An *Escherichia coli* ST131 pangenome atlas reveals population structure and evolution across 4,071 isolates

Arun Gonzales Decano, Tim Downing
School of Biotechnology, Dublin City University, Dublin 9, Ireland.
Corresponding author: tim.downing@dcu.ie

Abstract

Escherichia coli ST131 is a major cause of infection with extensive antimicrobial resistance (AMR) facilitated by widespread beta-lactam antibiotic use. This drug pressure has driven extended-spectrum beta-lactamase (ESBL) gene acquisition and evolution in pathogens, so a clearer resolution of ST131's origin, adaptation and spread is essential. Its ESBL genes are typically embedded in mobile genetic elements (MGEs) that aid transfer to new plasmid or chromosomal locations, which are mobilised further by plasmid conjugation and recombination, resulting in a flexible ESBL, MGE and plasmid composition with a conserved core genome. We used population genomics to trace the evolution of AMR in ST131 more precisely by extracting all available high-quality Illumina HiSeq read libraries to investigate 4,071 globally-sourced genomes, the largest ST131 collection examined so far. We applied rigorous quality-control, genome *de novo* assembly and ESBL gene screening to resolve ST131's population structure across three genetically distinct clades (A, B, C) and abundant subclades from the dominant clade C. We reconstructed their evolutionary relationships across the core and accessory genomes using published reference genomes, long read assemblies and k-mer-based methods to contextualise pangenome diversity. The three main C subclades have co-circulated globally at relatively stable frequencies over time, suggesting attaining an equilibrium after their origin and initial rapid spread. This contrasted with their ESBL genes, which had a less constrained pattern and stronger population structure. Within these subclades, diversity levels of the core and accessory genome were not correlated due to plasmid and MGE activity. Our findings emphasise the potential of evolutionary pangenomics to improve our understanding of AMR gene transfer, adaptation and transmission to discover accessory genome changes linked to emerging outbreaks and novel subtypes.

Keywords:
Genome, population structure, outbreak, phylogenetics, ST131, *Escherichia coli*, evolution, infection.

Significance

Multidrug-resistant *Escherichia coli* are a major global public health concern, among which ST131 is a pandemic subtype that is the most common cause of urinary tract infections. This study carefully assembled the genomes of 4,071 ST131 isolated between 1967 and 2018 to determine its subclades' epidemiological, evolutionary and genetic features relevant to their antibiotic resistance genes. We found that ST131 subclades relative frequencies were stable over time, suggesting they may spread rapidly during their origin before stabilising. In contrast to core genome analysis documenting the global co-circulation of subclades C1 and C2, key antibiotic resistance genes in the accessory genome had stronger geographic, genetic and temporal patterns. Additionally, extensive plasmid variation among isolates with nearly identical chromosomes was discovered using multiple methods. This population genomic study highlights the dynamic nature of the accessory genomes in ST131, suggesting that surveillance should anticipate genetically novel outbreaks with broader antibiotic resistance levels.

Introduction

Infections caused by multidrug-resistant (MDR) *Escherichia coli* sequence type (ST) 131 are increasing worldwide (de Kraker et al 2013, Poolman & Wacker 2016). ST131 are extraintestinal pathogenic *E. coli* (ExPEC) associated with bloodstream and urinary tract infections and typically possessing extended-spectrum β-lactamase (ESBL) genes (Banerjee & Johnson 2014, ECDC 2017, Findlay et al 2019), or more rarely carbapenemase genes (Peirano et al 2011). MDR ST131 is a major cause of ExPEC infection because it has an extensive range of virulence factors (Totsika et al 2011, Van der Bij et al 2012, Calhau et al 2013, Ben Zakour et al 2016, Goswami et al 2018) and may be highly pathogenic to hosts (Dautzenberg et al 2016). ST131 has been reported in healthcare and community settings around the globe, and its dominant lineage clade C is fluoroquinolone-resistant (FQ-R) (Price et al 2013, Petty et al 2014). Clade C has a type 1 fimbrial adhesin gene *H30* variant (*fimH*30) (Ben Zakour et al 2016, Stoesser et al 2016) and can offset the fitness costs of antimicrobial resistance (AMR), plasmid acquisition and maintenance through compensatory mutations at regulatory regions in contrast to FQ-susceptible clades A and B (McNally et al 2016).

Historically, *E. coli* population structure was inferred from allelic variation at seven housekeeping genes to assign ST complexes via MLST (multi-locus sequence typing) (Wirth et al 2006), or at 51 ribosomal genes for rST (ribosomal MLST) (Jolley et al 2012). Outbreak investigation necessitates sufficient biomarker density to allow isolate discrimination, which is only possible with genome sequencing to allow profiling of all AMR genes (Sintchenko & Holmes 2015, Revez et al 2017). Recent work applied cgMLST (core genome MLST) of 2,512 *E. coli* genes, but computational limitations meant examining 288 ST131 genomes where only a single specimen per rST was examined across 1,230,995 SNPs from a 2.33 Mb core genome, with a larger set of 9,479 diverse *E. coli* (Zhou et al 2019). Given that rST1503 alone may account for ~81% of ST131 and that outbreaks may comprise a single rST (Ludden, Decano et al 2019), our understanding of *E. coli* ST131 transmission dynamics and diversity within single STs may limit inferences of past, present and emerging MDR outbreaks.

A deeper investigation of MDR ST131's population structure, selective processes and ESBL gene evolution can illuminate its mechanisms of AMR, host colonisation and pathogenicity (Ben Zakour et al 2016, Stoesser et al 2016). Exploring the evolutionary origins, transmission and spread of outbreaks requires extensive sampling to link variation at AMR genes with inferred adaptive and epidemiological patterns (Croucher and Didelot 2015), and previous work suggests a high-resolution large-scale approach to bacterial epidemiological based on genomic data address these questions (Lees et al 2019).

Deducing the evolutionary relationships based on the core genome permits the discovery of novel accessory genome events (Downing 2015). ST131 evolution has been punctuated by plasmid conjugation, plasmid recombination and mobile genetic element (MGE) rearrangements, particularly of the cefotaximase (CTX-M) class of ESBL genes, *bla*$_{\text{CTX-M-14/15/27}}$ (Canton et al 2012, Decano et al 2019, Ny et al 2019) that allow third-generation cephalosporin-resistance (Mathers et al 2015). These *bla*$_{\text{CTX-M}}$ gene changes correlate strongly with ST131 subclade differentiation, such that the most common one (C2) is typically *bla*$_{\text{CTX-M-15}}$-positive (Kallonen et al 2017). ESBL and other virulence factor genes likely drive extraintestinal niche colonisation but vary across environments depending on MGE-driven mobility (Johnson et al 2010, Ben Zakour et al 2016, McNally et al 2016, Kallonen et al 2017). When coupled with host immunity, this environmental niche effect results in negative frequency-dependent selection (NFDS) in the ST131 accessory genome, leading to a variable AMR gene repertoire (McNally et al 2019) that has not yet to be explored within ST131's subclades. In addition, applying an evolutionary pangenomic approach with core and accessory genome variation within subclades may inform on the origin of new genetic ST131 lineages.

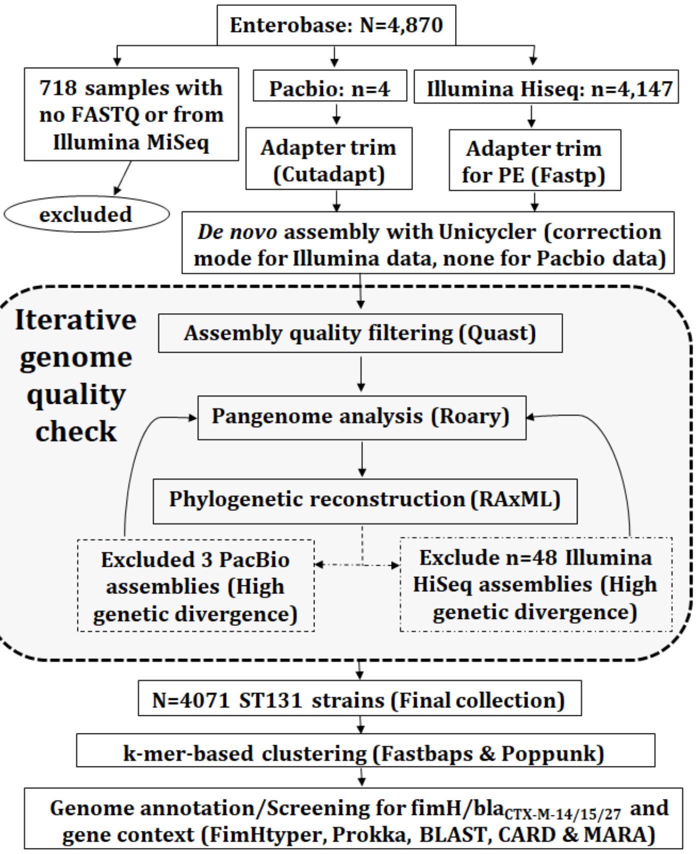
Here, we aggregated all available ST131 Illumina HiSeq read libraries, and automated quality-control, genome *de novo* assembly, DNA read mapping and ESBL gene screening in the largest ST131 set examined thus far to reconstruct a core genome phylogeny and evaluated the epidemiology of clades and subclades. We established that the two most common C subclades (C1 and C2) co-circulated globally and that their ESBL gene composition was flexible. We hypothesise that the diversity of accessory genomes in isolates with near-identical core genomes due to ST131's ability to retain newly acquired rare genes may be driven by environmental pressures.

Results

Collation, screening and generation of 4,071 high quality draft ST131 genome assemblies

We collated accession IDs and linked metadata for 4,071 high quality *de novo* genome assemblies whose DNA was isolated in 1967-2018 from diverse sources across 170 BioProjects (Table S1) following thorough filtering steps (Figure 1, see Methods). 4,070 genomes from Illumina HiSeq read libraries assembled using Unicycler had with N50s of 195,830±57,037 bp (mean ± standard deviation), lengths of 5,136,890±121,402 bp, 124.3±74.8 contigs and 4,829±142 genes (Table S2). The sole PacBio assembly (AR_0058) had five contigs with a N50 of 4,923,470 bp and was 5,132,452 bp long.

Figure 1. Methods summary: 4,870 read libraries were downloaded from Enterobase. 718 uninformative ones were excluded. Of those assessed, four were long read libraries (PacBio) and the rest were short paired-end reads (Illumina HiSeq). The adapters of the four PacBio and 4,147 Illumina reads were trimmed using Cutadapt and Fastp, respectively. The resulting adapter-free reads were assembled using Unicycler. An iterative genome quality check eliminated three PacBio and 77 Illumina libraries, yielding 4,071 as the final collection. Cleaned reads after Quast filtering were examined with Roary using Prokka annotation to evaluate the pangenomic diversity. Phylogenetic construction was performed by RAxML on the core genome. The assembled genomes were annotated and screened for AMR genes (including *bla*$_{\text{CTX-M-14/15/27}}$) and their context. Genetically distinct clusters from the phylogeny were determined using Fastbaps. Distances between the core and accessory genomes of isolate pairs were estimated using Poppunk based on k-mer differences.


Enterobase: N=4,870
718 samples with no FASTQ or from Illumina MiSeq
excluded
Pacbio: n=4
Illumina Hiseq: n=4,147
Adapter trim (Cutadapt)
Adapter trim for PE (Fastp)
De novo assembly with Unicycler (correction mode for Illumina data, none for Pacbio data)
Iterative genome quality check
Assembly quality filtering (Quast)
Pangenome analysis (Roary)
Phylogenetic reconstruction (RAxML)
Excluded 3 PacBio assemblies (High genetic divergence)
Exclude n=48 Illumina HiSeq assemblies (High genetic divergence)
N=4071 ST131 strains (Final collection)
k-mer-based clustering (Fastbaps & Poppunk)
Genome annotation/Screening for fimH/blaCTX-M-14/15/27 and gene context (FimHtyper, Prokka, BLAST, CARD & MARA)

We assembled the 4,071 assemblies' pangenome using NCTC13441 as a reference with Roary v3.11.2 (Page et al 2015) resulting in 26,479 genes, most of which were rare (Figure S1). 3,712 genes present in all samples formed the core genome, and of 22,525 CDSs in the accessory one, 1,018 were shell genes in

15-95% of samples and 21,507 (81% of the total) cloud genes in less than 15% of samples (Figure S2). Cloud gene rates were a function of sample size, which explained most (r^2=0.846, p=0.00012) of cloud gene number variation, but not that of core (r^2=0.162), soft core (r^2=0.258) or shell (r^2=0.001) genes.

Population structure classification shows three dominant ST131 C subclades

Clades A (n=414, 10.1%), B (n=420, 10.3%) and B0 (n=13, 0.3%) were relatively rare in comparison to the 3,224 in clade C (79%) based on *fimH* typing. This showed 91% of A had *fimH41*, 66% of B had *fimH22*, 99% of C had *fimH30*, and unexpectedly B0 had *fimH30*, not *fimH27* (Table 1). Nine isolates were *fimH54*, of which eight were in clade B (Matsumura et al 2017).

Clade/	Fastbaps	***fimH* allele**				Isolate
subclade	**Cluster IDs**	**41**	**22**	**30**	**Others**	**total**
A	2	376			38	414
B	1, 3, 5, 7, 8	8	277		135	420
B0	8			13		13
C0	-			51	1	52
C1	6			1,111	10	1,121
C2	4, 9			2,032	19	2,051
Total		384	277	3,207	203	4,071

Table 1. Number of ST131 in A, B, B0, C0, C1 and C2. Isolates from clade A mainly had *fimH*41 and were assigned to Fastbaps cluster 2. Clade B tended to have *fimH*22 as well as other *fimH* alleles, and were assigned to five Fastbaps groups (1/3/5/7/8). Clade C mainly had *fimH30* or *fimH*-like alleles, and were assigned to Fastbaps cluster 6 for C1 (aka C1_6), or clusters 4 (C2_4) or 9 (C2_9) for C2.

Clustering of the 4,071 isolates based on 30,029 core genome SNPs with Fastbaps identified nine genetically distinct subclades (clusters 1-9) and two groups of heterogeneous or rare isolates (clusters 10 and 11) (Figure 2). Clade A was mainly assigned to cluster 2 (n=407, 98.3%) and seven were unassigned (cluster 11) (Figure S3). Clade B isolated were in clusters 1 (n=90, 21.4%), 3 (n=96, 22.9%), 5 (n=64, 15.2%), 7 (n=115, 27.4%) and 8 (n=4, 1.0%), with an additional 51 (12.1%) unassigned (n=34 in cluster 10, n=17 in cluster 11). All of B0 was in cluster 8, suggesting that it was a component of clade B.

Clade C (n=3,224) had three main subclades determined by Fastbaps: C1_6, C2_4 and C2_9. C1 had 1,121 isolates: 1,113 in Fastbaps cluster 6 (referred to as C1_6) with eight unassigned in cluster 10 (Figure 3). C2 had 2,051 assemblies: 1,651 in cluster 9 (C2_9) and 386 in cluster 4 (C2_4). C0 (n=52) was mainly assigned to cluster 11, consistent with its heterogeneous nature (Ben Zakour et al 2016). 13 C2 genomes were assigned to cluster 10 and one to cluster 6.

Epidemic subclades C1 and C2 co-circulate globally but with stable frequencies

Accessory genome NFDS driven by AMR gene acquisition, ecological niche colonisation ability and host antigen recognition has stabilised the relative frequencies of ST131 and its clades over time relative to other STs (Kallonen et al 2017, McNally et al 2019). Here, this pattern was present for the clades (A, B, C) and three main C subclades, C1_6, C2_4 and C2_9 spanning 2002-2017 (for which year of isolation data was available), during which their relative rates soon stabilised after emergence, consistent with NFDS (Figure S5).

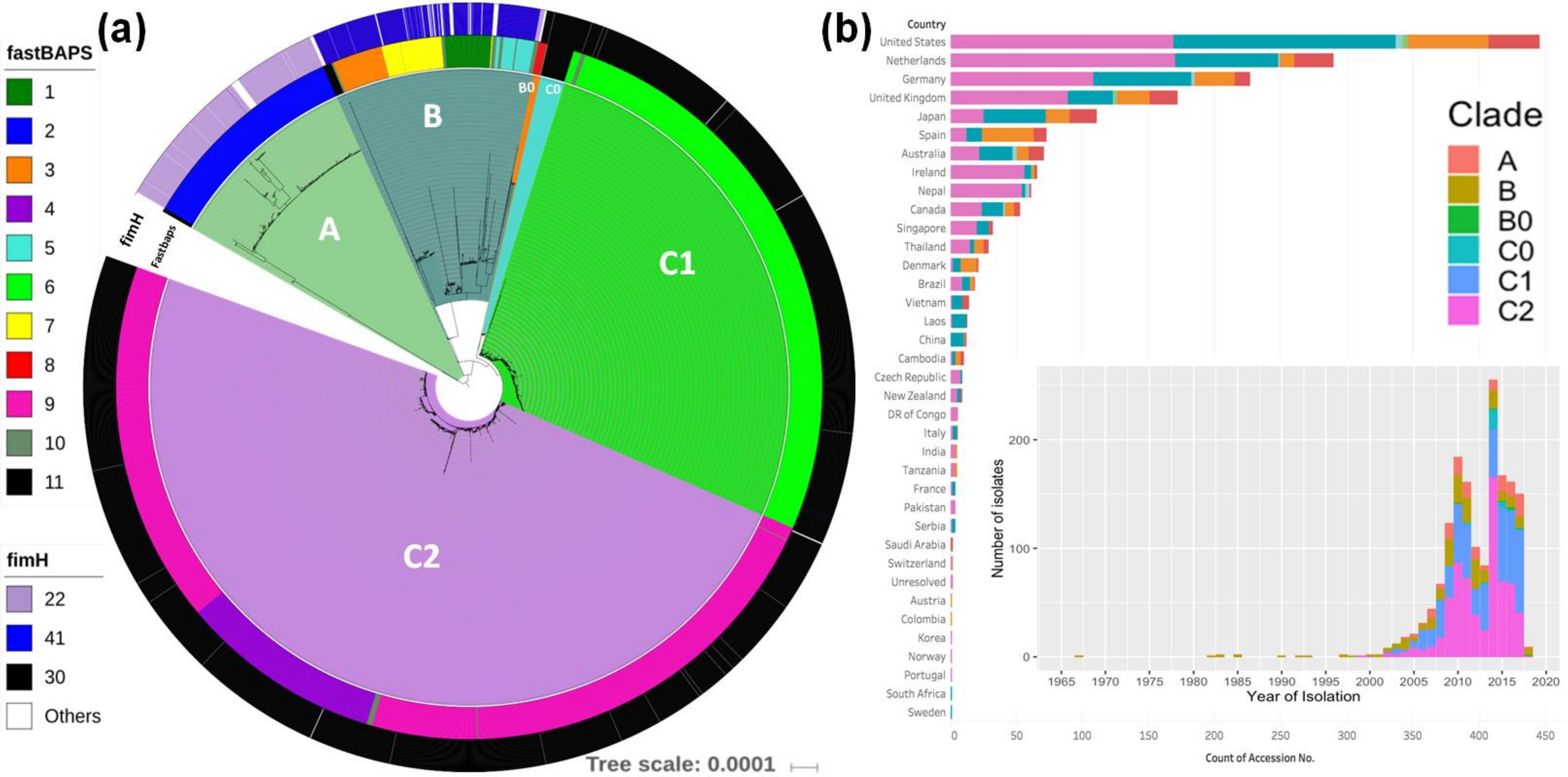


Figure 2. (a) A maximum likelihood phylogeny of 4,071 global ST131 and (b) the distribution of these samples across countries and over time. The phylogeny shows clades A (n=414, pale green), B (n=420, dark green), B0 (n=13, orange), C0 (n=52, blue), C1 (n=1,121, bright green) and C2 (n=2,051, purple). The phylogeny constructed with RAxML from the 30,029 chromosome-wide SNPs arising by mutation was visualized with iTol. The inner colored strip surrounding the tree represents the subgroups formed from Fastbaps clustering and the cluster (1-11) associated with each isolate. The outer colored strip surrounding the tree is the *fimH* allele (*H22* for clade A in pink, *H41* for clade B in blue, *H30* for clade C in black, with other alleles in white). The histograms in (b) show the distribution of sampling across countries, and that out of the 4,071 ST131 genomes isolated from 1999 to 2018, 2,051 belong to C2 (pink), with most samples coming from 2002-2017.

Subclades C1 and C2 were prevalent globally with no evidence of population structure, suggesting they have co-circulated for some time. This comes from the 819 isolates from Europe, 499 from North America, 294 from Asia, 80 from Oceania, 20 from South America, and 12 from Africa (Figure 3) - the remaining 2,347 (58%) had no geographic data (Table S3). C1_6 was more common in North America (OR=1.57, 95% CI 1.25-1.96, p=0.0004) but less so in Europe (OR=0.67, 95% CI 0.53-0.81, p=0.0004). C2_4 was more frequent in Asia (OR=1.75, 95% CI 1.18-2.56, p=0.019) and rarer in North America (OR=0.61, 95% CI 0.40-0.91, p=0.042).

Based on common ancestry with clade B, the origin of clade C was in North America because clade B samples from 1967-1997 were solely isolated in the USA until one isolate in Spain in 1998. This fitted previous work timing the origin of C to 1985, *fimH*30 to 1986, and the FQ-R C1/C2 ancestor to 1991 (Ludden, Decano et al 2019) (or 1986, Kallonen et al 2017), consistent with a North American source. However, the earliest C representative here from Norway in 1999 was a *bla*$_{\text{CTX-M}}$-negative *bla*$_{\text{TEM-1B}}$-positive FQ-R one from C2_9 (ERR1912633, Knudsen et al 2017).

The origin of C2_4 was unclear: although the earliest isolate was in 2008 from the USA, the most basal branches within C2_4 had isolates spanning a range of continents, and C2_4's long ancestral branch implied that it originated prior to 2008 (Figure S4). The most closely isolates to C2_4 were a group of 11 assigned to C2_10 that had limited country and year of isolation data bar one from the UK in 2011, one from the USA in 2009, one from the Netherlands (Figure S3). The next most closely related group was nine from C2_9 with no isolation data bar two from Australia in 2017.

Variable *bla*$_{\text{CTX-M-14/15/27}}$ gene prevalence across time, geography and ST131 subclades

Alignment of the 4,071 assemblies against *bla*$_{\text{CTX-M-14/15/27}}$ genes and CARD with BLAST showed that these ESBL genes were more common in C (75%) than A (45%) than B (4%) (Figure 4). Few isolates were both *bla*$_{\text{CTX-M-14/15}}$-positive (0.4%) or *bla*$_{\text{CTX-M-15/27}}$-positive (0.3%) (Table 2). Of the 2,408 *bla*$_{\text{CTX-M}}$-positive clade C samples, 1,782 were *bla*$_{\text{CTX-M-15}}$, 424 *bla*$_{\text{CTX-M-27}}$, 177 had *bla*$_{\text{CTX-M-14}}$, 15 *bla*$_{\text{CTX-M-14/15}}$, and 10 *bla*$_{\text{CTX-M-15/27}}$ (Figure S6) such that the rates were highest in C2_4 (93.8%) followed by C0 (90%), C2_9 (82.6%) and C1_6 (57%) (Figure S7). The earliest *bla*$_{\text{CTX-M}}$-positive clade C strain was from Canada in 2000 (ERR161284, C2_9, Petty et al 2014). 88% (339 of 386) of C2_4 and 81% (1,338 of 1,651) of C2_9 were *bla*$_{\text{CTX-M-15}}$-positive with limited geographic or temporal structure (Figure 4). This reiterated that the C2 ancestor was *bla*$_{\text{CTX-M-15}}$-positive whose gains of other *bla*$_{\text{CTX-M}}$ genes were likely rare local events.

Subclade	Number	***bla*$_{\text{CTX-M}}$ allele numbers per isolate**				
		14	**14+15**	**15**	**15+27**	**27**
A	414	65	1	66	2	51
B	420	7	1	9		1
C0	52			46		1
C1_6	1,113	149	6	59	3	418
C2_4	386	12	3	339	7	1
C2_9	1,651	16	6	1,338		4

Table 2. ST131 subclades' *bla*$_{\text{CTX-M-14/15/27}}$ genes. B0 (n=13) is not shown because it had no *bla*$_{\text{CTX-M}}$ genes.

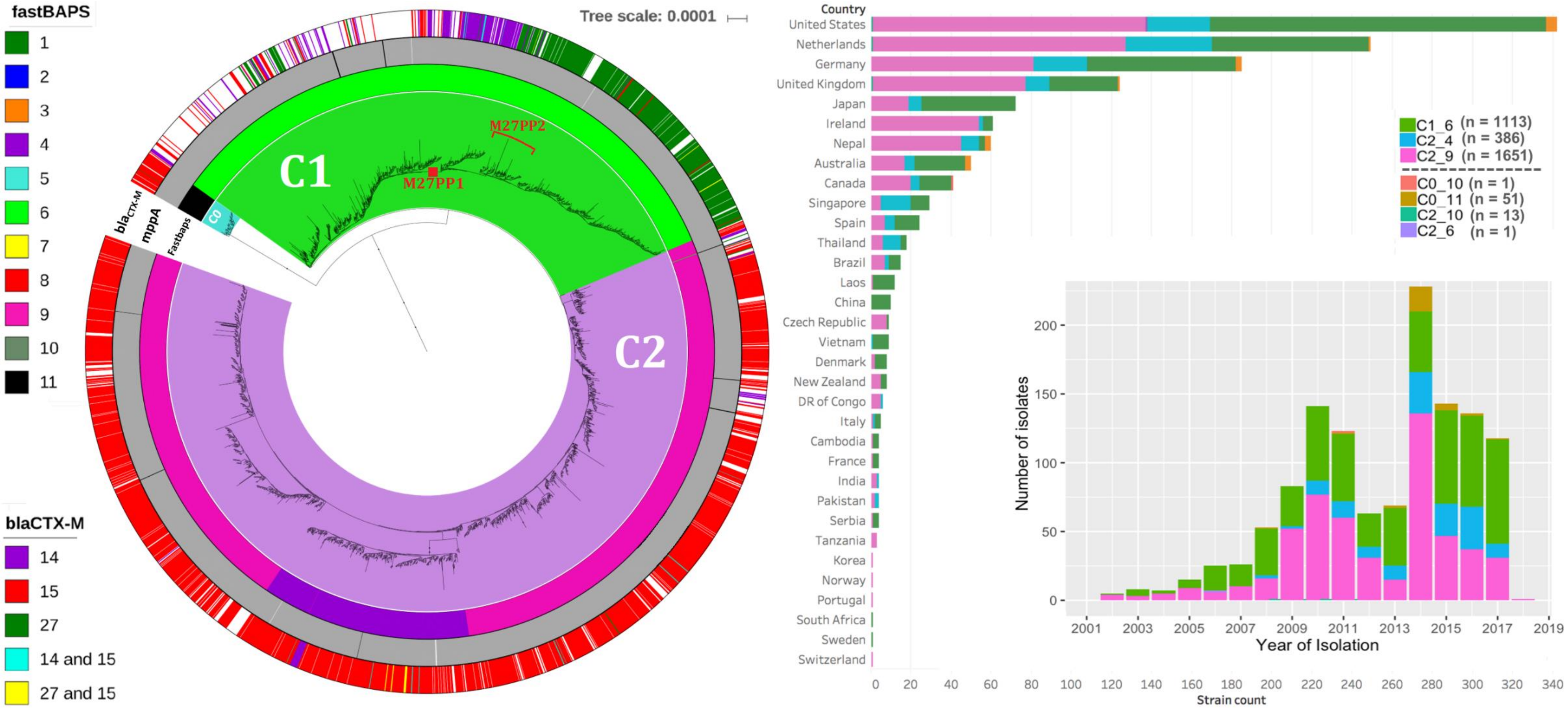


Figure 3. (a) A maximum likelihood phylogeny of 3,224 ST131 Clade C strains and (b) the distribution of these samples across countries and over time. The phylogeny shows C0 (n=52, blue), C1 (n=1,121, bright green) and C2 (n=2,051, purple). As per Figure 2, the phylogeny constructed with RAxML from the 30,029 chromosome-wide SNPs arising by mutation was visualized with iTol. The inner colored strip surrounding the tree represents the subgroups formed from Fastbaps clustering and the cluster (1-11) associated with each isolate. This indicated most C0 were in Fastbaps cluster 11 (n=52) with a single isolate in cluster 10 (grey). Of the 1,121 C1 samples, 1,113 formed Fastbaps cluster 6 (green) and eight were assigned cluster 10 (black). The C2 subclades corresponded to Fastbaps clusters 9 (C2_9, n=1,651 samples, pink) and 4 (C2_4, n=386, dark purple). The outer colored strip is the *bla*$_{\text{CTX-M}}$ allele: 2,416 had *bla*$_{\text{CTX-M-14}}$, *bla*$_{\text{CTX-M-15}}$, or *bla*$_{\text{CTX-M-27}}$ genes, 1,790 *bla*$_{\text{CTX-M-15}}$ (mainly C2), 177 *bla*$_{\text{CTX-M-14}}$ (mainly C1) and 424 *bla*$_{\text{CTX-M-27}}$ (mainly C1). The M27PP1 locus likely occurred on a single ancestral branch (red box) for 468 M27PP1-positive C1_6 isolates (though independent events occurred too) and a subgroup within these were mainly M27PP2-positive, suggesting a single gain ancestral to those highlighted. The histograms in (b) show sampling across countries, and that since 2002 both C1_6 (green) and C2_9 (pink) were common with the emergence of C2_4 (blue) and to a lesser extent C0 (biege).

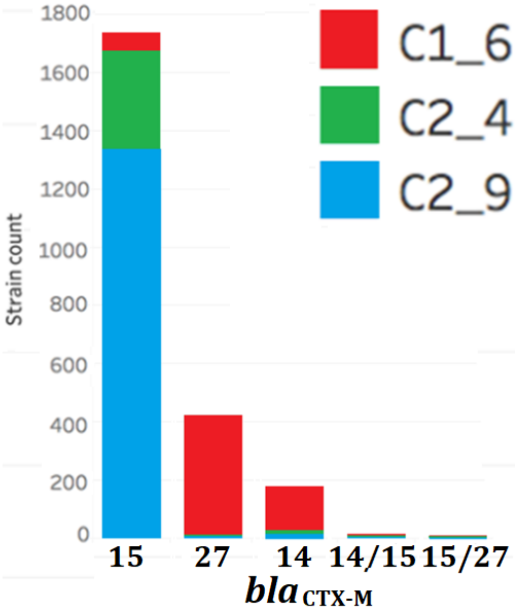


Figure 4. Frequencies of *bla*$_{\text{CTX-M}}$ alleles in C subclades C1_6 (red), C2_4 (green) and C2_9 (blue).

C1_6 had a different *bla*$_{\text{CTX-M}}$ gene rates to C: *bla*$_{\text{CTX-M-27}}$ (38%) was more common than *bla*$_{\text{CTX-M-14}}$ (14%) or *bla*$_{\text{CTX-M-15}}$ (6%) (Table 2). The earliest *bla*$_{\text{CTX-M-14}}$-positive C1 isolate was in 2002, followed by *bla*$_{\text{CTX-M-27}}$-positive ones from 2004 and *bla*$_{\text{CTX-M-15}}$-positive ones from 2008 (Figure S8). C1_6 was found in Japan only until detection in both China and Canada in 2005. *Bla*$_{\text{CTX-M-15}}$ was marginally more common in Europe (OR=3.3, 95% CI 1.38-8.70, p n/s), *bla*$_{\text{CTX-M-14}}$ was more common in Asia (OR=4.4, 95% CI 2.21-8.85, p=0.00007) as previously (Bevan et al 2017), and *bla*$_{\text{CTX-M-27}}$ was global.

A total 468 C1_6 represented the C1-M27 lineage because they had the M27PP1 prophage-like region (Table S3): 97% of these (397 of 410 with *bla*$_{\text{CTX}}$ gene data) were *bla*$_{\text{CTX-M-27}}$-positive, as expected (Matsumura et al 2016). Of 81 M27PP2-positive C1_6 found, 74 were M27PP1-positive and formed single lineage where all bar one with *bla*$_{\text{CTX}}$ gene data was *bla*$_{\text{CTX-M-27}}$-positive (Figure 3). This highlighted that the *bla*$_{\text{CTX-M-27}}$-positive C1-M27 ancestor acquired M27PP1 first, which coincided with wider spread of C1-M27 (Figure 3), and that the M27PP2 gain was independent in C1_6.

Here, the M27PP1 region was not unique to C1-M27: five C2 and one clade A samples had it. Likewise, 52 C2_9 and 7 clade B samples were M27PP2-positive (Table S3). The three M27PP1-positive C2_9 isolates with *bla*$_{\text{CTX}}$ gene data were all *bla*$_{\text{CTX-M-27}}$-positive and the earliest among them was from Japan in 2010 (DRR051016, Matsumura et al 2016). The set of 52 M27PP2-positive C2_9 isolates were paraphyletic, but all with *bla*$_{\text{CTX}}$ gene data were *bla*$_{\text{CTX-M-15}}$-positive, and only two were also M27PP1-positive. The earliest M27PP2-positive C2_9 isolate was in Nepal in 2010 (Stoesser et al 2014), before appearing Europe later (Matsumura et al 2017). This indicated that integration of M27PP1/2 and M27PP2 prophage-like regions were frequent, but only associated with a clonal expansion for C1-M27.

Screening of the 505,761 contigs from the 4,071 assemblies for *bla*$_{\text{CTX-M-14/15/27}}$-positive ones identified diverse structures and contexts based on annotation with MARA (Table S4). Most were plasmid-bound, though some included chromosomal insertions as observed using long reads (Decano et al 2019), such as a C2_4 isolate from Pakistan in 2012 (SRR1610051, Sheppard et al 2018) (Figure S9, Table S5).

Inter-clade but not intra-clade accessory genome divergence

Accessory genome composition varies across genetically distinct groups due to ecological niche specialisation (McNally et al 2016): this was supported here at the clade level across the 4,071 isolates by a positive correlation of pairwise core and accessory genome distances measured with Poppunk (Lees et al 2019) (Figure 5). This matched work on a *E. coli* dataset with 218 ST131 (Kallonen et al 2018), and was evident in a higher shell gene number in B compared to A and C (Table 3), suggesting higher diversity in B (Figure S10).

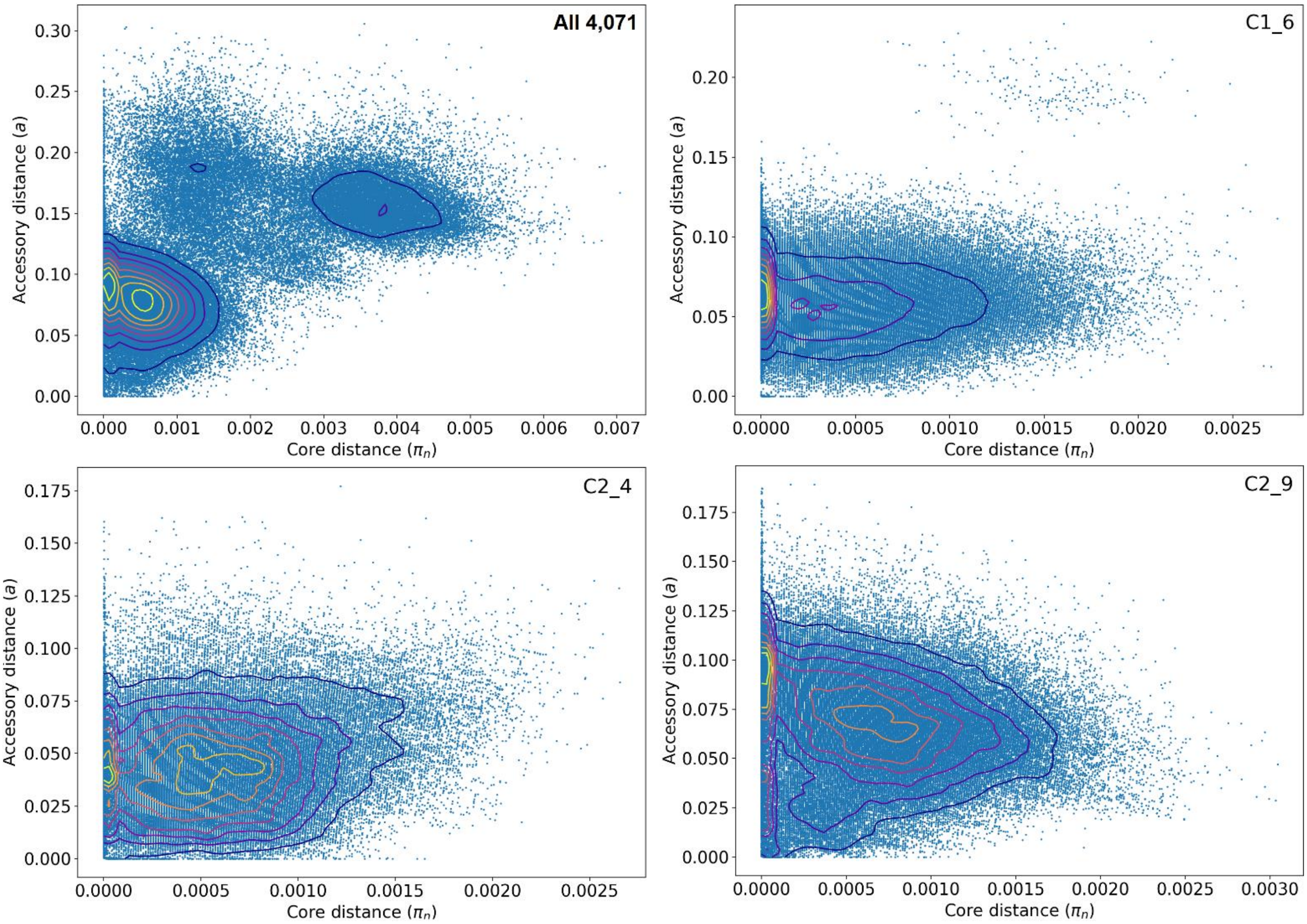


Figure 5. The distribution of core (π, x-axis) and accessory pairwise genome distances (a, y-axis) with blue dots indicating isolate pairs and the contours indicating dot density (higher in yellow). Top left: All 4,071 assemblies displayed pairwise differences such that the contours indicated the three main clades: A at π=0.0038, a=0.15; B at π=0.0014, a=0.18; C at both π=0.0005, a=0.08 and π=0.0001, a=0.09. Top right: 1,113 C1_6 assemblies had a peaks mainly at $\pi\leq$0.001, a=0.06. Bottom left: 386 C2_4 assemblies had peaks at π=0.0006, a=0.045 and π=0.0001, a=0.040. Bottom right: 1,651 C2_9 assemblies had peaks at π=0.0007, a=0.065 and π=0.0, a=0.090. Results for 2,416 *bla*$_{\text{CTX-M}}$-positive clade C assemblies and 52 C0 assemblies were similar. Within C1_6, C2_4, C2_9, isolates had more diverse accessory genomes compared to their core ones.

Fitting this clade-level accessory genome differentiation was a more open pangenome (*alpha*) in B (0.762) than A (0.807) or C (0.806), compared to a higher *alpha* for the whole collection (0.823). Previously, 648 clade C isolates had a more open pangenome than 140 from B and 70 from A (McNally et al 2019), but *alpha* was higher for small (<250) sample sizes here (Figure 6) as indicated before (Park et al 2019). *Alpha* estimates averaged across the sample size placed C as more open than A or B (Table S6), which was also found when adjusting for the differing clade sample sizes, highlighting a partial dependence of *alpha* on sample size. Combining clades A, B and C as pairs showed that B with C together had a more open pangenome (0.813) than A with C (0.835), pointing to less population structure between B and C relative to A (Table S6), in line with previous work.

Clades	All	A	B	C	C1_6	C2_4	C2_9
#isolates	4,071	414	433	3,200	1,113	384	1,651
alpha	0.823	0.807	0.762	0.806	0.755	0.696	0.822
Average *alpha*	0.812	0.780	0.796	0.766	0.749	0.754	0.799
Total genes	26,479	12,163	16,323	21,304	16,490	10,322	15,485
Core genes	3,712	3,798	3,771	3,916	3,843	4,109	4,019
Soft core genes	242	292	281	334	424	380	354
Shell genes	1,018	764	1,437	731	571	642	566
Cloud genes	21,507	7,309	10,834	16,323	11,652	5,191	10,546

Table 3. The pangenome composition of ST131 clades and subclades showed stable core, soft core and shell genomes with open pangenomes (*alpha*). The average *alpha* was determined for sample sizes from 30 to the maximum per group (see Figure 6): for all ST131 this was 0.812±0.024 (mean ± standard deviation); for A 0.780±0.045; B 0.796±0.046; C 0.766±0.031; C1_6 0.749±0.017; C2_4 0.754±0.077; and C2_9 0.799±0.038. The clades and subclades cloud gene rates correlated with sample size following a power law model. B0 (n=13) was included with B. Eight C1_10 and 16 C2_6 or C2_10 isolates not assigned to clear Fastbaps clusters were not examined here.

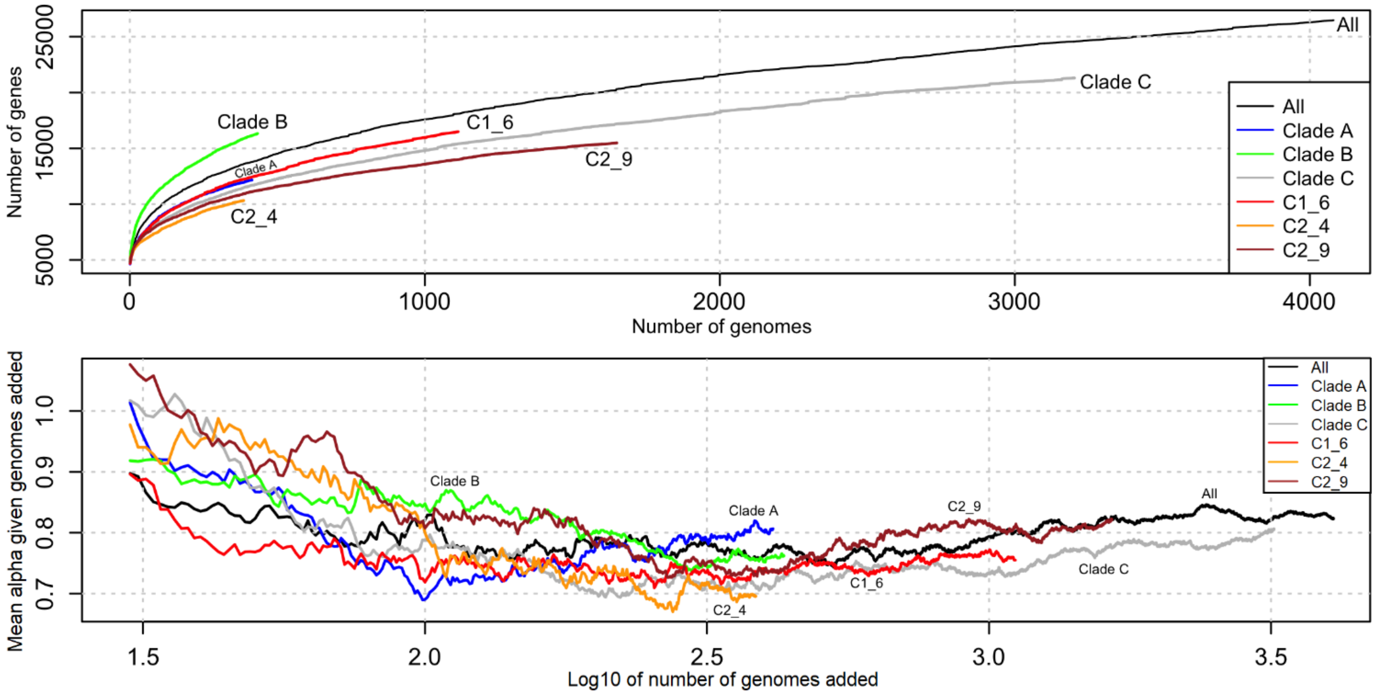


Figure 6. Top: The average number of genes in the ST131 pangenome (y-axis) increased as the 4,071 genomes were added (x-axis) indicating an open pangenome for the whole collection (black), clades and subclades: A (blue), B (green), C (grey), C1_6 (red), C2_4 (orange) and C2_9 (brown). Below: *Alpha* varied with numbers of genomes sampled (shown here for >30 genomes) and was more independent from sample number once the number of genomes examined about >250. Note that the x-axis' log$_{10}$ scale.

Within the C subclades, the pairwise core and accessory genome distances were not correlated: the accessory genomes varied extensively even with nearly identical core genomes (Figure 5). C2_4 had a more open pangenome (0.696) than C1_6 (0.755) or C2_9 (0.822), which was evident when adjusting for the differing sample sizes (Figure 6), though the average *alpha* placed C1_6 (0.749) and C2_4 (0.754) as about equally more open than C2_9 (0.799, Table 3).

Environmental context may drive NFDS, resulting in non-essential accessory genes remaining at intermediate (shell gene) frequencies (McNally et al 2019). The observed shell gene numbers were compared the expected values adjusted for sample number and gene frequency category change to investigate shell gene overlap across clades and subclades (see Methods). Pooled groups with divergent accessory genomes due to population structure should have more shell genes, whereas reduced population structure resulting in more similar accessory genomes should result in fewer shell genes. B and C together had 6% less shell genes, whereas A had an excess of 1% when combined with C and 6% with B (Table S6), consistent with less population structure between B and C. Within subclades, there was a small shell gene excess for C1_6 combined with C2_9 (3%), but C2_4's shell gene composition differed from both C1_6 (22% excess) and C2_9 (23% excess, Table S6). The same trend was observed when C2_4 was mixed with A (41% excess) or B (5%) in contrast to C1_6 (16% with A, -8% with B) and C2_9 (16% with A, -11% with B), indicative of a more unique shell gene set in C2_4. This posited that even though the high accessory genome diversity within subclades was independent of core genome composition, NFDS may result in a more open pangenome drive for new lineages like C2_4, which may mean C2_4's shell gene set may become more similar to other bacteria sharing the same environments.

Discussion

By collating all available ST131 genomes to produce 4,071 high-quality draft assemblies, we reconstructed their phylogenetic relationships to show that ST131 was dominated by subclades C1 and C2. For isolates with *bla*$_{\text{CTX-M}}$ gene data, C2 was 98% *bla*$_{\text{CTX-M-15}}$-positive in contrast to C1 that had either *bla*$_{\text{CTX-M-27}}$ (66%) or *bla*$_{\text{CTX-M-14}}$ (24%) genes. Although the C1 ancestor may have been *bla*$_{\text{CTX-M-14}}$-positive, the patterns here and *bla*$_{\text{CTX-M-27}}$'s higher ceftazidime resistance due to a D240G substitution also in *bla*$_{\text{CTX-M-15}}$ (Bonnet et al 2003) indicated it will become more common in C1. Although the subclades had different origins and ancestral ESBL gene compositions, they have both become common globally since 2002 with relatively consistent frequencies and minor differences in rates due to differing evolutionary patterns after likely emerging in North America (Stoesser et al 2016). This worldwide co-circulation coupled with NFDS suggested newer lineages with a distinct accessory genome will likely become globally disseminated, with implications for infection control if they have altered host adhesion abilities (like *fimH30*, Paul et al 2013) or AMR variants (like FQ-R or *bla*$_{\text{CTX-M-15}}$). This was highlighted by the emergence of C2_4, and many other contemporary examples like *bla*$_{\text{OXA-48}}$-producing ST131 (Mahon et al 2019). Tracking plasmid, MGEs and ESBL genes must be a key component of disease monitoring to consider potential future bacterial outbreaks' spectrum of AMR.

Horizontal DNA transfer allows *E. coli* adapt to new ecological niches and contributes to its dynamic accessory genome (Welch et al 2002) where the cloud gene number increases with isolate number and diversity. Our analysis of this larger collection's core (3,712) and accessory (22,525) genes extended previous work showing that 283 predominantly ST131 samples had 16,236 genes in an open pangenome with a core of 3,079 genes (Salipante et al 2015), 21% less than the core genome here. Nonetheless, ST131's accessory genome may be streamlined: a more genetically diverse set of 1,509 *E. coli* including 266 ST131 had a core genome of 1,744 genes and a 62,753 cloud genes (Kallonen et al 2017), and an *E. coli-Shigella* core genome had 2,608 genes among a total of 128,193 genes (Park et al 2019).

The NFDS hypothesis posits that intermediate-frequency shell genes assist with adaptation to new hosts and environments (McNally et al 2019). Pangenome openness and shell gene sharing across clades supported inter-clade structure resulting from ecological specialisation (Medini et al 2005), with clade A more different to B and C. Within subclades, isolates with minimal core genome differences could have divergent accessory genomes, implying that plasmid, ESBL gene and MGE changes may be detected better using pangenomic approaches than assessing the core alone (Lanza et al 2014).

Global coordination of data processing and bioinformatic interpretation can help identify, trace and control disease outbreaks (Pijnacker et al 2019), for which resolving recent transmissions may be limited by sampling (Hanage 2019). Expanding numbers of non-human isolates and the diversity of geographic regions sampled would help clarify environmental sources of *E. coli*'s ESBL genes, for which there was no evidence of retail meat (Randall et al 2017) or livestock (Ludden et al 2019) as sources for BSIs thus far, though transfer of bacteria may occur (Roer et al 2019). Better epidemiological information for genome-sequencing (Raven et al 2019) could allow inference of adaptations across lineages (Azarian et al 2018), such as *bla*$_{\text{CTX-M-15}}$-positive ST8313, a putative descendant of ST131 C2 (Findlay et al 2019).

Methods

Study selection and data extraction

Data on 4,870 *E. coli* ST131 genomes and linked metadata was collected using an automated text-mining algorithm using a Python implementation of Selenium (Selenium-python.readthedocs.io) from Enterobase (https://enterobase.warwick.ac.uk, Alikhan et al 2018) on the 10th of September 2018 as previously described (Kinderis et al 2018). This was used to download read libraries the European Nucleotide Archive (ENA) (www.ebi.ac.uk/ena) (Harrison et al 2018) and NCBI Short Read Archive (SRA) databases as FASTQ files, restricted to complete libraries not labelled as "traces" (Figure 1). Of the initial 4,870 read libraries, 4,264 were paired-end (PE) Illumina HiSeq ones and four were PacBio, in addition to the PacBio-sequenced NCTC13441 genome used as a reference in this study. 495 libraries predominantly from Illumina MiSeq platforms were not examined to avoid platform-specific artefacts.

Illumina HiSeq read data quality control, trimming and correction

Of the above 4,264 PE Illumina HiSeq read libraries, 4,147 passed stringent quality control. This was implemented using Fastp v0.12.3 (Chen et al 2018) to trim sequencing adapters, remove reads with low base quality scores (phred score <30) or ambiguous (N) bases, correct mismatched base pairs in overlapped regions and cut poly-G tracts at 3' ends (Table S2). Individual bases in reads were corrected by BayesHammer in SPAdes v3.11 (Nikolenko et al 2013). Quality control metrics were examined at each step: across the whole collection as a batch report using MultiQC v1.4 (Ewels et al 2016) and on individual FASTQ files using FastQC v0.11.8 (www.bioinformatics.babraham.ac.uk/projects/fastqc/). 117 (2.7%) Illumina HiSeq libraries did not quality control.

Illumina HiSeq read library genome assembly

The 4,147 Illumina HiSeq libraries passing quality control were *de novo* assembled using Unicycler v4.6 in bold mode to merge contigs where possible (Wick et al 2017). This used SPAdes v3.12 (Bankevich et al 2012) to generate an initial assembly polished by Pilon v1.22 (Walker et al 2014), which ran iteratively until no further corrections were required by the optimal assembly. This approach was similar to Enterobase's (Alikhan et al 2018), though Enterobase used BBMap in BBTools (Bushnell 2016), SPAdes v3.10 and BWA (Li and Durbin 2010) during assembly (Zhou et al 2019).

Reference PacBio genome quality control and assembly

The ST131 reference genome NCTC13441 was isolated in the UK in 2003 and was in subclade C2 (Brodrick et al 2018). It had a 5,174,631 bp chromosome with 4,983 protein-coding genes and one pEK499-like type IncFIA/FIIA plasmid with two *bla*$_{\text{CTX-M-15}}$ gene copies (accession ERS530440). Although four further PacBio read libraries were initially included to test genome assembly contiguity and ESBL gene context using longer read libraries, only one passed assembly annotation screening (AR_0058, accession SRR5749732, Sheppard et al 2018). Its adapters were removed using Cutadapt v1.18 (Martin 2011) followed by excluding duplicate reads with Unicycler v0.4.6. Base correction was implemented during genome assembly with Unicycler via SPAdes v3.12, and the genome assembly was iteratively polished by Racon v1.3.1 until no further corrections were required (Vaser et al 2017). This 5,132,452 bp assembly had five contigs, 5,506 genes and no IS*Ecp1*, and was assigned to C1.

Quality checking and annotation identifies 4,071 genome assemblies for investigation

For the 4,147 Illumina HiSeq assemblies and single PacBio assembly, quality was verified with Quast v5.0 (Gurevich et al 2013) based on the N50, numbers of predicted genes and open-reading frames, and numbers of contigs with mis-assemblies. The quality of the short read *de novo* assemblies was comparable to previous work whose requirements required assembly length in the range 3.7-6.4 Mb with <800 contigs and <5% low-quality sites (Zhou et al 2019). Initial annotation of 4,147 Illumina HiSeq assemblies using Prokka v1.10 (Seeman 2014) suggested 77 assemblies had a distinct gene composition and should be excluded because they were either genetically divergent, did not assemble adequately, or had sub-standard read libraries. As a result, 4,070 Illumina HiSeq genome assemblies were selected (Table S2) and aligned against the reference genome NCTC13441 and PacBio assembly AR_0058 (Table S3). This identified 4,829 genes on average per assembly (range 3,942 to 5,749, Figure S1). The variation in numbers of genes per assembly was largely explained by the total assembly length (r^2=0.959). 53% of the 4,071 had no source data and 12% of the remainder had a non-human source.

Pangenome analysis to identify the core and accessory genomes

We created a pangenome based on the 4,072 annotation files using Roary v3.11.2 (Page et al 2015) with a 100% BLAST v2.6.0 identity threshold using the MAFFT v7.310 setting (Katoh & Standley 2013). The resulting concatenated core CDS alignment spanning 1,244,619 bases and 3,712 genes scaffolded using NCTC13441 was used for core genome analyses. 242 soft core genes were found that may have been due to assembly errors or other artefacts. Pangenomes for each clade, C subclade and various combinations were also created for accessory genome evaluation.

Phylogenetic reconstruction, population structure and subclade assignment

A maximum likelihood phylogeny was generated based on the core genome alignment of 4,071 genome assemblies with NCTC13441 as a reference across 30,029 SNPs (with 26,946 alignment patterns) for 50 iterations of RAxML v8.2.11 with a GTR model and gamma substitution rate heterogeneity (Stamatakis 2014). 88% (3,585) of the genome assemblies were genetically unique. The total execution time on an Ubuntu v16.04 computer server with 256 Gb RAM using 52 threads was 24.4 days. Phylogenies were drawn and annotated using iTol v4.3.2.

Clade classifications were initially based on published ST131 *fimH* phylogenetic analyses associating clade A with *fimH41*, B with *fimH22*, B0 with *fimH27*, and C with *fimH30* (Price et al 2013). To classify the large number of isolates in the C subclades, we clustered the 30,029 core genome SNPs as a sparse matrix using a hierarchical Bayesian clustering algorithm implemented in Fastbaps v1.0 (Fast Hierarchical Bayesian Analysis of Population Structure, Tonkin-Hill et al 2018) in R v3.5.3 with packages ape v5.3, ggplot2 v3.1.1, ggtree v1.14.6 (Yu et al 2017), maps v3.3.0 and phytools v6.60. This used default parameters except for a Dirichlet prior variance of 0.006. C1-M27 was identified based on the prophage-like 11,894 bp M27PP1 region, with the 19,352 bp M27PP2 also examined (LC209430 with M27PP2 5’ and M27PP1 3’, Matsumura et al 2016).

ESBL gene screening and contig visualisation

We screened for contigs with *bla*$_{\text{CTX-M-14/15/27}}$ genes across the 4,071 assemblies' 505,761 contigs using BLASTn (Altschul et al 1990) alignment of these three genes individually, and the Comprehensive Antibiotic Resistance Database (CARD) v3.0 requiring 100% identity for any match with a contig. Selected *bla*$_{\text{CTX-M-14/15/27}}$-positive contigs were visualised using the Multiple Antibiotic Resistance Annotator (MARA) (Partridge & Tsafnat 2018), R v3.5.2 and EasyFig v2.2.2 (Sullivan 2011) to examine the local contig, MGE and gene annotation. A minority of isolates had incomplete contigs due to the small contig lengths. Frequencies of ST131 clades, subclades and their *bla*$_{\text{CTX-M-14/15/27}}$ genes across geographic regions and time were examined with R packages dplyr v8.0.1, forcats v0.4.0, ggplot2 v3.1.1, ggridges v5.1, grid v3.5.2, plotly v4.9.0, plyr v1.8.4, purr v0.3.2, questionr v0.7.0, readr v1.3.1, rentrez v1.2.1, stringr v1.4.0, tibble v2.1.1, tidyr v0.8.3, tidyverse v1.2.1 and XML v3.98-1.19.

Pangenome analysis to find shared and unique accessory genomes

Roary assigned numbers of genes to the core (c) and the accessory genomes, including the soft core (s), shell (a) and cloud (d). The expected shell gene number ($E[a_p]$) from the Roary output for a given pooled set of samples (p) originally from groups $i=1..k$ was determined based on the shell gene number of group i (a_i) weighted by the sample size (n_i) corrected for the deficit in core (c_i) and soft core (s_i) gene numbers: $\boldsymbol{E[a_p] = \frac{\sum_{i=1}^{k} n_i a_i}{\sum_{i=1}^{k} n_i} - \frac{\sum_{i=1}^{k} n_i (c_i - c_p)}{\sum_{i=1}^{k} n_i} - \frac{\sum_{i=1}^{k} n_i (s_i - s_p)}{\sum_{i=1}^{k} n_i}}$. The excess fraction of shell genes observed was $(a_p - E[a_p])/E[a_p]$. Similarly, the expected cloud gene number $E[d_p]$ was computed from the cloud gene number of group i (d_i) weighted by the sample size (n_i) adjusted for the difference in core (c_i), soft core (s_i) and shell (a_i) gene numbers: $\boldsymbol{E[d_p] = \frac{\sum_{i=1}^{k} n_i d_i}{\sum_{i=1}^{k} n_i} - \frac{\sum_{i=1}^{k} n_i (c_i - c_p)}{\sum_{i=1}^{k} n_i} - \frac{\sum_{i=1}^{k} n_i (s_i - s_p)}{\sum_{i=1}^{k} n_i} - \frac{\sum_{i=1}^{k} n_i (a_i - a_p)}{\sum_{i=1}^{k} n_i}}$. The excess fraction of cloud genes observed was $(d_p - E[d_p])/E[d_p]$.

Pangenome openness (*alpha*) was quantified from Roary results as $\Delta n = kN^{-alpha}$ where Δn was the number of new genes across N genome assemblies with n genes in total (Park et al 2019) with R packages poweRlaw v0.70.2, igraph v1.2.4.1 and VGAM v1.1.1 (Figure S11). This power-law regression approximated Heaps' law such that an open pangenome has *alpha* < 1 and a closed one *alpha* > 1 (Tettelin et al 2008). Previously, diverse *E. coli* had *alpha = 0.625* where *alpha* had a partial negative correlation with N (Park et al 2019). Similarly, *alpha* was ~0.877 for ST131 clade C, ~0.898 for B, ~0.958 for A, and ~0.951 for all ST131, suggesting *alpha* was higher when genetically distinct clades were combined (McNally et al 2019).

Accessory genome composition across clades and subclades

The relative pairwise genetic distances of the core (π) and accessory (a) genomes were compared for each clade, each C subclade and all *bla*$_{\text{CTX-M}}$-positive clade C samples using Poppunk (Population Partitioning Using Nucleotide Kmers), which can distinguish closely related genomes (Lees et al 2019). Poppunk used variable length pangenome k-mer comparisons with Mash v2.1 (Ondov et al 2016) and a Gaussian mixture model to examine the correlation of π and a per sample pair. This annotation- and alignment-free strategy complemented the approaches of Fastbaps, RAxML and Roary.

Acknowledgements

We thank Marius Kinderis (Dublin City University, Ireland) for assistance in implementing the text mining algorithm. This project was funded by a DCU O'Hare Ph.D. fellowship and a DCU Enhancing Performance grant.

Availability of data and materials

All raw sequence data (reads and/or assembled genomes) for the *E. coli* genomes analysed in this publication are in Table S3, with their Bioproject IDs and associated study DOIs in Table S1. The genome assemblies of the 4,071 *E. coli* ST131 are on Zenodo at DOI: 10.5281/zenodo.3341533 with the annotation files at 10.5281/zenodo.3341535. An interactive version of the phylogeny generated by Poppunk for the 4,071 ST131 assemblies is on MicroReact at https://microreact.org/project/oD6K_fL2d - this includes a Newick tree file available for download.

Supplementary Figures

Figure S1. Annotation of the 4,071 ST131 genomes (along with NCTC13441) using Prokka (a) identified 4,829 genes on average per assembly with a minimum of 3,942 and maximum of 5,749. (b) Of 26,479 gene clusters detected using Roary, 3,712 comprised the core genome (blue) spanning 1,244,619 bases based on pangenome analysis with 242 soft core genes (yellow), 1,018 shell genes (navy) and 21,507 cloud genes (light blue).

Figure S2. Pangenome analysis of the effect of increasing the number of genomes (x-axis) showed (left) that few new genes were discovered (black line), but that the number of unique genes associated with the cloud gene set increased consistently (dashed line). (Right) The core genome composition across all 4,071 assemblies was stable once >200 genomes were included (“Conserved genes”, solid line), whereas the total number of genes increased without plateauing (dashed line). There was a median of 2.1 additional genes per additional isolate in this collection.

Figure S3. Hierarchical sub-clustering of 4,071 strains using Fastbaps based on 30,029 SNPs. Clusters are indicated by numerical numbers in bold red on above the blue bars. There were nine major clusters found and two (clusters 10 and 11) were dispersed among the collection**.**

Figure S4. Phylogeny of 382 C2_4 strains rooted using C2_9 isolates (not shown). The first C2_4 isolate found was in 2008 in the USA, but C2_4’s long ancestral branch imply that it arose several years prior to this. 90% (349) of C2_4 isolates had a *bla*$_{CTX-M-15}$ gene. The outer ring shows isolates’ continents of origin, with Africa in orange (from the Democratic Republic of Congo); Asia in dark green (India n=1, Japan n=6, Nepal n=9, Pakistan n=2, Singapore n=15, Thailand n=9); Europe in bright green (Germany n=27, Ireland n=2, Italy n=1, the Netherlands n=43, Spain n=5, UK n=12, Vietnam n=1); Oceania in yellow (Australia n=5); and South America in light blue (both from Brazil). Four additional C2_4 isolates (n=386 in total) with long branches are not shown for clarity.

Figure S5. The annual fraction of samples from (top) C1_6 (red), C2_4 (green) and C2_9 (blue) and (bottom) clades A (red), B (beige), B0 (green), C0 (light blue), C1 (dark blue) and C2 (mauve) showed consistent levels with no clear evidence of periodic radiation and fixation of new lineages.

Figure S6. Discovery of at least one C1_6 (red), C2_4 (green) or C2_9 (blue) isolate per country (x-axis) per year from 2000-2019 (y-axis).

Figure S7. Bubble graph of C1_6, C2_4 and C2_9 colored by their *bla*$_{CTX-M}$ alleles (*bla*$_{CTX-M-14}$ in orange, *bla*$_{CTX-M-15}$ in turquoise, *bla*$_{CTX-M-27}$ in green, and *bla*$_{CTX-M-14/15}$ together in red) with country of origin shown within each bubble (where known). The area of each bubble corresponds with the relative frequency of that particular combination of subclade, country and *bla*$_{CTX-M}$ allele.

Figure S8. Distribution of *bla*$_{CTX-M-14}$ (purple circle), *bla*$_{CTX-M-15}$ (red cross) and *bla*$_{CTX-M-27}$ (green square) across C1_6, their geographic origin (country) over time (isolation year) during 2002-2017.

Figure S9. Representative examples of the *bla*$_{\text{CTX-M-14/15/27}}$-positive contigs' ESBL genes and MGEs annotated using Prokka and MARA. Some contigs were too short to show additional annotations, which can be estimated based on the longer contigs. (a) C1_6 had the most frequent incidence of *bla*$_{\text{CTX-M-14}}$-positive contigs that were typically in IS*Ecp1-bla*$_{\textit{CTX-M-14}}$-IS*903B* TUs, though with variations such as a 3' IS*903C* element instead. (b) C2 tended to have a *bla*$_{\text{CTX-M-15}}$ gene flanked by a 5' IS*Ecp1* and a Tn2 or the orf-477-Tn2 in tandem at the 3' as a 2,971 bp IS*Ecp1-bla*$_{\textit{CTX-M-15}}$-orf477Δ-*Tn2* TU. These were most likely on an IncF plasmid for the plasmid-encoded variants. (c) C1_6 had the highest incidence of *bla*$_{\text{CTX-M-127}}$-positive contigs that typically had a similar IS*Ecp1-bla*$_{\textit{CTX-M-14}}$-IS*903B* structure as shown in (a).

Figure S10. A phylogeny of all ST131 (left) with the corresponding pangenome-wide gene presence (blue) and absence (white) frequencies per isolate represented for each of the 26,479 genes discovered. In the latter matrix, the 3,712 core genes are shown first on the left side, followed by the 242 soft core genes, 1,018 shell genes in 15-95% of samples, and 21,507 cloud genes in <15% of samples. Clades B (top and bottom clusters separated by green lines) and A (second from bottom separated from C by a red line) had core genome differences compared to C (middle bounded by green and red lines).

Figure S11. Across the 4,071 genome assemblies, *alpha* was estimated as 0.8231 such that the median number of new genes added per isolated was 2.1: the blue line indicates the regression slope *alpha*.

Supplementary Tables

Table S1. The 170 BioProject accession numbers with study and source information for the 4,071 ST131 genomes examined.

Table S2. Quality statistics of the 8,140 Illumina read libraries and single PacBio read library associated with the 4,071 assemblies, including the proportion of duplicate reads, average GC content, mean sequence length (bp) and total number of sequences (millions) per library.

Table S3. Metadata of 4,071 high quality ST131 genomes assessed. This includes the accession numbers, strain name, Bioproject accession numbers, *fimH* allele, clade assignment, Fastbaps cluster, subclade, *bla*$_{\text{CTX-M}}$ allele, ISEcp1 count, year of isolation, city of isolation, country of isolation, continent of isolation, source niche type, M27PP1 presence, M27PP2 presence, total assembly length (bp), assembly N50, number of contigs in assembly, largest contig length (bp), GC content (%) and the number of genes in the assembly.

Table S4. Selected (n=28) isolated whose *bla*$_{\text{CTX-M}}$-positive contigs were visualised with MARA, including three from C0, nine from C1, two from C2_10, six from C2_4, and eight from C2_9.

Table S5. The contig identifiers of the 2,623 ST131 assemblies with IS*Ecp1* elements along with the IS*Ecp1* lengths (maximum 1,655 bp).

Table S6. Pangenome analysis for the collection and associated subsets, including the numbers of core (*c*), soft core (*s*), shell (*a*) and cloud (*d*) genes; the deficit in core, soft core and shell genes; the expected number of shell genes (*E[a]*) and percentage excess shell genes (*(a-E[a])/E[a]*); the expected number of cloud genes (*E[d]*) and percentage excess cloud genes (*(d-E[d])/E[d]*); and pangenome openness (*alpha*).

References

Alikhan NF, Zhou Z,Sergeant MJ, Achtman M. A genomic overview of the population structure of *Salmonella*. 2018 PLoS Genet 14 (4): e1007261

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990 215(3):403-10

Azarian T, Martinez PP, Arnold BJ, Grant LR, Corander J, Fraser C, Croucher NJ, Hammitt LL, Reid R, Santosham M, Weatherholtz RC, Bentley SD, O'Brien KL, Lipsitch M, Hanage WP. Predicting evolution using frequency-dependent selection in bacterial populations. 2018 Biorxiv doi: http://dx.doi.org/10.1101/420315

Banerjee R, Johnson JR. A new clone sweeps clean: the enigmatic emergence of *Escherichia coli* sequence type 131. Antimicrob Agents Chemother 2014 58:4997–5004. doi:10.1128/AAC.02824-14.

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012 19(5):455-77. doi: 10.1089/cmb.2012.0021

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:e00347. doi:10.1128/mBio.00347-16

Bevan ER, Jones AM, Hawkey PM. Global epidemiology of CTX-M β-lactamases: temporal and geographical shifts in genotype. J Antimicrob Chemother. 2017 72(8):2145-2155. doi: 10.1093/jac/dkx146

Bonnet R, Recule C, Baraduc R, Chanal C, Sirot D, De Champs C, Sirot J. Effect of D240G substitution in a novel ESBL CTX-M-27. J Antimicrob Chemother. 2003 52(1):29-35

Brodrick HJ, Raven KE, Kallonen T, Jamrozy D, Blane B, Brown NM, Martin V, Török ME, Parkhill J, Peacock SJ. Longitudinal genomic surveillance of multidrug-resistant *Escherichia coli* carriage in a long-term care facility in the United Kingdom. Genome Med. 2017 9(1):70. doi: 10.1186/s13073-017-0457-6

Bushnell B. BBMap short read aligner. 2016.

Calhau V, Ribeiro G, Mendonça N, Da Silva GJ. Prevalent combination of virulence and plasmidic-encoded resistance in ST 131 *Escherichia coli* strains. Virulence. 2013 4(8):726-9. doi: 10.4161/viru.26552

Cantón R, González-Alba JM, Galán JC. CTX-M Enzymes: Origin and Diffusion. Front Microbiol. 2012 3:110. doi: 10.3389/fmicb.2012.00110.

Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018 34(17):i884-i890. doi: 10.1093/bioinformatics/bty560

Croucher NJ, Didelot X. The application of genomics to tracing bacterial pathogen transmission. Current Opinion in Microbiology 2015 23:62–67. doi: 10.1016/j.mib.2014.11.004.

Dautzenberg MJ, Haverkate MR, Bonten MJ, Bootsma MC. Epidemic potential of *Escherichia coli* ST131 and *Klebsiella pneumoniae* ST258: a systematic review and meta-analysis. BMJ Open. 2016 6(3):e009971. doi: 10.1136/bmjopen-2015-009971.

de Kraker MEA, Jarlier V, Monen JCM, Heuer OE, van de Sande N, Grundmann H. 2013. The changing epidemiology of bacteraemias in Europe: trends from the European Antimicrobial Resistance Surveillance System. Clin Microbiol Infect 19:860–868. doi: 10.1111/1469-0691.12028

Decano AG, Ludden C, Feltwell T, Judge K, Parkhill J, Downing T. Complete assembly of *Escherichia coli* ST131 genomes using long reads demonstrates antibiotic resistance gene variation within diverse plasmid and chromosomal contexts. mSphere 4(3):e00130-19 DOI: 10.1128/mSphere.00130-19

Downing T. Tackling drug resistant infection outbreaks of global pandemic *Escherichia coli* ST131 using evolutionary and epidemiological genomics. Microorganisms 2015 3(2):236-267 doi: 10.3390/microorganisms3020236.

ECDC, European Centre for Disease Prevention and Control. European Centre for Disease Prevention and Control. Antimicrobial resistance surveillance in Europe 2015. 2017. Annual Report of the European Antimicrobial Resistance Surveillance Network (EARS-Net). Stockholm: ECDC.

Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics. 2016 32(19):3047-8. doi: 10.1093/bioinformatics/btw354

Findlay J, Gould VC, North P, Bowker KE, Williams OM, MacGowan AP, Avison MB. Characterisation of cefotaxime-resistant urinary *Escherichia coli* from primary care in South-West England 2017-2018. bioRxiv 2019 doi: http://dx.doi.org/10.1101/701383

Goswami C, Fox S, Holden M, Connor M, Leanord A, Evans TJ. 2018. Genetic analysis of invasive *Escherichia coli* in Scotland reveals determinants of healthcare-associated versus community-acquired infections. Microb Genom 4:e000190. doi:10.1099/mgen.0.000190.

Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 2013 29(8):1072-5. doi: 10.1093/bioinformatics/btt086

Hanage WP. Two health or not two health? That is the question. MBio. 2019 9:e00550-19. doi: 10.1128/mBio.00550-19.

Harrison PW, Alako B, Amid C, Cerdeño-Tárraga A, Cleland I, Holt S, Hussein A, Jayathilaka S, Kay S, Keane T, Leinonen R, Liu X, Martínez-Villacorta J, Milano A, Pakseresht N, Rajan J, Reddy K, Richards E, Rosello M, Silvester N, Smirnov D, Toribio AL, Vijayaraja S, Cochrane G. The European Nucleotide Archive in 2018. Nucleic Acids Res. 2019 47(D1):D84-D88. doi: 10.1093/nar/gky1078

Johnson JR, Johnston B, Clabots C, Kuskowski MA, Castanheira M. 2010. *Escherichia coli* sequence type ST131 as the major cause of serious multidrug-resistant E. coli infections in the United States. Clin Infect Dis 51:286–294. doi:10.1086/653932.

Jolley KA, Bliss CM, Bennett JS, Bratcher HB, Brehony C, Colles FM, Wimalarathna H, Harrison OB, Sheppard SK, Cody AJ, Maiden MC. Ribosomal multilocus sequence typing: universal characterization of bacteria from domain to strain. Microbiology. 2012 158(Pt 4):1005-15. doi: 10.1099/mic.0.055459-0.

Kinderis M, Bezbradica M, Crane M. Bitcoin Currency Fluctuation. 2018. In *Proceedings of the 3rd International Conference on Complexity, Future Information Systems and Risk (COMPLEXIS 2018)*, pages 31-41.

Kallonen T, Brodrick HJ, Harris SR, Corander J, Brown NM, Martin V, Peacock SJ, Parkhill J. Systematic longitudinal survey of invasive *Escherichia coli* in England demonstrates a stable population structure only transiently disturbed by the emergence of ST131. Genome Res. 2017 doi: 10.1101/gr.216606.116

Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013 30(4):772-80. doi: 10.1093/molbev/mst010

Knudsen PK, Brandtzaeg P, Høiby EA, Bohlin J, Samuelsen Ø, Steinbakk M, Abrahamsen TG, Müller F, Gammelsrud KW. Impact of extensive antibiotic treatment on faecal carriage of antibiotic-resistant enterobacteria in children in a low resistance prevalence setting. PLoS One. 2017 12(11):e0187618. doi: 10.1371/journal.pone.0187618

Lanza VF, de Toro M, Garcillán-Barcia MP, Mora A, Blanco J, Coque TM, de la Cruz F. Plasmid flux in *Escherichia coli* ST131 sublineages, analyzed by plasmid constellation network (PLACNET), a new method for plasmid reconstruction from whole genome sequences. PLoS Genet. 2014 10(12):e1004766. doi: 10.1371/journal.pgen.1004766

Lees JA, Harris SR, Tonkin-Hill G, Gladstone RA, Lo SW, Weiser JN, Corander J, Bentley SD, Croucher NJ. Fast and flexible bacterial genomic epidemiology with PopPUNK. Genome Research 29:304-316 (2019). doi:10.1101/gr.241455.118

Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics. 2010 26(5):589-95. doi: 10.1093/bioinformatics/btp698

Ludden C, Raven KE, Jamrozy D, Gouliouris T, Blane B, Coll F, de Goffau M, Naydenova P, Horner C, Hernandez-Garcia J, Wood P, Hadjirin N, Radakovic M, Brown NM, Holmes M, Parkhill J, Peacock SJ. One health genomic surveillance of *Escherichia coli* demonstrates distinct lineages and mobile genetic elements in isolates from humans versus livestock. MBio. 2019 10(1):e02693-18. doi: 10.1128/mBio.02693-18

Ludden C, Decano A, Jamrozy D, Pickard D, Morris D, Parkhill J, Peacock SJ, Achtman M, Dougan G, Cormican M, Downing T. Genomic surveillance of *Escherichia coli* ST131 reveals the evolutionary history of epidemic antimicrobial resistant clones. (in prep).

Mahon BM, Brehony C, Cahill N, McGrath E, O'Connor L, Varley A, Cormican M, Ryan S, Hickey P, Keane S, Mulligan M, Ruane B, Jolley KA, Maiden MC, Brisse S, Morris D. Detection of OXA-48-like-producing *Enterobacterales* in Irish recreational water. Sci Total Environ. 2019 690:1-6. doi: 10.1016/j.scitotenv.2019.06.480

Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 2011 17(1):10-12. DOI: https://doi.org/10.14806/ej.17.1.200

Mathers AJ, Peirano G, Pitout JD. The role of epidemic resistance plasmids and international high-risk clones in the spread of multidrug-resistant *Enterobacteriaceae*. Clin Microbiol Rev. 2015 28(3):565-91. doi: 10.1128/CMR.00116-14

Matsumura Y, Pitout JD, Gomi R, Matsuda T, Noguchi T, Yamamoto M, Peirano G, DeVinney R, Bradford PA, Motyl MR, Tanaka M, Nagao M, Takakura S, Ichiyama S. Global *Escherichia coli* Sequence Type 131 clade with bla(CTX-M-27) gene. Emerg Infect Dis. 2016 22(11):1900-1907. doi: 10.3201/eid2211.160519

Matsumura Y, Pitout JDD, Peirano G, DeVinney R, Noguchi T, Yamamoto M, Gomi R, Matsuda T, Nakano S, Nagao M, Tanaka M, Ichiyama S. Rapid identification of different *Escherichia coli* Sequence Type 131 clades. Antimicrob Agents Chemother. 2017 61(8):e00179-17. doi: 10.1128/AAC.00179-17

McNally A, Oren Y, Kelly D, Pascoe B, Dunn S, Sreecharan T, Vehkala M, Valimaki N, Prentice MB, Ashour A, Avram O, Pupko T, Dobrindt U, Literak I, Guenther S, Schaufler K, Wieler LH, Zhiyong Z, Sheppard SK, McInerney JO, Corander J. 2016. Combined analysis of variation in core, accessory and regulatory genome regions provides a super-resolution view into the evolution of bacterial populations. PLoS Genet 12:e1006280.

McNally A, Kallonen T, Connor C, Abudahab K, Aanensen DM, Horner C, Peacock SJ, Parkhill J, Croucher NJ, Corander J. Diversification of Colonization Factors in a Multidrug-Resistant *Escherichia coli* Lineage Evolving under Negative Frequency-Dependent Selection. MBio. 2019 10(2). pii: e00644-19. doi: 10.1128/mBio.00644-19

Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R. The microbial pan-genome. Curr. Opin. Genet. Dev. 2005 15:589–594 doi: 10.1016/j.gde.2005.09.006

Nikolenko SI, Korobeynikov AI, Alekseyev MA. BayesHammer: Bayesian clustering for error correction in single-cell sequencing. BMC Genomics. 2013 14 Suppl 1:S7. doi: 10.1186/1471-2164-14-S1-S7

Ny S, Sandegren L, Salemi M, Giske CG. Genome and plasmid diversity of Extended-Spectrum β-Lactamase-producing *Escherichia coli* ST131 – tracking phylogenetic trajectories with Bayesian inference. Scientific Reports 2019 9:10291 doi: https://doi.org/10.1038/s41598-019-46580-3

Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, Phillippy AM. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol. 2016 17(1):132. doi: 10.1186/s13059-016-0997-x

Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MT, Fookes M, Falush D, Keane JA, Parkhill J. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics. 2015 31(22):3691-3. doi: 10.1093/bioinformatics/btv421

Park SC, Lee K, Kim YO, Won S, Chun J. Large-scale genomics reveals the genetic characteristics of seven species and importance of phylogenetic distance for estimating pan-genome size. Front Microbiol. 2019 10:834. doi: 10.3389/fmicb.2019.00834.

Partridge SR, Tsafnat G. Automated annotation of mobile antibiotic resistance in Gram-negative bacteria: the Multiple Antibiotic Resistance Annotator (MARA) and database. J Antimicrob Chemother. 2018 73(4):883-890. doi: 10.1093/jac/dkx513

Paul S, Linardopoulou EV, Billig M, Tchesnokova V, Price LB, Johnson JR, Chattopadhyay S, Sokurenko EV. Role of homologous recombination in adaptive diversification of extraintestinal *Escherichia coli*. J Bacteriol. 2013 195(2):231-42. doi: 10.1128/JB.01524-12

Peirano G, Schreckenberger PC, Pitout J. 2011. Characteristics of NDM-1-producing *Escherichia coli* isolates that belong to the successful and virulent clone ST131. Antimicrob Agents Chemother 55:2986–2988. doi:10.1128/AAC.01763-10.

Petty NK, Ben Zakour ZN, Stanton-Cook M, Skippington E, Totsika M, Forde BM, Phan MD, Gomes Moriel D, Peters KM, Davies M, Rogers BA, Dougan G, Rodriguez-Baño J, Pascual A, Pitout JD, Upton M, Paterson DL, Walsh TR, Schembri MA, Beatson SA. 2014. Global dissemination of a multidrug resistant *Escherichia coli* clone. Proc Natl Acad Sci U S A 111:5645–5649. doi:10.1073/pnas.1322678111

Pijnacker R, Dallman TJ, Tijsma ASL, Hawkins G, Larkin L, Kotila SM, Amore G, Amato E, Suzuki PM, Denayer S, Klamer S, Pászti J, McCormick J, Hartman H, Hughes GJ, Brandal LCT, Brown D, Mossong J, Jernberg C, Müller L, Palm D, Severi E, Gołębiowska J, Hunjak B, Owczarek S, Le Hello S, Garvey P, Mooijman K, Friesema IHM, van der Weijden C, van der Voort M, Rizzi V, Franz E; International Outbreak Investigation Team. An international outbreak of *Salmonella* enterica serotype *Enteritidis* linked to eggs from Poland: a microbiological and epidemiological study. Lancet Infect Dis. 2019 pii: S1473-3099(19)30047-7. doi: 10.1016/S1473-3099(19)30047-7

Poolman JT, Wacker M. Extraintestinal Pathogenic *Escherichia coli*, a Common Human Pathogen: Challenges for Vaccine Development and Progress in the Field. J Infect Dis. 2016 213(1):6-13. doi: 10.1093/infdis/jiv429

Price LB, Johnson JR, Aziz M, Clabots C, Johnston B, Tchesnokova V, Nordstrom L, Billig M, Chattopadhyay S, Stegger M, Andersen PS, Pearson T, Riddell K, Rogers P, Scholes D, Kahl B, Keim P, Sokurenko EV. 2013. The epidemic of extended-spectrum-β-lactamase-producing *Escherichia coli* ST131 is driven by a single highly pathogenic subclone, H30-Rx. mBio 4:e00377-13. doi: 10.1128/mBio.00377-13

Randall LP, Lodge MP, Elviss NC, Lemma FL, Hopkins KL, Teale CJ, Woodford N. Evaluation of meat, fruit and vegetables from retail stores in five United Kingdom regions as sources of extended-spectrum beta-lactamase (ESBL)-producing and carbapenem-resistant *Escherichia coli*. Int J Food Microbiol. 2017 241:283-290. doi: 10.1016/j.ijfoodmicro.2016.10.036

Raven KE, Blane B, Leek D, Churcher C, Kokko-Gonzales P, Pugazhendhi D, Fraser L, Betley J, Parkhill J, Peacock SJ. Methodology for Whole-Genome Sequencing of Methicillin-Resistant *Staphylococcus aureus* Isolates in a Routine Hospital Microbiology Laboratory. J Clin Microbiol. 2019 24;57(6). pii: e00180-19. doi: 10.1128/JCM.00180-19

Revez J, Espinosa L, Albiger B, Leitmeyer KC, Struelens MJ; ECDC National Microbiology Focal Points and Experts Group. Survey on the Use of Whole-Genome Sequencing for Infectious Diseases Surveillance: Rapid Expansion of European National Capacities, 2015-2016. Front Public Health. 2017 5:347. doi: 10.3389/fpubh.2017.00347

Roer L, Overballe-Petersen S, Hansen F, Johannesen TB, Stegger M, Bortolaia V, Leekitcharoenphon P, Korsgaard HB, Seyfarth AM, Mossong J, Wattiau P, Boland C, Hansen DS, Hasman H, Hammerum AM, Hendriksen RS. ST131 *fimH*22 *Escherichia coli* isolate with a blaCMY-2/IncI1/ST12 plasmid obtained from a patient with bloodstream infection: highly similar to *E. coli* isolates of broiler origin. J Antimicrob Chemother. 2019 74(3):557-560. doi: 10.1093/jac/dky484

Salipante SJ, Roach DJ, Kitzman JO, Snyder MW, Stackhouse B, Butler-Wu SM, Lee C, Cookson BT, Shendure J. Large-scale genomic sequencing of extraintestinal pathogenic Escherichia coli strains. Genome Res. 2015 25(1):119-28. doi: 10.1101/gr.180190.114

Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014 30(14):2068-9. doi: 10.1093/bioinformatics/btu153

Sheppard AE, Stoesser N, German-Mesner I, Vegesana K, Sarah Walker A, Crook DW, Mathers AJ. TETyper: a bioinformatic pipeline for classifying variation and genetic contexts of transposable elements from short-read whole-genome sequencing data. Microb Genom. 2018 4(12). doi: 10.1099/mgen.0.000232

Sintchenko V, Holmes EC. The role of pathogen genomics in assessing disease transmission. BMJ. 2015 350:h1314. doi: 10.1136/bmj.h1314

Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014 30(9):1312-3. doi: 10.1093/bioinformatics/btu033

Stoesser N, Sheppard AE, Pankhurst L, de Maio N, Moore CE, Sebra R, Turner P, Anson LW, Kasarskis A, Batty EM, Kos V, Wilson DJ, Phetsouvanh R, Wyllie D, Sokurenko E, Manges AR, Johnson TJ, Price LB, Peto TEA, Johnson JR, Didelot X, Walker AS, Crook DW, Modernizing Medical Microbiology Informatics Group (MMMIG). 2016. Evolutionary history of the global emergence of the *Escherichia coli* epidemic clone ST131. mBio 7:e02162. doi:10.1128/mBio.02162-15

Sullivan MJ, Petty NK, Beatson SA. Easyfig: a genome comparison visualizer. Bioinformatics. 2011 27(7):1009-10. doi: 10.1093/bioinformatics/btr039

Tettelin H, Riley D, Cattuto C, Medini D. Comparative genomics: the bacterial pan-genome. Curr Opin Microbiol. 2008 11(5):472-7

Tonkin-Hill G, Lees JA, Bentley SD, Frost SDW, Corander J. Fast hierarchical Bayesian analysis of population structure. Nucleic Acids Res. 2019 47(11):5539-5549. doi: 10.1093/nar/gkz361

Totsika M, Beatson SA, Sarkar S, Phan MD, Petty NK, Bachmann N, Szubert M, Sidjabat HE, Paterson DL, Upton M, Schembri MA. 2011. Insights into a multidrug resistant *Escherichia coli* pathogen of the globally disseminated ST131 lineage: genome analysis and virulence mechanisms. PLoS One 6:e26578. doi:10.1371/journal.pone.0026578.

Van der Bij AK, Peirano G, Pitondo-Silva A, Pitout JD. 2012. The presence of genes encoding for different virulence factors in clonally related Escherichia coli that produce CTX-Ms. Diagn Microbiol Infect Dis 72:297–302. doi:10.1016/j.diagmicrobio.2011.12.011. Calhau V, Ribeiro G, Mendonça N, Da Silva GJ. 2013. Prevalent combination of virulence and plasmidic-encoded resistance in ST131 *Escherichia coli* strains. Virulence 4:726–729. doi:10.4161/viru.26552.

Vaser R, Sović I, Nagarajan N, Šikić M. Fast and accurate *de novo* genome assembly from long uncorrected reads. Genome Res. 2017 27(5):737-746. doi: 10.1101/gr.214270.116

Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, Earl AM. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One. 2014 9(11):e112963. doi: 10.1371/journal.pone.0112963

Welch RA, Burland V, Plunkett G 3rd, Redford P, Roesch P, Rasko D, Buckles EL, Liou SR, Boutin A, Hackett J, Stroud D, Mayhew GF, Rose DJ, Zhou S, Schwartz DC, Perna NT, Mobley HL, Donnenberg MS, Blattner FR. Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. Proc Natl Acad Sci U S A. 2002 99(26):17020-4 doi:10.1073/pnas.252529799

Wick RR, Judd LM, Gorrie CL, Holt KE. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. PLoS Comput Biol. 2017 13(6):e1005595. doi: 10.1371/journal.pcbi.1005595

Wirth T, Falush D, Lan R, Colles F, Mensa P, Wieler LH, Karch H, Reeves PR, Maiden MC, Ochman H, Achtman M. Sex and virulence in *Escherichia coli*: an evolutionary perspective. Mol Microbiol. 2006 60(5):1136-51.

Yu, Guangchuang, David K Smith, Huachen Zhu, Yi Guan, and Tommy Tsan-Yuk Lam. 2017. Ggtree: An R Package for Visualization and Annotation of Phylogenetic Trees with Their Covariates and Other Associated Data. Methods Ecol. Evol. 8(1): 8-36. doi:10.1111/2041-210X.12628.

Zhou Z, Alikhan NF, Mohamed K, the Agama study group, Achtman M. The user’s guide to comparative genomics with EnteroBase. Three case studies: micro-clades within *Salmonella enterica serovar Agama*, ancient and modern populations of *Yersinia pestis*, and core genomic diversity of all *Escherichia*. 2019. Biorxiv doi: http://dx.doi.org/10.1101/613554