS) across the whole invasive *E. coli* population in England (Supplemental Figure S4). T6SS was specific to the B2 phylogroup. The T6SS is used as an anti-competition mechanism to enhance survival in a competitive niche such as the gut (Chatzidaki-Livanis et al. 2016; Sana et al. 2016). This may be one explanation for the prevalence of the B2 phylogroup in our collection.

Antibiotic resistance

Phenotypic antibiotic resistance to ciprofloxacin increased from 10.5% to 28.8% (probably due to the emergence of ST131) and peaked in 2006, but remained under 20% from 2008–2012. It is striking that there was no consistent change in the phenotypic antibiotic resistance of the two most prevalent sequence types, ST73 and ST131, except for ampicillin resistance in ST73 (Figure 8). Furthermore, there was no clear increase in antibiotic resistance over time for the whole collection. ST131 was more resistant to most antimicrobials than ST73 (and often the rest of the collection) and contained the most antibiotic resistance genes (Figure 4). However, the equally successful ST73 was one of the least resistant lineages to most antimicrobials. Phenotypic antibiotic susceptibility results are summarised in Table 1.



371

372 Figure 7. Comparison of antibiotic resistance of ST131, ST73 and the whole collection.
373 Phenotypic antibiotic resistance data is represented by % of non-susceptible (resistant +

intermediate) isolates per year. Each subfigure represents one antibiotic class. Carbapenems (imipemen, meropenem and ertapenem) and tigecycline are not shown due to the lack of resistance against these classes in this collection. AMK=amikacin, GEN=gentamicin, TOB=tobramycin, TMP=trimethoprim, CEF=cefalotin, FOX=cefoxitin, CXM=cefuroxime, FEP=cefepime, CTX=cefotaxime, CAZ=ceftazidime, AMX=amoxicillin, AMP=ampicillin, CIP=ciprofloxacin, ATM=aztreonam, AMC=amoxicillin-clavulanic acid, TZP=piperacillin-tazobactam.

DISCUSSION

We analysed whole-genome sequence data for 1509 *E. coli* blood isolates taken from a systematic sentinel-based surveillance program across England, as well as unbiased sampling from a University hospital, isolated in 2001–2012. This eleven-year period enabled us to analyse the temporal trends in the population structure and changes in the antibiotic resistance of invasive *E. coli* as well as characterize in detail the most prevalent sequence types causing bacteraemia. The most predominant STs were ST73, ST131, ST95, ST69 and ST12. These belong to *E. coli* phylogroups B2 and D, which have been previously associated with virulent and pathogenic UPEC and ExPEC strains (Picard et al. 1999; Johnson and Stell 2000; Johnson et al. 2001). The fact that some of the genotypes from the (*in silico*) PCR did not present monophyletic clades made it difficult to assign all isolates to phylogroups without the use of a phylogenetic tree. The pattern of prevalence of STs is consistent with previous studies of isolates associated with urinary sepsis or bloodstream infection; for example, during a similar time period in Ireland the most frequent STs were ST131, ST73 and ST69 (Miajlovic et al. 2016) and the *E. coli* from the BSAC Bacteraemia Resistance Surveillance Programme have the same most common profiles (CC73, CC131 and CC95) as this combined collection from BSAC and CUH (Day et al. 2016). This also shows that the collections from BSAC and CUH are similar when STs are considered. Information from hierBAPS clustering showed that BAPS clusters were more often monophyletic than STs

since single and double locus variants often disturbed the uniformity of ST clades. The presence of phylogroup F isolates within group D, already reported in the original article presenting the method (Clermont et al. 2013), made the use of the current PCR-based method for assigning isolates to phylogroups problematic for some clades.

The study period captured the emergence of ST69 and ST131 into our study population, but the introduction of these lineages only disturbed the population structure transiently after which it quickly reached a new equilibrium, with the new lineages subsequently maintaining a stable proportion of the population. Despite its apparent success, the globally disseminated MDR lineage ST131 failed to become dominant in the whole population. The most common lineage prior to the emergence of ST131, ST73, only reached a proportion of over 20% of the whole population in a single year. This suggests that the driver for the overall structure of the population and proportions of successful clones may be a form of negative frequency-dependent selection. New STs have an advantage when rare, but this is lost when they become more common. One hypothesis to explain this is that the *E. coli* causing bloodstream infections do not form a discrete population, but represent a spill-over of *E. coli* that occupy a commensal niche in the wider human population. This is supported by evidence that both drug-resistant (ST131) and drug-susceptible (ST73) lineages are equally successful in being maintained in this reservoir, and that drug resistance as a whole in the population is not increasing, demonstrating that antibiotic resistance is not a primary driver of success or prevalence in this niche. In the case of ST131, this is supported by the findings of Ben Zakour et al. (2016) who reported that that virulence determinants were acquired before the emergence of the fluoroquinolone resistant C clade (Ben Zakour et al. 2016). According to this hypothesis, the primary forces shaping the population are not those within the hospital environment, but are due to competition within the gut commensal niche in the broader

426 human population, where antibiotic use is more sporadic than in the hospital population. This
427 is a strong contrast with the population structure of true nosocomial pathogens such as
428 methicillin-resistant *Staphylococcus aureus* (MRSA), where specific drug-resistant clones
429 sequentially dominate the population within a niche where antibiotic exposure is common
430 and drug resistance is a strong selective advantage.
431
432 ST73 in our collection was susceptible to the antibiotics tested, but a recent report from the
433 UK found several MDR ST73 isolates associated with an MDR plasmid (Alhashash et al.
434 2016). Further studies are needed to define whether this MDR phenotype will become more
435 disseminated in ST73 over time.
436
437 Although previously reported to be present in the UK and Ireland from 2001 (Day et al.
438 2016), we observed the appearance of ST69 and ST131 within our sampling framework. We
439 used BEAST to analyse their emergence and found that the events leading to their spread
440 seem quite different. A Bayesian skyline analysis of ST69 showed several sequential minor
441 increases in population size, which started in the late 1970s with the last and most rapid one
442 coinciding with the beginning of this study. The most recent common ancestor of ST69 could
443 be dated to around 60 years ago. A similar analysis of the ST131 lineage identified that the
444 split between the O25:H4 and O16:H5 lineages occurred around 142 years ago. The
445 phylogeny and the Bayesian skyline analysis support the observation that there was a single
446 rapid expansion of this clade in the last few years, starting around 1995. However, rather than
447 being due to a single sub-lineage, this expansion seems to have happened in the entire
448 O25:H4 clade and possibly the whole ST. According to our analysis, the *fimH30* carrying
449 clade C diverged from clade B around 1960, the fluoroquinolone resistant clade C1 (H30-R)
450 diverged from clade C around 1982 and clade C2 (H30-Rx) diverged from the rest of clade C

around 1990. These dates are somewhat different from the recently published analyses of Stoesser *et al* (2016) and Ben Zakour *et al* (2016). The emergence of the fluoroquinolone resistant subset of the C clade in the 1980s found here is consistent with two previous publications. Similarly, the divergence of the C2 clade from the C1 in 1990 reported here is similar to the year of 1987 reported by Ben Zakour et al. (2016) and Stoesser et al (2016). This indicates a strong temporal signal in the sequences for these genetic events. The difference in other dates are most likely due to weaker temporal signals in the data for these events e.g. the divergence of the C clade from B clade is dated close to the year 1960 by Ben Zakour et al. (2016) and in this study, but to 1985 by Stoesser et al (2016). The fact that all of the ST131 clades were present during the whole study period indicates that the entire ST131 lineage is successful, and not just clade C or the fluoroquinolone resistant and ESBL-expressing clades C1 and C2.

For ST131 the presence of $bla_{\text{CTX-M-15}}$ was initially reported to be exclusive to the H30-Rx (C2) clade (Price et al. 2013). In our collection, $bla_{\text{CTX-M-15}}$ was found in both the C1 and C2 clades. We also identified that $bla_{\text{CTX-M-15}}$ was acquired and/or lost several times in the ST131 population studied here. Stoesser et al. had previously hypothesised that this could occur based on the diverse contexts in which the gene is found (Stoesser et al. 2016). It is also noteworthy that our unbiased sample of *E. coli* causing BSI was dominated by C2 isolates, with over 80% of clade C isolates and almost 60% of all the isolates belonging to clade C2. This is in contrast to the previous reports of the ST131 population structure where the C1 and C2 have been more equally distributed (Price et al. 2013; Petty et al. 2014; Ben Zakour et al. 2016; Stoesser et al. 2016).

The pan-genome of our collection of invasive *E. coli* included almost 70,000 genes. The core was very small and most of the genes in the pan-genome were present in a small subset of strains, reflecting the massive diversity of *E. coli*. Previous reports on the size of the *E. coli* pan-genome have been markedly smaller, but also the data sets that the analyses were performed on have been smaller. With an open pan-genome this will have an effect on the pan-genome size. Rasko *et al.* reported a pan-genome of over 13,000 genes and a core of 2,200 genes from an analysis of a diverse set of 17 genomes from different *E. coli* pathovars (Rasko et al. 2008). Chen *et al.* analysed the pan and core-genomes of seven UPEC isolates and reported a core genome of 2,865 genes (Chen et al. 2006). Both of these estimates of the core are considerably larger than ours. This may be due to the high similarity cut-off in our analysis, which will make the core genome appear smaller and the pan-genome larger. However, even with a cut-off of 90% the pan-genome is 46,022 genes in this dataset and the core genome 1170 genes. Recently, a core of 1,080 gene clusters was reported in a study investigating 70 EPEC isolates using large-scale BLAST score ratio analysis (Hazen et al. 2016). This is closer to our number of 885 core genes even though the pan-genome in the analysis by Hazen *et al.* was only 12,964 gene clusters. This is likely due to the different methods that were used in the analyses.

The proportion of antibiotic resistant isolates in the data set did not increase substantially during the 11 years that were included in this study. The resistance patterns fluctuated over time, which was probably due to the relatively small sample size per location per each year. Despite the fact that we did not observe a clear increase in the proportion of resistance, there have been many recent reports suggesting an increase in the proportion of resistant isolates (ESPAUR 2015; Ironmonger et al. 2015). This may be due to differences at the regional

level, while this study was performed at a nationwide level, and therefore smaller increases at a local level might not have been apparent.

The analysis of virulence factors in this diverse population enabled us to compare the most successful lineages to each other and to the rest of the population. We identified a number of virulence factors that were differentially present between lineages. One gene, the serine protease autotransporter (SPATE) *espC*, was specific to ST131 and closely related lineages. This is situated in a pathogenicity island that is restricted almost exclusively to ST131 and its closest clades. This island was first reported by Totsika et al. (2011) as a region of difference 3 (ROD3), and has previously been reported to be present throughout ST131 by Petty et al. (2013), but not to be conserved in clade A (Totsika et al. 2011; Petty et al. 2014). Since the ST131 *espC* island does not have an integrase or other mobility related genes it is not clear if it is a self- mobile element. Its sporadic acquisition in different branches of the tree could reflect acquisition by mechanisms other than self-mobility (such as homologous recombination in flanking sequences) or could potentially represent lineage specific deletion. In EPECs *espC* has been shown to play a role in cell death by causing apoptosis and necrosis (Serapio-Palacios and Navarro-Garcia 2016) and the ST131 *espC* island has recently been reported to harbour a gene encoding the regulatory protein RegA (annotated as *cfaD* by Prokka in our analysis) that is present in *Citrobacter rodentium* and *Escherichia* clades III, IV and V (Tan et al. 2015). Incorporation of this island may be one reason for the success of ST131. There were several genes present almost exclusively in ST73 that could in part contribute to its success. These genes were *focAGH* (encoding F1C fimbriae genes), *pic* (encoding another SPATE gene), *set1AB* (encoding *Shigella* enterotoxin 1) and *upaH* which is an autotransporter that induces biofilm formation and bladder colonisation (Allsopp et al. 2010; Allsopp et al. 2012). The secreted autotransporter toxin encoded by *sat* (another

524 SPATE gene) was present in both ST131 and ST73 and could be contributing to the success
525 of both lineages.
526
527 One limitation of this study is that although the collection was drawn from 11 centres over 11
528 years, it only comprised the first 10 isolates per site each year except in the case of CUH. It
529 could therefore potentially include isolates from temporally limited local epidemics, which
530 could skew the results and interpretation. In addition, the limitation of short-read sequencing
531 is evident when analysing plasmids. The dynamic nature of plasmids, in combination with
532 short reads generated by the Illumina HiSeq, means that many differences in plasmid
533 structure are unlikely to be captured by our analyses. We are also unable to assemble
534 complete plasmid sequences and so the presence of genes in plasmids with given inc-types is
535 based on association alone. More detailed analysis would require the use of long read
536 technologies. Determining the presence of genes by association alone adds a degree of
537 uncertainty to the results due to untypeable plasmids and the mobile nature of the genetic
538 elements that can be present in the chromosome.
539
540 In summary, we have analysed the population structure of *E. coli* associated with
541 bloodstream infection over an 11-year period in England. During this time we observed the
542 emergence of ST131 and ST69, but this introduction did not disturb the population structure
543 for long, and a new equilibrium was quickly established. The globally disseminated MDR
544 lineage ST131 was not the most frequently identified lineage in this collection; this was
545 ST73, which is generally susceptible to most antibiotics. This indicates that antibiotic
546 resistance is not a primary driver of success in the niche occupied by these *E. coli*. The
547 relatively static structure of the population suggests that it is instead driven by negative
548 frequency-dependent selection occurring in the commensal niche in the wider human

549 population, and that bacteraemia represents a spill-over from this population. This
550 emphasises the importance of surveillance of the wider human population to understand the
551 dynamics and structure of invasive *E. coli*.
552
553 **MATERIALS AND METHODS**
554 **Bacterial isolates**
555 A total of 1522 *E. coli* isolates were initially included in the study. Of these, 1098 were from
556 the British Society for Antimicrobial Chemotherapy (BSAC) Bacteraemia Resistance
557 Surveillance Programme (www.bsacsurv.org) (Reynolds et al. 2008) between 2001 and 2011
558 (Supplemental Table S2). Up to 10 isolates (when available) were obtained for each year
559 from 11 contributing laboratories distributed across England. The 11 centres were selected in
560 order to provide geographical and temporal diversity. A further 424 isolates were sourced
561 from the diagnostic laboratory at the Cambridge University Hospitals NHS Foundation Trust,
562 Cambridge. Using the laboratory database, we selected every third isolate associated with
563 bacteraemia that had been stored in the -80°C freezer archive between 2006 and 2012.
564 Thirteen isolates were subsequently excluded (4 and CUH isolates and 9 BSAC isolates)
565 based on low quality of sequence data or species misidentification, giving a final sample size
566 of 1509 isolates. Antimicrobial susceptibility testing was performed using the Vitek2
567 instrument with the N206 card (Biomerieux, Marcy l'Etoile, France) for isolates from CUH,
568 and using the agar dilution method for the BSAC collection (Andrews 2001). For the
569 purposes of this analysis we combined phenotypic antibiotic resistance data from BSAC and
570 CUH, and grouped together the intermediate and resistant isolates in the analyses to represent
571 the non-susceptible part of the population. Since the isolates from BSAC and CUH have been
572 tested against different antibiotic combinations we have antibiotic resistance data from 2001–
573 2011 for amoxicillin and imipenem; from 2006–2012 for amikacin, tobramycin, ampicillin,

ertapenem, meropenem, aztreoman, cefalotin, cefoxitin, cefepime and trimethoprim and throughout the study period (2001–2012) for gentamicin, tigecycline, cefuroxime, ceftazidime, cefotaxime, ciprofloxacin, amoxicillin-clavulanic acid and piperacillin-tazobactam.

The National Research Ethics Service (ref. 12/EE/0439) and the Cambridge University Hospitals NHS Foundation Trust (CUH) Research and Development (R&D) Department approved the study protocol.

DNA extraction and sequencing

Genomic DNA was extracted using a QIAxtractor (QIAGEN) and library preparation performed according to the Illumina protocol. Index-tagged libraries were created, 96 isolates multiplexed per lane and sequenced using the Illumina HiSeq 2000 platform (Illumina Inc.) to generate 100 base pair (bp) paired-end reads. Average sequencing depth was 77-fold, with a minimum of 48-fold.

Sequence data analysis

MLST was performed using an in-house script (https://github.com/sanger-pathogens/mlst_check) and sequence types (ST) defined using the Warwick MLST scheme (Wirth et al. 2006). *De novo* assembly was performed using Velvet (Zerbino and Birney 2008), and scaffolds were generated using SSPACE (Boetzer et al. 2011) and GapFiller (Boetzer and Pirovano 2012). Reads were mapped back to the assemblies using SMALT (Ponstingl 2013). Assemblies were annotated with an in-house pipeline based on Prokka (Seemann 2014). Annotated assemblies were used in a pan-genome analysis in Roary, from

598 which a core genome alignment was generated (Page et al. 2015). List of genes in the core
599 genome and the core genome alignment are can be downloaded from Figshare.
600
601 Bayesian analysis of population structure (hierBAPS) (Corander et al. 2008; Cheng et al.
602 2013) was used to analyse the population structure. A core genome alignment was produced
603 with Roary and a SNP alignment was generated using SNP-sites (Page et al. 2016) and used
604 in hierBAPS which was run 5 times with the prior upper bound for the number of clusters
605 varying between 100 and 300. All runs converged to the same estimate of the global posterior
606 mode partition, indicating a strong support for the obtained clustering solution. Phylogenetic
607 trees were generated using SNP sites determined by SNP-sites from the core genome
608 alignments, or from SNPs identified by mapping to reference genomes, using RAxML 7.8.6
609 with 100 bootstraps (Stamatakis 2006). For the reference-based SNP tree the sequences were
610 mapped against a selected reference genome using SMALT and SNPs were called using
611 SAMtools (Li et al. 2009; Ponstingl 2013). Reference genomes used in the analyses were
612 CFT073 (AE014075.1) and EC958 (NZ_HG941718.1) (Forde et al. 2014), for ST73 and
613 ST131, respectively Phage sequences were recognized using Phast (Zhou et al. 2011) and
614 were masked from the analysis. Phast results can be retrieved from the program website
615 (phast.wishartlab.com). Gubbins with default settings was used to identify recombination
616 (Croucher et al. 2015), and the regions detected were masked from the alignment. The
617 resulting alignment was used to produce a phylogenetic tree with RAxML.
618
619 Temporal analysis of ST69 and ST131 was performed with BEAST 1.7.5.1-1 (Drummond
620 and Rambaut 2007; Drummond et al. 2012) on a reference-based alignment of 50 randomly
621 selected isolates from both STs. This approach was used since BEAST did not converge with
622 the complete collections or 100 isolates in the available running time on the computer cluster.

References in the analysis were UMN026 (NC_011751.1) and EC958 (NZ_HG941718.1), for ST69 and ST131, respectively. Phast was used to identify phage regions and Gubbins was used to identify regions of recombination in the alignment and these regions were masked from the alignment before running BEAST. The nucleotide substitution model used was GTRGAMMA and we ran three replicates of all combinations for strict clock and lognormal relaxed clock and three tree priors; coalescent: constant population, exponential growth and Bayesian skyline. To estimate what was the best fitting model for each ST we compared Bayes Factors from marginal likelihood estimations calculated by path and stepping-stone sampling (Baele et al. 2012; Baele et al. 2013). Only the models that converged well and had effective sample size (ESS) over 200 for each parameter were included in the test. The best fitting model was used in the subsequent analyses. For ST131 this was Bayesian skyline model under the log-normal relaxed clock and for ST69 it was constant population model under the strict clock followed by Bayesian skyline model under the strict clock. For the construction of the Bayesian skyline for ST69 we used data generated with Bayesian skyline model under the strict clock. The temporal analysis for ST131 was confirmed with Least-Squares Dating (LSD, version 0.3beta) using all isolates (To et al. 2016).

In silico PCR was used to assign isolates to *E. coli* phylogroups A, B1, B2, D, E and F using the Clermont method (Clermont et al. 2000; Clermont et al. 2013), to assign ST131 isolates to B, H30-R (C1) and H30-Rx (C2) clades (Price et al. 2013; Ben Zakour et al. 2016) and to perform plasmid incompatibility group/replicon typing (Carattoli et al. 2005). Primers designed to detect clade specific SNPs reported by Ben Zakour et al. are presented in Supplementary Table S3. *In silico* serotyping was performed with SRST2 according to the author's instructions with the database provided (Ingle et al. 2015). Here, serotype is defined

by presence of known genes encoding serotype-determining enzymes. The results required minor manual curating when the gene typing resulted in discrepancy between the gene pairs defining the serotype. This occurred mostly with the novel sequences the authors had included in the database based on their own results. Antibiotic resistance genes were detected using SRST2 with 98% identity. An in-house curated database based on ResFinder of antibiotic resistance genes was used as reference (Zankari et al. 2012). Parsimony reconstruction of the presence of $bla_{CTX\text{-}M\text{-}15}$ in ST131 was performed with Fitch algorithm (Fitch 1971). Virulence genes were analysed with SRST2 using the database and protocols described by the authors and using an *Escherichia* genus-specific database clustering together genes with 90% similarity and detecting genes with 90% identity and at least 90% coverage (Inouye et al. 2014). Gene typing for *gyrA*, *parC* and *fimH* was performed by clustering sequences acquired from Roary using Usearch (Edgar 2010). Clustering was based on reference sequences for seven *fimH* genes, seven *gyrA* genes and ten *parC* genes. The alleles tested were as described previously (Johnson *et al.* 2013) with the following exceptions: *fimH*15 was omitted and *fimH*31 was added to the analysis (Johnson et al. 2013).

Statistical testing

To test if a specific gene is more often found in certain STs, we used Pearson's chi-squared test statistic using prop.test in R (R Core Team 2015).

DATA ACCESS

Sequence reads have been submitted to the European Nucleotide Archive (ENA) (www.ebi.ac.uk/ena) under the accession numbers listed in Supplemental Data S1. The lists of genes present in the core genome and core genome alignment can be downloaded from

Figshare (https://figshare.com/s/20dfe5842f952497619b and https://figshare.com/s/3a12b011ff3c291a271b).

ACKNOWLEDGEMENTS

We thank Beth Blane for laboratory assistance, and the library construction, sequencing and core informatics teams at the Wellcome Trust Sanger Institute. We thank the BSAC for providing isolates from the BSAC Resistance Surveillance Project. This publication presents independent research supported by the Health Innovation Challenge Fund (HICF-T5-342 and WT098600), a parallel funding partnership between the UK Department of Health and Wellcome Trust. The views expressed in this publication are those of the authors and not necessarily those of the Department of Health, Public Health England or the Wellcome Trust.

DISCLOSURE DECLARATION

The authors report no conflicts of interest.

References

Adams-Sapper S, Diep BA, Perdreau-Remington F, Riley LW. 2013. Clonal composition and community clustering of drug-susceptible and -resistant Escherichia coli isolates from bloodstream infections. *Antimicrob Agents Chemother* **57**(1): 490-497.

Alhashash F, Wang X, Paszkiewicz K, Diggle M, Zong Z, McNally A. 2016. Increase in bacteraemia cases in the East Midlands region of the UK due to MDR Escherichia coli ST73: high levels of genomic and plasmid diversity in causative isolates. *J Antimicrob Chemother* **71**(2): 339-343.

Alhashash F, Weston V, Diggle M, McNally A. 2013. Multidrug-resistant Escherichia coli bacteremia. *Emerg Infect Dis* **19**(10): 1699-1701.

Allsopp LP, Beloin C, Moriel DG, Totsika M, Ghigo JM, Schembri MA. 2012. Functional heterogeneity of the UpaH autotransporter protein from uropathogenic Escherichia coli. *J Bacteriol* **194**(21): 5769-5782.

Allsopp LP, Totsika M, Tree JJ, Ulett GC, Mabbett AN, Wells TJ, Kobe B, Beatson SA, Schembri MA. 2010. UpaH is a newly identified autotransporter protein that contributes to biofilm formation and bladder colonization by uropathogenic Escherichia coli CFT073. *Infect Immun* **78**(4): 1659-1669.

Andrews JM. 2001. Determination of minimum inhibitory concentrations. *J Antimicrob Chemother* **48 Suppl 1**: 5-16.

Baele G, Lemey P, Bedford T, Rambaut A, Suchard MA, Alekseyenko AV. 2012. Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. *Mol Biol Evol* **29**(9): 2157-2167.

Baele G, Li WL, Drummond AJ, Suchard MA, Lemey P. 2013. Accurate model selection of relaxed molecular clocks in bayesian phylogenetics. *Mol Biol Evol* **30**(2): 239-243.

Banerjee R, Johnston B, Lohse C, Chattopadhyay S, Tchesnokova V, Sokurenko EV, Johnson JR. 2013. The clonal distribution and diversity of extraintestinal Escherichia coli isolates vary according to patient characteristics. *Antimicrob Agents Chemother* **57**(12): 5912-5917.

Ben Zakour NL, Alsheikh-Hussain AS, Ashcroft MM, Khanh Nhu NT, Roberts LW, Stanton-Cook M, Schembri MA, Beatson SA. 2016. Sequential Acquisition of Virulence and Fluoroquinolone Resistance Has Shaped the Evolution of Escherichia coli ST131. *MBio* **7**(2): e00347-00316.

Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W. 2011. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* **27**(4): 578-579.

Boetzer M, Pirovano W. 2012. Toward almost closed genomes with GapFiller. *Genome Biol* **13**(6): R56.

Carattoli A, Bertini A, Villa L, Falbo V, Hopkins KL, Threlfall EJ. 2005. Identification of plasmids by PCR-based replicon typing. *J Microbiol Methods* **63**(3): 219-228.

Chatzidaki-Livanis M, Geva-Zatorsky N, Comstock LE. 2016. Bacteroides fragilis type VI secretion systems use novel effector and immunity proteins to antagonize human gut Bacteroidales species. *Proc Natl Acad Sci U S A* **113**(13): 3627-3632.

Chen L, Hu H, Chavda KD, Zhao S, Liu R, Liang H, Zhang W, Wang X, Jacobs MR, Bonomo RA et al. 2014. Complete sequence of a KPC-producing IncN multidrug-resistant plasmid from an epidemic Escherichia coli sequence type 131 strain in China. *Antimicrob Agents Chemother* **58**(4): 2422-2425.

Chen SL, Hung CS, Xu J, Reigstad CS, Magrini V, Sabo A, Blasiar D, Bieri T, Meyer RR, Ozersky P et al. 2006. Identification of genes subject to positive selection in uropathogenic strains of Escherichia coli: a comparative genomics approach. *Proc Natl Acad Sci U S A* **103**(15): 5977-5982.

738 Cheng L, Connor TR, Siren J, Aanensen DM, Corander J. 2013. Hierarchical and spatially
739 explicit clustering of DNA sequences with BAPS software. *Mol Biol Evol* **30**(5):
740 1224-1228.
741 Clark G, Paszkiewicz K, Hale J, Weston V, Constantinidou C, Penn C, Achtman M, McNally
742 A. 2012. Genomic analysis uncovers a phenotypically diverse but genetically
743 homogeneous Escherichia coli ST131 clone circulating in unrelated urinary tract
744 infections. *J Antimicrob Chemother* **67**(4): 868-877.
745 Clermont O, Bonacorsi S, Bingen E. 2000. Rapid and simple determination of the
746 Escherichia coli phylogenetic group. *Appl Environ Microbiol* **66**(10): 4555-4558.
747 Clermont O, Christenson JK, Denamur E, Gordon DM. 2013. The Clermont Escherichia coli
748 phylo-typing method revisited: improvement of specificity and detection of new
749 phylo-groups. *Environ Microbiol Rep* **5**(1): 58-65.
750 Corander J, Marttinen P, Siren J, Tang J. 2008. Enhanced Bayesian modelling in BAPS
751 software for learning genetic structures of populations. *BMC Bioinformatics* **9**: 539.
752 Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, Bentley SD, Parkhill J, Harris
753 SR. 2015. Rapid phylogenetic analysis of large samples of recombinant bacterial
754 whole genome sequences using Gubbins. *Nucleic Acids Res* **43**(3): e15.
755 Day MJ, Doumith M, Abernethy J, Hope R, Reynolds R, Wain J, Livermore DM, Woodford
756 N. 2016. Population structure of Escherichia coli causing bacteraemia in the UK and
757 Ireland between 2001 and 2010. *J Antimicrob Chemother* **71**(8): 2139-2142.
758 de Kraker ME, Jarlier V, Monen JC, Heuer OE, van de Sande N, Grundmann H. 2013. The
759 changing epidemiology of bacteraemias in Europe: trends from the European
760 Antimicrobial Resistance Surveillance System. *Clin Microbiol Infect* **19**(9): 860-868.
761 Drummond AJ, Rambaut A. 2007. BEAST: Bayesian evolutionary analysis by sampling
762 trees. *BMC Evol Biol* **7**: 214.
763 Drummond AJ, Suchard MA, Xie D, Rambaut A. 2012. Bayesian phylogenetics with
764 BEAUti and the BEAST 1.7. *Mol Biol Evol* **29**(8): 1969-1973.
765 Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST.
766 *Bioinformatics* **26**(19): 2460-2461.
767 Elixhauser A, Friedman B, Stranges E. 2011. Septicemia in U.S. Hospitals, 2009: Statistical
768 Brief #122. In *Healthcare Cost and Utilization Project (HCUP) Statistical Briefs*,
769 Rockville (MD).
770 ESPAUR. 2015. English surveillance programme for antimicrobial utilisation and resistance
771 (ESPAUR) 2010 to 2014: report 2015. Public Health England.
772 Fitch WM. 1971. Toward Defining Course of Evolution - Minimum Change for a Specific
773 Tree Topology. *Systematic Zoology* **20**(4): 406-416.
774 Forde BM, Ben Zakour NL, Stanton-Cook M, Phan MD, Totsika M, Peters KM, Chan KG,
775 Schembri MA, Upton M, Beatson SA. 2014. The complete genome sequence of
776 Escherichia coli EC958: a high quality reference sequence for the globally
777 disseminated multidrug resistant E. coli O25b:H4-ST131 clone. *PLoS One* **9**(8):
778 e104400.
779 Gerver R, Mihalkova M, Abernethy J, Bou-Antoun S, Nsonwu O, Kausar S, Wasti S, Apraku
780 D, Davies J, Hope R. 2015. Annual Epidemiological Commentary: Mandatory
781 MRSA, MSSA and E. coli bacteraemia and C. difficile infection data, 2014/15. p. 81.
782 Publich Health England, London, United Kingdom.
783 Gibreel TM, Dodgson AR, Cheesbrough J, Fox AJ, Bolton FJ, Upton M. 2012. Population
784 structure, virulence potential and antibiotic susceptibility of uropathogenic
785 Escherichia coli from Northwest England. *J Antimicrob Chemother* **67**(2): 346-356.

786 Guyer DM, Henderson IR, Nataro JP, Mobley HL. 2000. Identification of sat, an
787 autotransporter toxin produced by uropathogenic Escherichia coli. *Mol Microbiol*
788 **38**(1): 53-66.
789 Hazen TH, Donnenberg MS, Panchalingam S, Antonio M, Hossain A, Mandomando I,
790 Ochieng JB, Ramamurthy T, Tamboura B, Qureshi S et al. 2016. Genomic diversity
791 of EPEC associated with clinical presentations of differing severity. *Nature*
792 *Microbiology* **1**: 15014.
793 Horner C, Fawley W, Morris K, Parnell P, Denton M, Wilcox M. 2014. Escherichia coli
794 bacteraemia: 2 years of prospective regional surveillance (2010-12). *J Antimicrob*
795 *Chemother* **69**(1): 91-100.
796 Ingle D, Valcanis M, Kuzevski A, Tauschek M, Inouye M, Stinear T, Levine MM, Robins-
797 Browne RM, Holt KE. 2015. EcOH: In silico serotyping of E. coli from short read
798 data. *bioRxiv*: http://dx.doi.org/10.1101/032151.
799 Ingle DJ, Tauschek M, Edwards DJ, Hocking DM, Pickard DJ, Azzopardi KI, Amarasena T,
800 Bennett-Wood V, Pearson JS, Tamboura B et al. 2016. Evolution of atypical
801 enteropathogenic E. coli by repeated acquisition of LEE pathogenicity island variants.
802 *Nature Microbiology* **1**: 15010.
803 Inouye M, Dashnow H, Raven LA, Schultz MB, Pope BJ, Tomita T, Zobel J, Holt KE. 2014.
804 SRST2: Rapid genomic surveillance for public health and hospital microbiology labs.
805 *Genome Med* **6**(11): 90.
806 Ironmonger D, Edeghere O, Bains A, Loy R, Woodford N, Hawkey PM. 2015. Surveillance
807 of antibiotic susceptibility of urinary tract pathogens for a population of 5.6 million
808 over 4 years. *J Antimicrob Chemother* **70**(6): 1744-1750.
809 Johnson JR, Delavari P, Kuskowski M, Stell AL. 2001. Phylogenetic distribution of
810 extraintestinal virulence-associated traits in Escherichia coli. *J Infect Dis* **183**(1): 78-
811 88.
812 Johnson JR, Stell AL. 2000. Extended virulence genotypes of Escherichia coli strains from
813 patients with urosepsis in relation to phylogeny and host compromise. *J Infect Dis*
814 **181**(1): 261-272.
815 Johnson JR, Tchesnokova V, Johnston B, Clabots C, Roberts PL, Billig M, Riddell K, Rogers
816 P, Qin X, Butler-Wu S et al. 2013. Abrupt emergence of a single dominant multidrug-
817 resistant strain of Escherichia coli. *J Infect Dis* **207**(6): 919-928.
818 Lecointre G, Rachdi L, Darlu P, Denamur E. 1998. Escherichia coli molecular phylogeny
819 using the incongruence length difference test. *Mol Biol Evol* **15**(12): 1685-1695.
820 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin
821 R, Genome Project Data Processing S. 2009. The Sequence Alignment/Map format
822 and SAMtools. *Bioinformatics* **25**(16): 2078-2079.
823 Liu YY, Wang Y, Walsh TR, Yi LX, Zhang R, Spencer J, Doi Y, Tian G, Dong B, Huang X
824 et al. 2016. Emergence of plasmid-mediated colistin resistance mechanism MCR-1 in
825 animals and human beings in China: a microbiological and molecular biological
826 study. *Lancet Infect Dis* **16**(2): 161-168.
827 Mellies JL, Navarro-Garcia F, Okeke I, Frederickson J, Nataro JP, Kaper JB. 2001. espC
828 pathogenicity island of enteropathogenic Escherichia coli encodes an enterotoxin.
829 *Infect Immun* **69**(1): 315-324.
830 Miajlovic H, Mac Aogain M, Collins CJ, Rogers TR, Smith SG. 2016. Characterization of
831 Escherichia coli bloodstream isolates associated with mortality. *J Med Microbiol*
832 **65**(1): 71-79.
833 Nicolas-Chanoine MH, Bertrand X, Madec JY. 2014. Escherichia coli ST131, an intriguing
834 clonal group. *Clin Microbiol Rev* **27**(3): 543-574.

Olesen B, Frimodt-Moller J, Leihof RF, Struve C, Johnston B, Hansen DS, Scheutz F, Krogfelt KA, Kuskowski MA, Clabots C et al. 2014. Temporal Trends in Antimicrobial Resistance and Virulence-associated Traits within Escherichia coli Sequence Type 131 Clonal Group and its H30 and H30-Rx Subclones, 1968 - 2011. *Antimicrob Agents Chemother* **58**(11): 6886-6895.

Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MT, Fookes M, Falush D, Keane JA, Parkhill J. 2015. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **31**(22): 3691-3693.

Page AJ, Taylor B, Delaney AJ, Soares J, Seemann T, Keane JA, Harris SR. 2016. SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. *Microbial Genomics* **2**(4): doi:10.1099/mgen.1090.000056.

Petty NK, Ben Zakour NL, Stanton-Cook M, Skippington E, Totsika M, Forde BM, Phan MD, Gomes Moriel D, Peters KM, Davies M et al. 2014. Global dissemination of a multidrug resistant Escherichia coli clone. *Proc Natl Acad Sci U S A* **111**(15): 5694-5699.

Picard B, Garcia JS, Gouriou S, Duriez P, Brahimi N, Bingen E, Elion J, Denamur E. 1999. The link between phylogeny and virulence in Escherichia coli extraintestinal infection. *Infect Immun* **67**(2): 546-553.

Ponstingl H. 2013. SMALT 0.7.5. http://www.sanger.ac.uk/resources/software/smalt/.

Price LB, Johnson JR, Aziz M, Clabots C, Johnston B, Tchesnokova V, Nordstrom L, Billig M, Chattopadhyay S, Stegger M et al. 2013. The epidemic of extended-spectrum-beta-lactamase-producing Escherichia coli ST131 is driven by a single highly pathogenic subclone, H30-Rx. *MBio* **4**(6): e00377-00313.

R Core Team. 2015. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rasko DA, Rosovitz MJ, Myers GS, Mongodin EF, Fricke WF, Gajer P, Crabtree J, Sebaihia M, Thomson NR, Chaudhuri R et al. 2008. The pangenome structure of Escherichia coli: comparative genomic analysis of E. coli commensal and pathogenic isolates. *J Bacteriol* **190**(20): 6881-6893.

Reynolds R, Hope R, Williams L, Surveillance BWPoR. 2008. Survey, laboratory and statistical methods for the BSAC Resistance Surveillance Programmes. *J Antimicrob Chemother* **62 Suppl 2**: ii15-28.

Salipante SJ, Roach DJ, Kitzman JO, Snyder MW, Stackhouse B, Butler-Wu SM, Lee C, Cookson BT, Shendure J. 2015. Large-scale genomic sequencing of extraintestinal pathogenic Escherichia coli strains. *Genome Res* **25**(1): 119-128.

Sana TG, Flaugnatti N, Lugo KA, Lam LH, Jacobson A, Baylot V, Durand E, Journet L, Cascales E, Monack DM. 2016. Salmonella Typhimurium utilizes a T6SS-mediated antibacterial weapon to establish in the host gut. *Proc Natl Acad Sci U S A* **113**(34): E5044-5051.

Schmidt H, Hensel M. 2004. Pathogenicity islands in bacterial pathogenesis. *Clin Microbiol Rev* **17**(1): 14-56.

Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**(14): 2068-2069.

Serapio-Palacios A, Navarro-Garcia F. 2016. EspC, an Autotransporter Protein Secreted by Enteropathogenic Escherichia coli, Causes Apoptosis and Necrosis through Caspase and Calpain Activation, Including Direct Procaspase-3 Cleavage. *MBio* **7**(3): e00479-00416.

Skov RL, Monnet DL. 2016. Plasmid-mediated colistin resistance (mcr-1 gene): three months later, the story unfolds. *Euro Surveill* **21**(9).

884 Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses
885 with thousands of taxa and mixed models. *Bioinformatics* **22**(21): 2688-2690.
886 Stein M, Kenny B, Stein MA, Finlay BB. 1996. Characterization of EspC, a 110-kilodalton
887 protein secreted by enteropathogenic Escherichia coli which is homologous to
888 members of the immunoglobulin A protease-like family of secreted proteins. *J*
889 *Bacteriol* **178**(22): 6546-6554.
890 Stoesser N, Sheppard AE, Pankhurst L, De Maio N, Moore CE, Sebra R, Turner P, Anson
891 LW, Kasarskis A, Batty EM et al. 2016. Evolutionary History of the Global
892 Emergence of the Escherichia coli Epidemic Clone ST131. *MBio* **7**(2): e02162-02115.
893 Tan A, Petty NK, Hocking D, Bennett-Wood V, Wakefield M, Praszkier J, Tauschek M,
894 Yang J, Robins-Browne R. 2015. Evolutionary adaptation of an AraC-like regulatory
895 protein in Citrobacter rodentium and Escherichia species. *Infect Immun* **83**(4): 1384-
896 1395.
897 To TH, Jung M, Lycett S, Gascuel O. 2016. Fast Dating Using Least-Squares Criteria and
898 Algorithms. *Syst Biol* **65**(1): 82-97.
899 Totsika M, Beatson SA, Sarkar S, Phan MD, Petty NK, Bachmann N, Szubert M, Sidjabat
900 HE, Paterson DL, Upton M et al. 2011. Insights into a multidrug resistant Escherichia
901 coli pathogen of the globally disseminated ST131 lineage: genome analysis and
902 virulence mechanisms. *PLoS One* **6**(10): e26578.
903 von Mentzer A, Connor TR, Wieler LH, Semmler T, Iguchi A, Thomson NR, Rasko DA,
904 Joffre E, Corander J, Pickard D et al. 2014. Identification of enterotoxigenic
905 Escherichia coli (ETEC) clades with long-term global distribution. *Nat Genet* **46**(12):
906 1321-1326.
907 Wirth T, Falush D, Lan R, Colles F, Mensa P, Wieler LH, Karch H, Reeves PR, Maiden MC,
908 Ochman H et al. 2006. Sex and virulence in Escherichia coli: an evolutionary
909 perspective. *Mol Microbiol* **60**(5): 1136-1151.
910 Zankari E, Hasman H, Cosentino S, Vestergaard M, Rasmussen S, Lund O, Aarestrup FM,
911 Larsen MV. 2012. Identification of acquired antimicrobial resistance genes. *J*
912 *Antimicrob Chemother* **67**(11): 2640-2644.
913 Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de
914 Bruijn graphs. *Genome Res* **18**(5): 821-829.
915 Zhang C, Xu X, Pu S, Huang S, Sun J, Yang S, Zhang L. 2014. Characterization of
916 carbapenemases, extended spectrum beta-lactamases, quinolone resistance and
917 aminoglycoside resistance determinants in carbapenem-non-susceptible Escherichia
918 coli from a teaching hospital in Chongqing, Southwest China. *Infect Genet Evol* **27**:
919 271-276.
920 Zhou Y, Liang Y, Lynch KH, Dennis JJ, Wishart DS. 2011. PHAST: a fast phage search tool.
921 *Nucleic Acids Res* **39**(Web Server issue): W347-352.
922
923

924 Table 1. Proportion of antibiotic non-susceptible isolates in each year.

		2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
Aminoglycosides													
	Gentamicin	10.5	5.7	2.8	4.6	7.5	11.7	12.7	12.1	4.1	7.2	9.4	5.0
	Amikacin						0.0	3.8	7.0	5.6	3.0	1.4	3.0
	Tobramycin						0.0	15.4	10.5	9.7	7.6	12.7	8.9
Antifolates													
	Trimethoprim						50.0	42.3	31.6	34.7	28.8	33.8	36.6
Beta-lactams													
Extended spectrum penicillins													
	Amoxicillin	55.3	58.5	63.0	46.8	59.8	66.4	62.2	71.0	65.0	63.2	53.0	
	Ampicillin						50.0	59.6	50.9	56.9	54.5	63.4	65.3
Carbapenems													
	Imipenem	0.0	0.0	0.0	0.0	0.0	1.9	0.0	0.0	0.0	0.0	0.0	

	Ertapenem						0.0	0.0	0.0	1.4	0.0	0.0	0.0
	Meropenem						0.0	0.0	0.0	0.0	0.0	0.0	0.0
Monobactams													
	Aztreonam						0.0	0.0	8.8	9.7	3.0	9.9	8.9
1st & 2nd generation cephalosporins													
	Cefalotin						75.0	63.5	49.1	47.2	50.0	53.5	59.4
	Cefuroxime	13.2	10.4	8.3	14.7	15.0	20.7	15.3	22.3	22.7	14.4	13.5	13.9
	Cefoxitin						25.0	5.8	5.3	8.3	9.1	5.6	5.9
	Cefuroxime Axetil						25.0	11.5	17.5	20.8	12.1	14.1	13.9
3rd & 4th generation cephalosporins													
	Ceftazidime	0.0	4.7	2.8	3.7	10.3	12.6	8.0	10.8	8.7	2.6	9.9	9.9
	Cefotaxime		3.8	2.8	6.4	9.3	13.5	9.3	13.4	7.0	3.9	11.1	8.9
	Cefepime						0.0	7.7	7.0	1.4	3.0	7.0	5.0

Quinolones													
	Ciprofloxacin	10.5	8.5	12.0	21.1	16.8	28.8	26.7	19.7	15.7	17.0	19.9	17.8
Tetracyclines													
	Tigecycline		0.9	0.0	0.0	0.0	0.0	0.0	0.0	0.6	0.0	0.0	1.0
Combinations													
	AmoxiClav*	19.7	27.4	27.8	21.1	29.0	42.3	30.0	33.1	36.6	28.1	26.9	27.7
	PipTaz**	6.6	9.4	9.3	27.5	4.7	19.8	4.7	12.1	9.4	6.5	2.3	7.1
*Amoxicilline clavulanic acid													
**Piperacillin tazobactam													
NA=no information/not tested													

925