A Population Phylogenomic Analysis of the Origin and Spread of *Escherichia coli* Sequence Type 131 (ST131)

Arun Gonzales Decano
BS Biology, MSc Medical Microbiology

A thesis presented to Dublin City University for the Degree of
Doctor of Philosophy

PhD Adviser: Dr. Tim Downing
School of Biotechnology
Dublin City University

September 2019

Declaration

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of PhD is entirely my own work, and that I have exercised reasonable care to ensure that the work is original, and does not, to the best of my knowledge, breach any law of copyright, and has not been taken from the work of others and to the extent that such work has been cited and acknowledged within the text of my work.

Signed: Arun G. Decano Candidate ID No. 1521-2072 Date: 13/09/19

"Education breeds confidence. Confidence breeds hope. Hope breeds peace."

– Confucius

Acknowledgements

The completion of this work will not be possible without the unparalleled support of the following people in my life:

My family, who always has my back and has supported my decision to go abroad and pursue my studies in a foreign country despite their constant concern of my safety.

My best friend, Kayla, who has sacrificed so much just to be by my side during my most difficult times.

My former housemates in Dublin especially Marc, Andre and Aurora, who made my homesickness bearable while I'm away from my family in the Philippines.

My former labmates Ray, Simone and Ann, whose dedication in their individual fields of study, influenced me to be passionate about my own.

All the undergrad students and interns in the lab, who patiently listened while I endlessly talk about ST131 and the advantages of coding.

My fellow DCU postgrads and friends especially Flavio, Burcu, Thayse and Catherine, whom I have shared laughter, prosecco and sentiments of failed experiments with and spent long hours of working in the lab with. I thank them for helping me keep my sanity while navigating through the roller coaster of events during my PhD study.

Most of all, I am extremely grateful to Dr. Tim Downing for his leadership, guidance and infinite support. I cannot imagine having a better PhD mentor and friend in this journey. Tim, thank you for everything.

This PhD thesis is dedicated to my late brother, Clark, who has inspired me to rise above the challenges I face every day.

Table of Contents

Thesis Abstract

The incidence of infections caused by extraintestinal Escherichia coli (ExPEC) is rising globally due to their increasing resistance to standard antibiotics. This results in the use of broader-spectrum drugs, prolonged patient ill-health and more nosocomial infections. E. coli sequence type 131 (ST131) is the predominant ExPEC clone worldwide. The antimicrobial resistance (AMR) gene repertoire of ST131 is evolving rapidly due to the widespread use of β-lactam (bla) antibiotics. Here, we performed a genomic investigation of an ST131 outbreak in a long-term care facility (LTCF) to describe transmission, within-host clonal diversity, genetic diversity of antibiotic resistance and the evolution of ST131 in the LTCF over a seven-year period. We analyzed the population structure and inferred the genealogical history of the LTCF isolates in the context of local hospital and global collections of ST131 to elucidate the epidemiology of ST131. We confirmed our initial hypotheses by reconstructing the evolutionary history of a much larger population consisting of >4000 global ST131 genomes This provided a deeper resolution of their evolutionary trajectories and the adaptive mechanisms of AMR driven by their ESBL genes, particularly cefotaximase (blaCTX). We further investigated the intersection of the AMR genes (AMRGs) found in ST131 with that of the human microbiome to understand the extent of their loss, gain and spread across different bacterial species. Across all strains, a large number of ST131's AMRGs were found in a total of 794 genes in the human microbiome. Various gene families were represented, including transporters, transcription factors, β-lactamases and cell wall biosynthesis enzymes. To establish the main culprit for the dynamic nature of the blaCTX-M genes, we performed long read sequencing using a GridION X5 instrument. Analysis of long read-only assemblies revealed a clear and robust result on the genetic flanking context of blaCTX-M genes in both plasmid and chromosomes. Overall, our findings underpin the tremendous potential power for improving our current treatment of bacterial infections using high-throughput analysis of whole genome sequence data.

Abbreviations

AMR	Antimicrobial resistance
AMRG	Antimicrobial resistance gene
Bla	β-lactamase
***Bla*$_{\text{CTX-M}}$**	Cefotaximase, a type of β-lactamase
BLAST	Basic Local Alignment Search Tool
Bp	Base pair
CDS	Coding sequence
ENA	European Nucleotide Archive
ExPEC	Extraintestinal pathogenic *Escherichia coli*
fimH	fimbrial protein type H
FQ	Fluoroquinolone
FQ-R	Fluoroquinolone resistant
FQ-S	Fluoroquinolone susceptible
Gbp	Gigabase pairs
GC content	Guanine-Cytosine content
HPD	Highest Posterior Density
ICHEC	Irish Centre for High-End Computing
Inc	Incompatibility group of plasmids
IncF	Incompatibility plasmid type F
IS	Insertion sequence
Kb/Kbp	Kilobase pair
LR	Long read
LTCF	Long-term care facility
MGE	Mobile genetic element
ML	Maximum likelihood
MLST	Multi-locus sequence type
N50/NG50	The length (for a set of sequences with varying sizes), such that 50% of the genome is contained in that length
NFDS	Negative frequency-dependent selection
NGS	Next Generation Sequencing (technology)
ONT	Oxford Nanopore Technology
ORF	Open reading frame
PFGE	Pulse-filed gel electrphoresis
QC	Quality control
QRDR	Quinolone resistance-determining region
r/m	Recombination to mutation ratio
RGD	Relative Genetic Diversity
rMLST	Ribosomal multi-locus sequence type
rST	Ribosomal sequence type
SNP	Single nucleotide polymorphism
SR	Short read
ST131	*E. coli* sequence type 131
TE	Transposable element
TMRCA	The most recent common ancestor
Tn	Transposon
TU	Transposable unit
UPEC	Uropathogenic *E. coli*

Thesis Overview

In this thesis, we generated a series of phylogenomic workflows to investigate the evolution, population structure and antimicrobial resistance in ST131 populations. The sample datasets in the succeeding chapters were interrelated but were used independently of each other.

A thorough background study on ST131 genomics was first conducted and recorded in the first chapter of the thesis. In Chapter 2, N=100 Irish/UK ST131 isolates were examined as an extension of the pilot study of the project. This collection served as the testing dataset for optimising the methods we used to analyse ST131 populations discussed in succeeding chapters. Using these 100 strains, we quantified recombination events in ST131 using three most commonly used platforms for examining HGT events in bacteria.

A larger N=794 ST131 collection was then examined in Chapter 3. In this part of the study, we investigated an outbreak of extended spectrum β-lactamase (ESBL) producing *E. coli* ST131 in a long-term care facility (LTCF) in Ireland (n=90) and combined this data with Irish (n=48) and global (n=690) ST131 genomes to reconstruct the evolutionary history and further understand changes in population structure and recombination patterns over time. We uncovered an evidence for an extensive rearrangement of ESBL genes in plasmids and chromosome which contributed to the spread of diverse clones worldwide and a local outbreak in a LTCF in Ireland which spanned 4 years.

As short DNA reads do not fully resolve the architecture of repetitive elements i.e. on plasmids, we performed long read sequencing of six *E. coli* ST131 isolated from six patients in Chapter 4. Majority of our long read assemblies revealed entire chromosomes and plasmids as single contigs in contrast to highly fragmented ones from short reads. Our results here highlighted diverse core and accessory genomes with *bla*CTX-M-15, *bla*CTX-M-14 and *bla*CTX-M-27 genes and showed that AMR genes exist in multiple different chromosomal and plasmid contexts even between closely-related isolates within a clonal group such as *E. coli* ST131.

er the years, ST131 has radiated into genetically distinct subclades. In Chapter 5, we extracted all available high-quality global ST131 Illumina HiSeq read libraries, automated quality-control, genome de novo assembly and ESBL gene screening to explore the largest ST131 sample collection examined so far. We used published reference genomes, Nanopore and PacBio assemblies as well as k-mer-based methods to contextualise the core and accessory genome diversity to identify the key features in each main clade and subclade. Our findings here provided a deeper and more refined resolution of the hypervariable accessory genome, including plasmids and key ESBL genes in ST131.

Finally, in chapter 6, the extent of AMR gene transfer between the human gut microbiome isolates and ST131 was explored. Additionally, a dynamic gene repertoire associated with the mobile resistome in the pathogen *E. coli* ST131 was examined. This assesses horizontal DNA transfer between E. coli ST131 and gut E. coli regarding plasmids, transposons and other mobile genetic elements (MGEs).

Chapter 1: Introduction to *E. coli* ST131 Genomics

1.1 Background on the Origin, Evolution and Structure of ST131 Population

Antimicrobial resistance (AMR) poses an increasing challenge for treating infectious diseases. Recent advances in sequencing technologies and molecular genetics equip us with tools to assess the origins of infections and their transmission. The most common bacterial infection is brought about by *Escherichia coli* and its subgroup sequence type 131 (ST131), which possesses extensive resistance properties against antibiotics, is responsible for cases of global pandemic outbreaks (Nicolas-Chanoine et al. 2013). It also retains a broad arsenal of genes promoting antimicrobial tolerance. Preliminary data shows that ST131 is commonly isolated in nursing homes and hospitals, which may serve as reservoirs of drug-resistant bacteria that spread into the community (Price et al. 2013, Petty et al. 2014, Johnson, et al. 2013).

Escherichia coli sequence type 131 (*E. coli* ST131 or ST131) is a pandemic multidrug-resistant (MDR) *E. coli* sublineage. ST131 has a variety of virulence-associated genetic elements and has a broad capacity to cause urinary tract and bloodstream infections in both hospital and community settings (Olesen et al. 2014; Johnson, et al. 2013; Ludden et al. 2015; Zhong et al. 2015). *E. coli* isolates are generally classified into four phylogenetic groups: A, B1, B2 and D (Selander et al, 1986); ST131 corresponds to subgroup 1 of phylogenetic group (virulent) B2 (Figure 1.1).

Previous genomic studies initially elucidated the complex clonal structure of small samples of ST131 (Price et al. 2013; Petty et al. 2014) and identified subclades with specific marker allele for the type 1 fimbriae *fimH* (Dr-binding fimbrial adhesin gene: H subclone assignments): *H41* in clade A, H22 in clade B, and *H30* in clade C (Johnson et al, 2013). *H30* is the most prevalent, followed by H22 and then H41 (Johnson, et al. 2013). Although these three are the most frequent types of *fimH* among isolated ST131 (Adams-Sapper, et al. 2013; Stoesser et al. 2016; Ben Zakour et al. 2016), other

types such as *fimH35*, *H27*, *H31* and *H94* were also recently observed in discrete B subclades, B1 to B5 (Nicolas-Chanoine et al. 2013) (Stoesser et al. 2016; Ben Zakour et al. 2016). This classification provides more opportunities to further investigate the unprecedented expansion of ST131 clone.

One review (Pitout, JDD and Devinney, R., 2017) summarized the step-wise evolution of ST131 (Figure 1.2), which involves the insertion of prophages in the B subclade giving rise to B0. Recombination at *parC1a* and *fimH30* and insertion of genomic islands, prophages facilitated by IncF2 plasmid gave rise to the C0. The start of fluoroquinolone (FQ) treatment gave rise the mutant genotypes *parC1aAB* and *gyrA1AB* associated with fluoroquinolone resistance (FQ-R) and eventual diversification of C into two highly virulent subclusters, C1 and C2. C2 has a gene encoding an extended spectrum beta-lactamase (ESBL) called cefotaximase-15 ($bla_{\text{CTX-M-15}}$). ST131 is comprised of strains with O25b (Rogers et al. 2011; Woodford, 2011; Olesen et al. 2013) or O16 serotype, which have greater prevalence of another similar ESBL $bla_{\text{CTX-M-14}}$, which are marginally less FQ-R but more commonly resistant to trimethoprim-sulfamethoxazole (Johnson et al. 2014; Matsumura et al. 2012), (Matsumura et al. 2013; Nicolas-Chanoine et al 2008; Petty et al 2014; Totsika et al 2011, 2012).

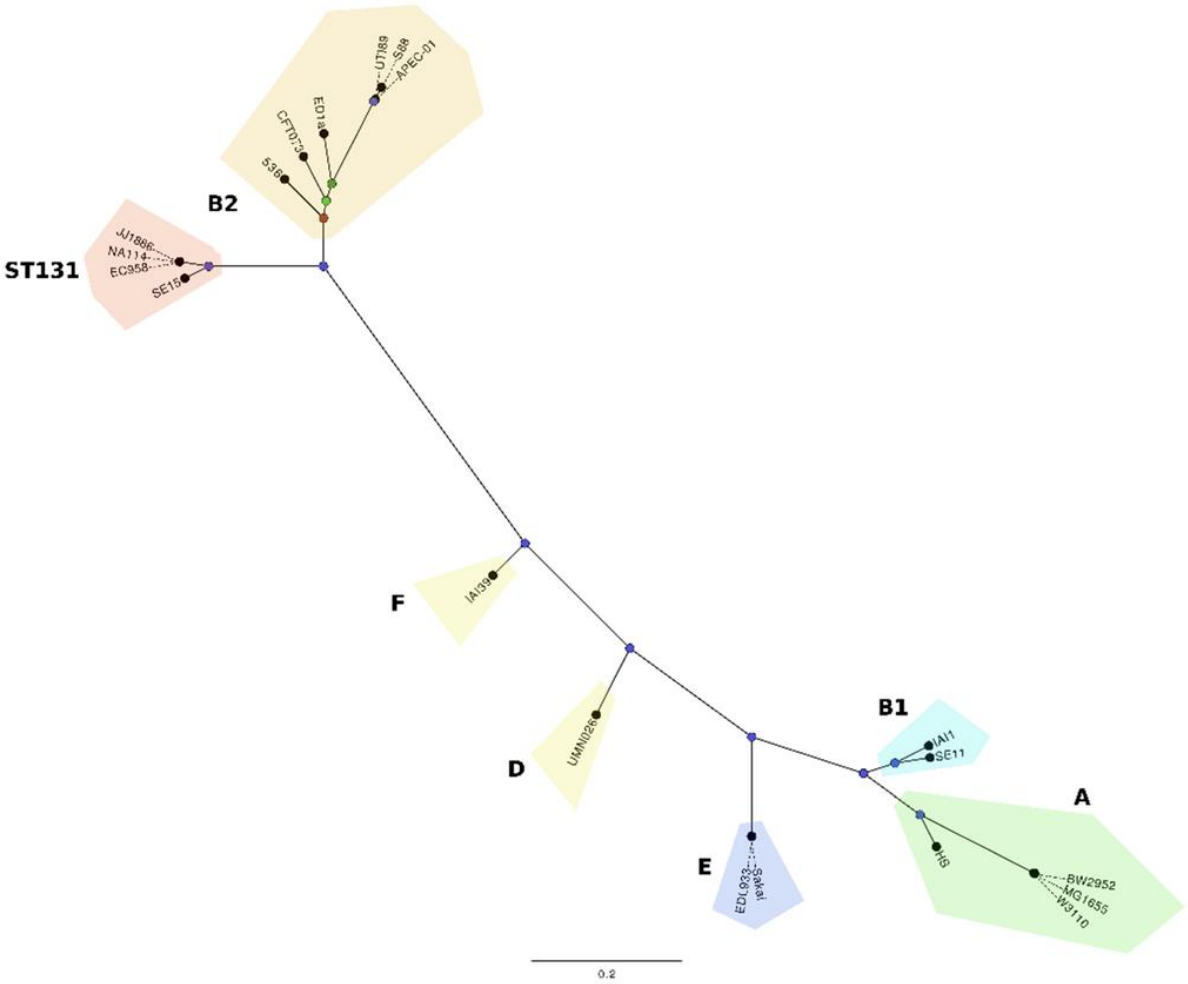


Figure 1.1. Phylogenetic tree showing the four general phylogroups of *E. coli*. ST131 strains (highlighted in pink) cluster together with other virulent strain in clade B2. The figure was adapted from Forde et al. 2014 and Schembri et al. 2015.

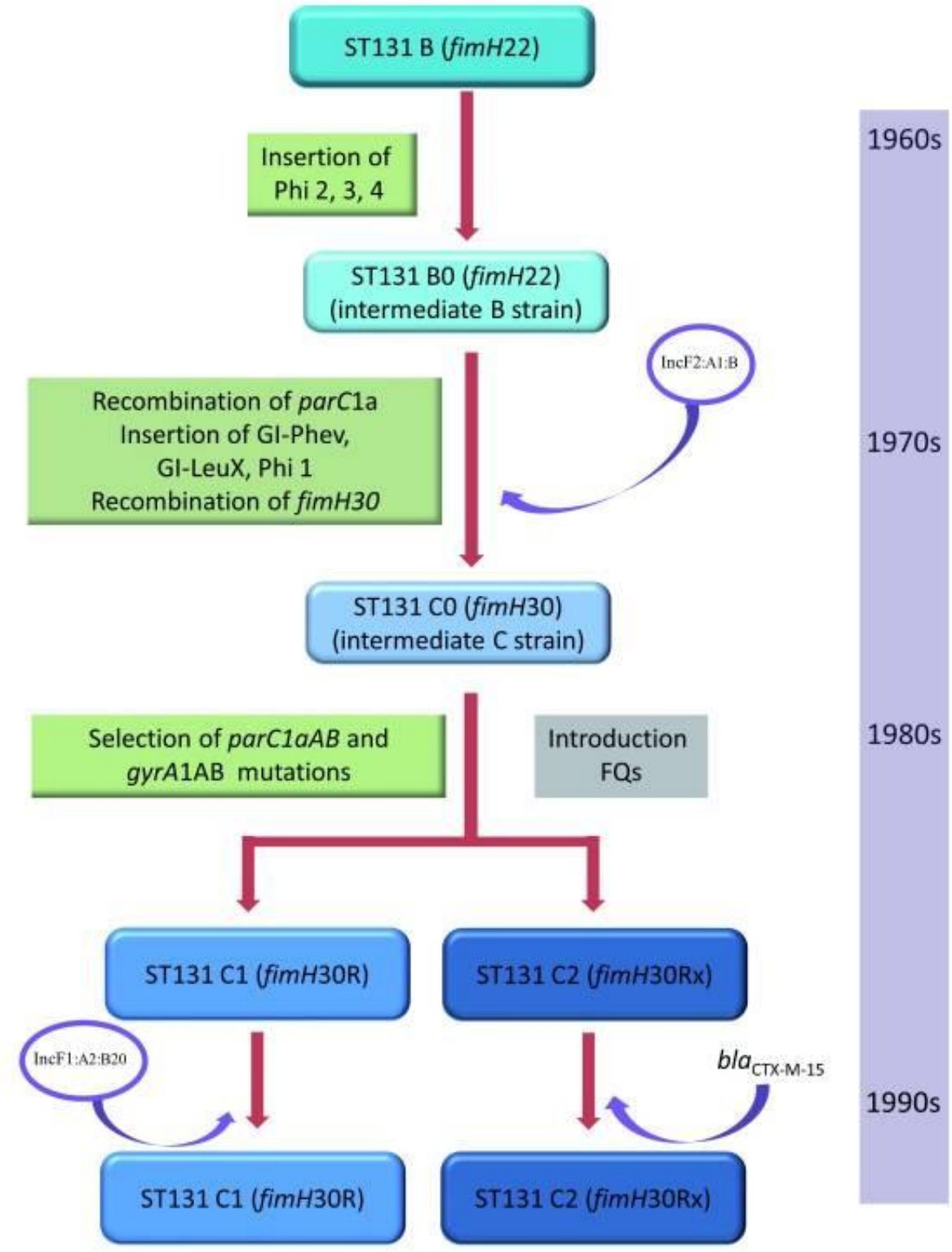


Figure 1.2. Stepwise evolution of *Escherichia coli* ST131 clades B, B0, C0, and C. Diagram is adapted from Pitout, JDD and Devinney, R., 2017. FQ, fluoroquinolones; GI, genomic islands; Inc, incompatibility; Phi, prophages; ST, sequence types.

Loss and gain of resistance-conferring genes via horizontal gene transfer (HGT) particularly that of a β-lactamase gene called *bla*$_{CTX\text{-}M\text{-}15}$, has played a huge role in shaping ST131 (Petty et al. 2014). β-lactamases are bacterial enzymes that can degrade antibiotics with β-lactam compounds aimed to inhibit cell wall synthesis. The widespread use of β-lactams caused selection pressures in bacterial populations producing ESBLs that hydrolyze penicillins, early cephalosporins, oxyimino-thiazolyl cephalosporins (including third and fourth generation cephalosporins) and monobactams, but not cephamycins or carbapenems. Seven different SNPs at β-lactamase genes are believed to have given rise to ESBLs. Early ESBLs were only able to hydrolyze ceftazidime but have eventually evolved to confer broad resistance to cefotaxime (a compound found in cephalosporins). An *E. coli* isolate obtained from a dog in Japan (used in a pharmacokinetic study of cephalosphorin) in 1986 was the first report of insensitivity to cefotaxime, though it was still susceptible to ceftazidime. Three years later in 1989, the first clinical case with cefotaximase-resistance was reported in Munich, Germany from an *E. coli* culture from an ear discharge of a four-month old child. The enzyme in the isolate was hence given the name *bla*$_{CTX\text{-}M\text{-}1}$ (CTX for cefotaximase and M means Munich). Since then, successive cases of *bla*$_{CTX\text{-}M\text{-}1}$ detection in bacteria have been reported globally.

Bla$_{CTX\text{-}M\text{-}15}$ is most common ESBL gene among ST131. A study of *Enterobacteriaceae* in Poland in 1998 first revealed its presence in *E. coli* samples (Baraniak, et al. 2002). Indeed, *bla*$_{CTX\text{-}M\text{-}15}$ is the most common type of ESBL-producing genes in *E. coli* (Pitout and Laupland, 2008). These genes are embedded in transposon-like structures often contained in plasmids that carry additional armoury of resistance-associated elements (Canton and Coque, 2006) and when acquired, cause insusceptibility to other β-lactams, FQ, aminoglycosides, and trimethoprim-sulfamethoxazole (Johnson, et al. 2014).

Greater resolution genomic analyses identified virulence factor acquisition events and revealed that the clonal expansion associated with drug resistance in ST131 and was estimated to have emerged in North America around 25 years ago, coinciding with the first use of FQ in 1986 (Stoesser et al. 2016; Ben Zakour et al. 2016). Strains belonging to Clade C are characterized by their high FQ-R due to double mutations

(Stoesser et al. 2016) at the "quinolone resistance-determining region" (QRDR) of the chromosomal genes that encode the FQ targets DNA gyrase and topoisomerase IV: *gyrA* and *parC*, respectively (Hooper, 2001; Ruiz et al. 2003; Johnson, et al. 2014). The presence or absence of $bla_{\text{CTX-M-15}}$, defined further diversification of clade C into C1 and C2, dated to 1987: some C1 strains contain $bla_{\text{CTX-M-3/14/27}}$ while majority in C2 (H30-Rx) had the $bla_{\text{CTX-M-15}}$ gene (Ben Zakour et al. 2016) and display reduced susceptibility to third-generation cephalosporins limiting prophylactic options for this globally-disseminated clone.

1.1.1. Reference genomes used for analysing *E. coli* ST131 populations

The availability of reliable and high-quality reference genomes for *E. coli* ST131 provides a better understanding of the characteristics of this multidrug-resistant pathogen. I used three well-characterized genomes to analyse *E. coli* ST131 populations in the succeeding chapters, in addition to NA114 in Chapter 2: the genome of the commensal strain, SE15 as the negative control of our comparative analysis, the highly-pathogenic EC958 as the positive control and the recently assembled and characterized high-quality genome of NCTC13441 as the main reference sequence (Table 1.1).

- SE15

The *E. coli* SE15 genome is a 4,717,338-bp circular chromosome with 4,338 predicted protein-coding genes and a 122-kb plasmid (pSE15) that encodes 150 protein-coding genes (Toh et al. 2010). It belongs to the phylogenetic group B2 although it lacks many virulence-associated genes such as α-hemolysin and cytotoxic necrotizing factor -- known toxins encoded by pathogenicity islands of uropathogenic *E. coli* strains. The complete genome of SE15 was determined using cappilliary (Sanger) libraries and 454 pyrosequencing and assembles with Newbler assembler software (Margulies et al. 2005) and Phred-Phrap-Consed program (Gordon and Green 2013).

- EC958

In contrast to SE15, EC958 strain is fluoroquinolone-resistant (FQ-R), encodes *fimH30* and $bla_{\text{CTX-M-15}}$. The strain was isolated in March 2005 from a urine sample collected from an eight-year old girl presenting in the community in the United Kingdom (Totsika et al. 2011). The genome of EC958 is one of the most well-defined and well-studied *E. coli* ST131 strains. EC958 consists of several virulence-associated genes that encode adhesins, autotransporter proteins and siderophore receptors. EC958 was also shown to cause acute and chronic UTI (Totsika et al. 2013) and impairment of uterine contractility in mice (Floyd et al. 2012). Further, the serum resistome of *E. coli* EC958 has been extensively reviewed in (Phan et al. 2013).

Reference genome	**Pathogenicity**	**Sequencing platform**	**Assembly type**	**Genome size (bp)**	**Number of chromosomal protein-coding genes**
SE15	Commensal	Sanger (capilliary) and 454 Pyrosequencing	Newbler/Phred -Phrap-Consed	4,717,338	4,338
EC958	Uropathogenic	PacBio RS I	PacBio's SMRT Portal v2.0.0	5,109,767	4,982
NA114	Uropathogenic	Illumina Genome Analyzer (GA2x)	Velvet	4,935,666	4,875
NCTC13441	Uropathogenic	PacBio RS	PacBio	5,174,631	4,983

Table 1.1. Reference genomes used in ST131 genomic studies.

- NA114

NA114 is another uropathogenic ST131 isolated from western Indian city of Pune (Jadhav et al. 2011), which genome sequence was identified using Illumina Genome Analyzer (GA2x v.1.6). Its chromosome is 4,935,666-bp long with a coding percentage of 88.4% with 4,875 protein coding sequences and has a GC content of 51.16% (Avasthi et al. 2011). NA114 was also found to have a 3.5 Kb plasmid and have multiple virulence–associated genes including *sfa*, *aer*, *cnf*, and an intact polyketide synthase (*pks*) island (Johnson et al. 2008).

- NCTC13441

The genome of NCTC13441 is the most recent high-quality assembled *E. coli* ST131 reference genome. NCTC13441 is a uropathogenic *E. coli* ST131 strain that has an ESBL *bla*CTX-M-15 gene and was collected from a clinical isolate in the UK (Public Health England). It belongs to the group of strains with serotype O25:H5 and contains the sequenced plasmid pEK499 (Woodford et al. 2009). NCTC13441 was previously used to show that the frequency of horizontal gene transfer via conjugation was ten times higher and more stable in *E. coli* ST131 $bla_{\text{CTX-M-15}}$-producing than *K. pneumoniae* with New Delhi Metallo-β-lactamase-1 (NDM-1) (Warnes et al. 2012).

1.1.2 Plasmids common to *E. coli* ST131

Plasmids are circular self-replicating double-stranded DNA molecules that compose the majority of the bacterial accessory genome (Juhas et al. 2009; Hinnebusch and Tilly, 1993; Frost et al. 2005). Plasmids are classified as conjugative, non-conjugative, and mobilisable plasmids. Plasmids mediate antibiotic resistance gene circulation in bacteria. Cell-to-cell transfer of plasmids occurs by conjugation, where a plasmid from a donor cell is transferred to a recipient cell through a pilus. Pilus-encoding genes that are expressed in the plasmid of the donor cell. Another important plasmid protein is the relaxase that nicks one strand of the double-stranded plasmid; the nicked one strand is transported to the recipient cell via the pilus and replicated into double-stranded plasmid. Upon acquisition of the new plasmid, the recipient cell then becomes conjugative as well. Although mobilisable plasmids also encode the relaxase, they are incapable of independent conjugation and rely on other plasmids' machinery for transfer. Non-conjugative plasmids do not encode pilus- or relaxase genes and are unable to use other plasmids mechabinsms for transport.

Plasmids are also categorised according to their replicon type. Plasmids encoding the same or similar origins of replication (*ori*) are rarely found in the same cell as they compete for the replication machinery. These competing plasmids are grouped in the same incompatibility (Inc) group (Shintani et al., 2015). The plasmids from the IncF

group were found to carry and mobilise AMR genes in *E. coli* ST131 (Johnson et al. 2016), particularly IncF, IncI1, IncN and IncA/C groups (Carattoli, 2009; Nicolas-Chanoine et al. 2008; Nicolas-Chanoine et al. 2014). These plasmid types associated with of β-lactamase genes such as *bla*$_{\text{CTX-M}}$, *bla*$_{\text{OXA-1}}$ and *bla*$_{\text{TEM-1}}$.

1.2 Types of sequence libraries used in this study

DNA sequencing is the process of identifying the exact order of nucleotide bases (adenine (A), thymine (T), cytosine (C), guanine (G) in given segment(s) of DNA using different molecular methods. Whole genome sequencing determines the order of all of the DNA in an organism's cell. In bacteria, it entails the identification of all the nucleotide bases in both the chromosome and plasmid(s). Whole genomic sequence data is the raw material used in the interdisciplinary field of genomics, which largely involves analysis of whole genome sequence data using high throughput bioinformatics algorithms to decipher gene functions and analyse the structure of the entire set of protein-coding regions in an organism.

Sanger sequencing (aka the chain termination method), was the pioneer of all sequencing technologies. This strategy developed by Fred Sanger et al. in 1977 was used to complete the Human Genome Project. The process involves generation of several copies of the target DNA and making multiple fragments of various lengths. To mark the end of each fragment, Sanger sequencing uses fluorescent nucleotides as chain terminators (Sanger and Coulson 1975; Sanger, Nicklen and Coulson 1977). This traditional DNA sequencing method, however, failed to keep up with growing demand for deeper genome analyses and the increasing complexity of questions asked by researchers. This gap led to the development of second (more popularly known as the next-generation sequencing or NGS) and third generation sequencing technologies. Fundamentally, these two modern sequencing techniques revolutionized genome-sequencing. Among others, they allowed rapid and deeper sequencing of whole genomes and gave rise to the field of metagenomics, which investigates diverse microbial communities in humans, animals or the environment (https://www.illumina.com/science/technology/next-generation-sequencing.html). Below are the types of DNA sequence reads we employed in the analyses of ST131.

1.2.1 Short reads

The most commonly used DNA sequence library type is the short paired-end (PE) read library commercialized by Illumina. PE reads are the result of sequencing from both ends of a DNA fragment; the aligned forward and reverse reads are also called read pairs. Depending on the machine type/model used, PE reads are about 50-700 nucleotides long and are cheaper than Sanger-sequenced libraries. A more accurate read alignment and identification of single-nucleotide polymorphisms/variations (SNPs/SNVs) are more easily achieved with the analysis of differential read-pair spacing because PE sequencing allows the removal of artefacts from routine PCR library preparation (Head et al. 2014).

1.2.2 Long reads

While short PE reads can cover majority of the genome with high accuracy, they lack the contextual information in resolving complex and repetitive regions. In achieving optimal benefits from whole genome sequencing, one could use a strategy that provides longer read lengths. Two widely used strategies developed by Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT), can generate reads with lengths of several thousand base pairs (Jayakumar and Sakakibara 2017) and can fill the fill the gap that the PE reads may not resolve, though with less nucleotide-specific accuracy: Pacbio reads were estimated to exhibit about 20% of error rates (Travers et al. 2010, Thompson et al. 2011) while ONT libraries were recorded to have about 35% (Goodwin et al. 2015).

Longer reads do not necessarily mean more accurate genome assembly and mapping. It is essential to note that the quality of high molecular weight DNA as a starting material is the primary key to achieve high quality results more than the type of sequencer used. This includes prevention of DNA damage and fragmentation during library preparation (Pollard et al. 2018).

- PacBio sequences

The first type of long-read sequencing technology is the single molecule real time (SMRT) sequencing developed by PacBio. This technology typically produces sequences that are about 100 Kb. As implemented in their Sequel and RS- II platforms, this SMRT system utilizes a scheme of massively parallel polymerases; each polymerase is bound to a single molecule of target DNA. This DNA molecule is circularized using a pair of hairpin sequencing adaptors called the SMRTbell. A CCD camera, powered by a zero-mode waveguide (Levene et al. 2003, Eid et al. 2009) is used to detect the signal from the fluorescent system of tagged bases and polymerase on the template DNA (Pollard et al. 2018).

- ONT sequencing

Oxford Nanopore Technologies (ONT) is a long read technology typically called Nanopore sequencing (Eisenstein 2012), and is implemented in their Flongle, MinION (Mikheyev and Tin 2014), GridION X5 and PromethION devices. In Nanopore sequencing, a nanopore is positioned in an electrically resistant polymer membrane (Deamer, Akeson and Branton 2016) An ionic current is passed through the pore by setting a voltage across this membrane. If a molecule such as ssDNA passes through the pore or near its aperture, a disruption in current with a distinct signal is created. Evaluating this time-dependent signature leads to the identification of the molecule that entered through the pore (https://nanoporetech.com/how-it-works).

1.3 Methods for Investigating Bacterial Genome Evolution

The lack of an adequate number of informative markers limits the resolution for profiling bacteria strains from identifying novel variations at other loci (Achtman 2008). Although alternative methods have been proposed for distinguishing microbial strains (Coughlan et al. 2013; Downing 2015), bacterial genome sequencing has largely replaced classical tools e.g. Multilocus sequence typing (MLST) and Pulse field gel electrophoresis (PFGE), for microbial screening and analysis. Indeed, genomic sequencing is now becoming a routine in clinical microbiology (Page et al. 2016).

Deep investigation of bacterial genomes is the most optimal strategy to tackle ST131. It allows prediction of major virulence markers and antimicrobial susceptibilities and elucidate this ST's evolutionary and transmission history (Price et al. 2013; Petty et al. 2014; Salipante et al. 2015; Stoesser et al. 2013 and 2016; Ben Zakour et al. 2016). In this context, reviewed here are the methods we used to analyse ST131 genomes. Moreover, with the exponential growth of publicly available genome databases, the aim is to provide a better understanding of computationally robust, open-source and state-of-the-art methodologies from analysing many read libraries, including the processing of reconstructed prokaryotic genomes, particularly of *E. coli* ST131; the citations are however limited to studies that are of fundamental importance to the development of a certain strategy and to platforms that are either recent or best illustrates how that strategy works.

1.3.1 Assessing read quality

To get optimal genomic assessment results, raw read libraries processed by sequencing platforms need to eliminate adapter sequences, low-quality reads and other contaminants (Aronesty, 2011; Magoc et al. 2013). Often, these platforms' varied sequencing chemistry and contaminants including DNA from library preparation reagents (de Goffau et al. 2018) lead to issues that affect the quality of the reads: either read quality plummets at the end of the read or the quality of the second read is suboptimal compared to the first (http://www.ark-genomics.org/events-online-training-eu-training-course/assessing-quality-illumina-data). Hence, it is highly important to perform quality control checks to ensure that the data is both accurate and reliable.

The FASTX-Toolkit is a collection of Linux command line tools for processing FASTQ files and allows evaluation of base quality and nucleotide distribution in a sequence file (http://hannonlab.cshl.edu/fastx_toolkit/). A superior package that gives a more detailed read quality report is FastQC, developed by Babraham Bioinformatics Institute. FASTQC is a modular tool (www.bioinformatics.babraham.ac.uk/projects/fastqc/) that provides a

straightforward way to do quality control on raw sequence data obtained from high-throughput sequencing projects. It generates summary graphs and tables showing the quality of the sequencing library. Included in the report is the Phred quality score (Q score) that measures base calling accuracy and indicates if a given base is incorrectly called by the sequencer. Thus, a Q score is a function that is logarithmically related to the base calling error probability (P) such that Q =−log10P (Ewing B and Green P., 1998). Compiling all the reports produced by FastQC is best implemented by MultiQC, which parses the summary statistics obtained from the results and log files produced after running other bioinformatics tools (Ewels et al. 2016).

1.3.2 Sequence trimming, filtering and error correction

The first step to avoid false positive and false negative results is to filter out duplicated reads, remove sequences that are short (<50 bp) or with low GC content, and exclude those with many ambiguously-called bases. Several assembly tools incorporate adapter removal and error-correction steps in their package, like Fastp, which includes a comprehensive features to do all-in-one pre-processing of raw fastq files: it removes low quality (Q<30), short (<50 bp; preventing the occurrence of sequence duplicates) or reads with many unknown bases in a single file, it cuts adapters and correct mismatched base pairs in overlapping regions of paired-end reads (Chen et al. 2018A).

Most genome assemblers involve the most basic step of generating a *de Bruijn* graph that represents all k-mers occurring in the input read data together with the overlap between them (Compeau et al., 2011). Thus, the resulting assembled genomic sequence can be traced back through the path of its *de Bruijn* graph (Heydari et al. 2017). However, this task is significantly complicated with the presence of errors in sequence reads. This is because, if a single sequencing error in a read occurs in up to k erroneous k-mers in the *de Bruijn* graph, these k-mers produce artefacts in the graph resulting to spurious dead ends, parallel paths and chimeric connections (Zerbino and Birney 2008). In some instances, erroneous k-mers can greatly outnumber true k-mers blurring the process of calling the right sequence in a read file (Heydari et al. 2017). The tools for correcting these errors thus try to identify and correct these by creating a *k*-mer coverage frame from the input reads and replacing those with very

low coverage *k*-mers by similar *k*-mers with a much higher coverage. Further details of assembly software and process will be discussed in the succeeding sections.

A number of independent read error correction tools were evaluated in great detail in a review by (Alic et al. 2016). Of all the software assessed in this paper, BayesHammer is the best perfoming. BayesHammer is designed for error-correction and uses algorithms based on Hamming graphs and Bayesian sub-clustering. Although it operates without an assumption of uniformity, it yields significant improvements over other available error correction tools (Nikolenko, Korobeynikov and Alekseyev 2013). In addition, BayesHammer performs excellent correction of both short and long reads that have non-uniform coverage—an advantage as to why this tool is incorporated in highly reliable assembly pipelines like SPAdes (Bankevich et al. 2012) and Unicycler (Wick et al. 2017).

1.3.3 De novo assembly of genomes

The complex genetic identity of bacterial strains is defined by their core (chromosomal) and variable accessory (usually contained in plasmids) genomes (Segerman et al. 2012). AMR in *E. coli* ST131 are influenced by several core genome mutations and plasmid genes encoding different ESBLs (Downing et al. 2015). A crucial step to understand ST131 genome diversity is the reconstruction of their genome sequences or *de novo* assembly.

De novo genome assembly is done without prior information about the sequence length, layout or composition of the sample DNA (Alkan et al. 2011). A high-quality genome assembly serves as a foundation for research into broader scale studies on bacterial genomics. Most studies that explored ST131 genomes used de Bruijn graph-based platforms to perform assemblies (Petty et al. 2014; Hargreaves et al. 2015; Sidjabat et al. 2015; Stoesser et al. 2013, 2016) likely because the method is straightforward and is much more efficient than OLC for piecing short reads together (Zerbino and Birney, 2008). De Bruijn graph methods use smaller sub-fragments (k-mer) to lessen computational memory requirements through a smaller search space (Li et al. 2012). *De novo* assembly of genomes is achieved by linking contigs together

into scaffolds (Hunt et al. 2014) and finally correcting errors in the completed assembly (polishing).

- Read assembly

The first step in assembling a genome is to merge together of DNA reads into contiguous sequences (contigs) with the same base composition as the template DNA where the sequences were derived from. This process is often done in variable ways, often depending on several factors i.e. depth of sequence coverage, the number of times sequencing is done over each base in a genome. A coverage (or depth) of 30-fold means that a particular nucleotide in a certain genomic position is read by 30 fragments of the total sequence; the higher the coverage, the more accurate the inference of the base. Assessment of gene order and synteny, performing comparative and functional genomics or identifying recombination patterns are all dependent on an assembly with high quality continuity.

- Scaffolding

Scaffolding is joining contigs together into longer sequences termed as scaffolds. This step is computationally intensive and can be problematic with the presence of repetitive sequences (repeats) in the library. The solution is often to use multiple assemblers or run several parameters and select the one that gives the best summary statistics (Hunt et al. 2013). Genome assemblers are classified according to three methods: overlap layout-consensus (OLC; Flicek and Birney, 2009), de Bruijn graph (Li et al. 2012) or String graph (Myers, 2005). Further significant improvements have been made in assembly algorithms regarding reliability and efficiency of reconstructing genome architecture (Namiki et al. 2012; Bankevich et al. 2012; Boisvert et al. 2012; Peng, et al. 2012; Pell et al. 2012): with greater N50s, fewer rearrangements and break points, whereby the assembler cannot resolve the underlying sequence when a repeat region is larger than the fragment size (Page et al. 2016). The N50 is the size of the largest contig for which 50% of the total size is found in contigs of at least that length, otherwise known as the weighted median size of a contig.

- Contiguation

Contiguation is the alignment, ordering and orienting of contigs/scaffolds to close gaps in the draft assembly. This process usually involves finding alignment positions and identification of syntenies in assembled contigs (Assefa et al. 2009).

- Assembly polishing

The availability of sufficient data produced by modern sequencing platforms allow us to conduct deeper genomic investigations of hundreds of bacterial genomes in a day. NGS technologies have proven their contribution by providing low-cost, high-throughput short reads and the use of single molecule sequencing is now evolving to become the routine for generating whole genome assemblies. However, previous reviews of these technologies have extensively shown the unsuitability of NGS reads to assemble complex genomes and the susceptibility of long reads to cause frameshifts in ORFs (Watson 2018). Further, *en masse* typing of sequence variation in genes and genomic DNA challenges the robustness, sensitivity and scalability of current platforms for data processing (Mir and Southern, 2000).

Ensuring the completeness and accuracy of a genome assembly is essential to avoid errors in downstream analyses. Incorporating this step-in assembly of genomes promotes correct identification of gene content and appropriate inference of genetic evolution and helps researchers make the most of the available sequence data type. Several assembly polishers were hence developed to improve draft assemblies and eliminate single bases and small insertion/deletion events (Ronen et al. 2012), gaps (Swain et al. 2012) and alignment discontinuities (Hunt et al. 2013). Pilon was developed to correct all these error types and merge multiple assemblies to a consensus high-quality assembly--more optimally when supplied with PE reads. Although computationally extensive, this tool was proven to accurately distinguish small variations as well as resolve large insertions and identify large sequence variants e.g. duplications (Walker et al. 2014). For long read assemblies, Racon has an independent consensus module for ensuring a high-quality assemblies from Pacbio and ONT reads (Vaser et al. 2017). Results of data simulation showed that Racon works best when partnered with Miniasm-assembled (Li 2016) genomes.

1.3.4 Genome assemblers

Three main strategies can be used to improve assemblies: using PE reads from libraries with various insert size lengths, combining different types of short and long reads from Illumina and Pacbio/ONT (Wick et al. 2017) and utilizing a reference genome to fill the gaps in *de novo* assemblies (Nishito et al. 2008). The last strategy is termed as comparative assembly, which gets better results when the target genome is highly similar to the reference genome because it can resolve repeats more easily (Pop et al. 2004).

Choosing the best assembler (or set of assemblers) is key to obtaining the most accurate reconstructed genome. GAGE-B (Genome Assembly Gold-standard Evaluations) evaluated the performance of assembly tools to identify which generated best assemblies of bacterial genomes from a single shotgun library. Moreover, GAGE-B assessed the appropriate coverage depths and other parameters used to produce optimal results, determining the difference between using a high coverage single library with those of multiple libraries, and analysed the effect of longer 250 bp MiSeq reads compared to 100 bp HiSeq reads regards the final assembly output. Assemblers assessed with GAGE-B included AByss v1.3.4 (Simpson et al. 2009), CABOG v7.0 (Miller et al. 2008), MIRA v3.4.0 (Chevreux et al. 2004), MaSuRCA v1.8.3 (Zimin et al. 2013), SGA v0.9.34 (Simpson and Durbin, 2012), SOAPdenovo2v2.04 (Luo et al. 2012), SPAdes v2.3.0 (Bankevich et al. 2012) and Velvet v1.2.08 (Zerbino and Birney, 2008). The comparative evaluations revealed that MASurCA and SPAdes consistently generated contigs with the highest contig N50s with relatively few errors for both Illumina MiSeq and HiSeq assemblies of various bacterial species. However, GAGE-B highly depends on a template genome, which is unavailable when performing a *de novo* assembly (Magoc et al. 2013).

More recently, a superior pipeline called Unicycler was developed to perform short read-only, long read-only and hybrid (short and long reads) assemblies. Included in this platform is an assembly polisher tested to work best with either SPAdes or Miniasm for short or long read assemblies respectively (Wick et al. 2017).

1.3.5 Assessment of assembly quality

A way to assess assemblies without a reference sequence is by computing the maximum likelihood of an assembly given the error in the reads, the insert size distribution and the extent of unassembled data (Rahman and Pachter, 2013). This method had been initially proposed in earlier studies but did not take into account important parameters such as sequencing error (Myers, 2005; Medvedev and Brudno, 2009).

ALE (Assembly Likelihood Evaluation) and CGAL (Computing Genome Assembly Likelihoods) produce summary likelihood score of an assembly. CGAL does so by initially describing a probabilistic generative sequencing model that highlights different aspects of sequencing experiments (Rahman and Pachter, 2013) and ALE produces four likelihood scores for each base. FRCbam is another reference-free assembler evaluator can be employed to achieve the best assembly. FRCbam detects misassemblies or errors and ranks assembler performances by computing for the read and spanning coverages directly from BAM files. The misassemblies and errors are used to plot a feature response curve (FRCurve), overlaying these curves determines the best assembly (Vezzi et al. 2012).

Calculating for the fragment coverage distribution (FCD) is another tuning parameter to evaluate assemblies: the FCD is the distribution of coverage depths for fragments that contain the base and is measured on a per-site basis. REAPR (Recognition of Errors in Assemblies using Paired Reads) is a reference-free algorithm developed to improve assembly quality evaluation of by constructing a FCD plot of the fragment depth taken from the fragments mapped to a target base. A fragment (*f*) is the sum of the reads and their insert; that is *f = read1 + insert + read2*. The difference between the expected (ie chromosome median) coverage (*e*) to the observed coverage (*o*) at a given base is the FCD error. This statistic determines whether the scaffold should be broken or merged at that base and computes *e - f/b*, where *f* is the local coverage based on a mean fragment length *i*, *b* is the coverage at the base such that *o = f/b*. The resulting FCD score is the sum of the *e - f/b* scores for *-1.5*i* bases to *+1.5*i* bases

around the given base. High FCD error across regions indicates errors in the assembly (Hunt et al. 2013).

Quast works in a similar manner and extends its assembly evaluation by merging all the algorithms and parameters used by similar tools. It extends its evaluation power by including new metrics such as NA50, computing for the total number of misassemblies and misassembled contigs, and by quantifying the total predicted genes (Gurevich et al. 2013).

1.3.6 Plasmid reconstruction

Salient to the adaptation of a microorganism to its environment and adjustment to selective pressures are the genes that code for drug resistance and virulence, which are often encoded by plasmids. Difficulties in plasmid analysis and reconstruction arise due to their high sequence variation and repetitive sequences. Moreover, current genome-sequencing analysis protocols fail to evaluate genomic segments exchanged between plasmids and the chromosome, limiting the full evaluation of plasmid sequences and the pangenome (core and accessory genome) as a whole.

A method called PLACNET (plasmid constellation networks) has shed light on this issue by identifying, visualizing and analysing plasmids by creating a network of contig interactions, thus allowing comprehensive plasmid analysis. PLACNET uses three types of information to identify plasmids: (i) information about scaffold links and coverage in the WGS assembly, (ii) comparison to reference plasmid sequences, and (iii) plasmid-diagnostic sequence features such as replication initiator proteins. PLACNET combines these three types of data to produce a network that needs to be pruned manually to eliminate confounders. To identify the genomic sources, PLACNET searches or aligns assembled contigs against a database of ~6,000 publicly available chromosomal and plasmid genomes (Lanza et al. 2014; De Toro et al. 2014).

Although understanding the information stored in scaffolds is essential, abundant details on plasmids can be mined from decoding the structure of the *de Bruijn* graph. Given a situation when there are no long reads, this information can be uncovered by

assembling plasmids from sequence reads. PlasmidSPAdes can improve plasmid recovery from across species by reassemblying their genomes, determining their plasmids and obtaining the corresponding GenBank entries with the plasmid annotations (Antipov et al. 2016).

Plasmids may be underrepresented in long reads: in PacBio reads, a standard size-selection protocol may exclude short DNA fragments. Unicycler's graph-based scaffolding circularizes sequences, particularly those that are plasmid-derived, and results did not show duplicated sequences at the start/end of circular replicons (Wick et al. 2017).

In cases when long read assembly pipelines fail to complete assembly, an application such as Circlator (Hunt et al. 2015) can be used to manually circularize plasmid-derived scaffolds and distinguish them from chromosomal ones. Another solution used by mlplasmids is to use machine-learning (Arredondo-Alonso et al. 2018) to classify scaffolds as plasmid-derived or chromosome-associated contigs/scaffolds. Using support-vector machine (SVM) models on short read sequences, mlplasmids accurately classified contigs from assemblies of *E. faecium, pneumoniae* and *E. coli* as plasmid and chromosome; the results of their experiments showed that mlplasmids was the best classifier for the three bacterial species (Arredondo-Alonso et al. 2018).

1.3.7 Read mapping

Mapping is the alignment of sequence reads to reference sequence(s) such as a gene, a contig, a complete genome, transcriptome, or *de novo* assembly. The process involves predicting the position of each read relative to the reference genome. The quality of the alignment depends on the optimization of certain parameters such as the number of differences allowed between reference and query, the number of differences allowed in the seed, the number allowed and penalty for gap openings, and the number and penalty for gap extensions. The number of nucleotide differences should match the expected number of differences between two sequences being compared. Setting the value for this parameter means changing the number of mutations necessary to convert one string to another.

Read mapping algorithms are divided into two main categories: (i) hash table-based and (ii) Burrows-Wheeler Transform-based algorithms. Hashing algorithms transform a string into a key that allows a fast search during alignment (Figure 1.3). This process is computationally intensive and requires huge disk space: storing all *k*-mers in a list, using a value of *k* significantly less than the read size is viewed as a solution. Burrows-Wheeler Transform-based algorithms include the use of suffix trees and suffix arrays. Suffix tree present one-to-one correspondence between the paths from root to tips, with the suffixes existing in a string, so that the string suffixes serve as a path joining the root to the tip in a given tree. A suffix array is used as an alternative to the suffix tree, which can use a lot of memory. A suffix array is the set of suffixes of the genome sorted lexicographically (Figure 1.4).

Determining the best read-mapping tool for bacterial species is a continuing topic of research. One reason is due to the presence of genetic heterogeneity in a clonal population of cells, the quality control steps undertaken, differing coverage levels and sample-reference genome divergence. The most informative way of approach this is to *de novo* assemble and map the reads back to the assembly, where no difference should be expected if the assembly is perfect. A well-studied reference genome and its read libraries can be used to do this. Other studies have performed titrations of read mixes from multiple samples with known SNPs or simulated read errors/variants. Quantifying the percentage of reads mapped for simulated reads show that Smalt is better than BWA, especially as read diversity increases. Likewise, simulated *Listeria* reads (Ponstingl et al. 2010) show that Smalt, Burrows-Wheeler Aligner (BWA) (Li and Durbin 2009), MOSAIK (Lee et al. 2014) and SequenceSearch and Alignment by Hashing Algorithm (SSAHA) (Ning, Cox and Mullikin 2001) had no consistent differences for true SNP detection, and that Smalt was much better for eliminating false SNPs. Using Bowtie (Langmead et al. 2009) or Bowtie2 (Langmead and Salzberg 2012) is not recommended for bacterial genomes as both were they were designed to meet computational speed requirements i.e. in read mapping human reads rather than bacterial. Although the various *E. coli* genomics studies have used several different mappers, Smalt remains to be the most effective at present followed by BWA.

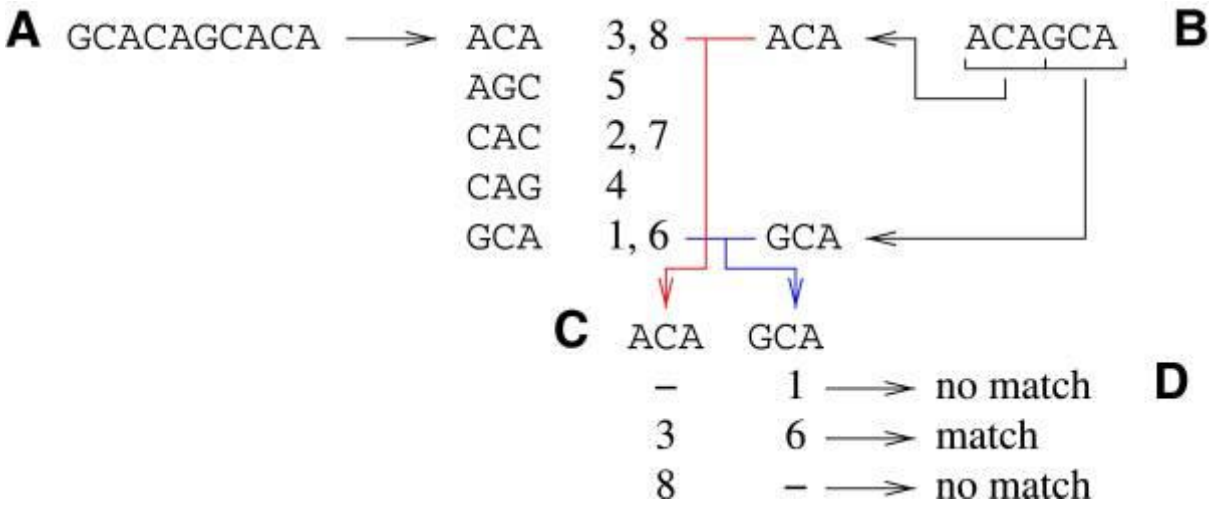


Figure 1.3. A hashing algorithm. (A) The genome is cut into overlapping 3-mers, and their respective positions in the genome are stored. (B) The read is cut into 3-mers. The 3-mers from the reads are compared to 3-mers from the genome using a hashing procedure. (C) Positions for each seed are sorted and compared to the other seeds. (D) Compatible positions are kept. Adapted from Schbath et al. 2012.

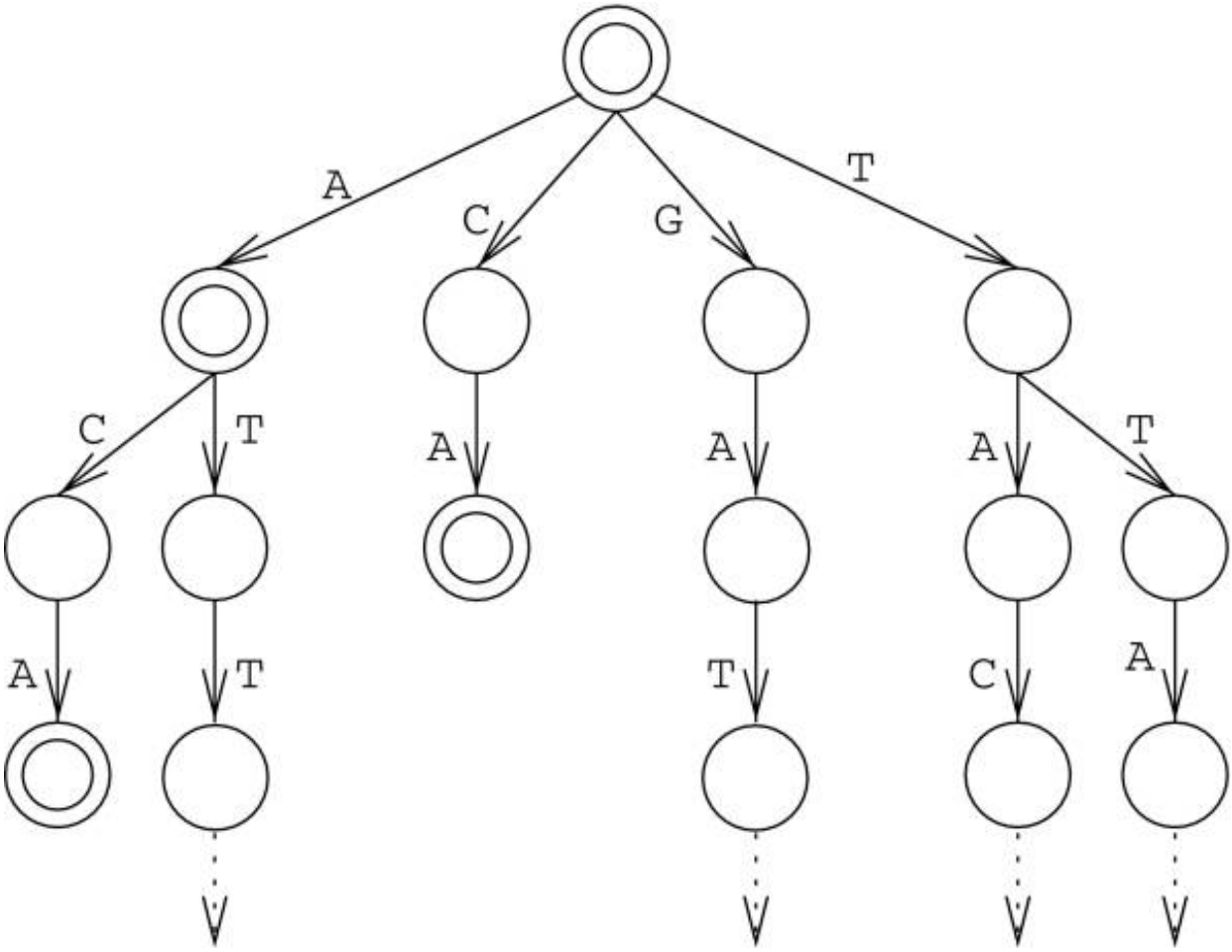


Figure 1.4. A suffix tree of the genome GATTACA. Dotted arrows indicate that the tree continues there. Double circle indicate that a suffix ends there. Adapted from Schbath et al. 2012.

1.3.8 Variant calling

Relative to eukaryotes, bacterial genomes are small and less complex, yet they remain to be one of the most difficult targets of genomic analyses (Audano, Ravishankar and Vannberg 2017). Acquisition of novel DNA segments through horizontal gene transfer (HGT) causes significant genomic changes (Vos et al. 2015) in a sample genome when compared with a reference. The standard method to identify single nucleotide polymorphisms (SNPs) or variant alleles (known as SNP/variant calling) is achieved by either mapping the raw reads to a reference genome or by assembling the query genome *de novo* then aligning it to the reference sequence.

Using genome assemblies to call SNPs as is useful for analysing individual genes, for constructing whole genome phylogenies Gardner and Hall, 2013) or if raw read data are not available. The Cortex variation assembler pipeline claims to perform this process best (Iqbal, Turner and McVean 2013). But mapping reads to a high-quality reference genome provides a deeper resolution when identifying SNPs because it is more feasible to quantify the depth of coverage and the proportion of mixed alleles using raw reads than an assembled genome. A review by (Sandmann et al. 2017) evaluated the top SNP calling applications, most of which are applicable to bacterial genome analyses such as the GATK HaplotypeCaller (DePristo et al. 2011), Platypus (Rimmer et al. 2014), LoFreq (Wilm et al. 2012), and SAMtools (Li et al. 2009). The assessment showed that none of these succeeded in calling all true SNPs and highlighted the need to improve variant calling strategies.

In a situation where there is no available reference genome, or one is trying to map reads to a highly divergent region (Bertels et al. 2014), the approaches above cannot be applied. A possible solution to this is to use a program such as Kestrel, which takes the information within a set of k-mer frequencies over the read data. Kestrel can also characterize densely packed SNPs and large indels without performing mapping, assembly or generation of de Bruijn graphs (Audano, Ravishankar and Vannberg 2017).

1.3.9 Genome annotation

Genome annotation or gene prediction opens doors to further investigate functional properties. Annotating genomes has to take into account several dimensions (Bryson et al. 2006). It involves identifying gene sequences including open reading frames (ORFs) and stop codons. The problem of genomic annotation is clearly defined as having an input sequence of DNA (X) -- each element (x) can be either of the four nucleotides A, T, C, G, that is X = (x1,...xn) $\sum^*$, where $\sum$ = {A, T, C, G} and should be correctly labelled as a gene-coding region (Bandyopadhyay et al. 2008). The easiest way to identify genes is to look for ORFs -- continuous stretches of codons that code for a protein and are halted by a stop codon (UAA, UAG or UGA).

In contrast to eukaryotes, bacterial genomes have higher gene density with more than 90% protein-coding regions (Wang et al. 2004). Determining the real genes within six frames of bacterial genome is therefore the main task of gene predictors. Identifying ORFs with the mean size of proteins (roughly 900 bp) is a way to locate genes (Allen et al. 2004): this strategy is effective for pin-pointing small genes but not those with long ORFs, so annotation pipelines require manual verification to predict tRNA, rRNA and ORFs.

False positive annotations remain a concern, so predicted genes should be screened manually to correct start codon positions, gene names, gene products, frameshifts in the alignment and resolving issues on ORFs caused by overlapping sequences that are coding forward or reverse. Various annotation programs are available online (Goel et al. 2013; Seemann, 2013) and most run on several Hidden Markov Models (HMMs) and BLAST-based gene prediction models.

The Rapid Annotations using Subsystems Technology (RAST) Server was designed to predict which sets of assertions of gene and protein function from a report to genes derived from other genomes (subsystems) match the genome of interest and uses this to construct the metabolic network. The results can be viewed in an environment that supports comparative analysis with other annotated genomes (Aziz et al. 2008). An extension and a modular implementation of this program called RASTtk (RAST tool kit) was created to handle annotation of batch genomes (Brettin et al. 2015).

More recent studies on ST131 evolution used the Rapid Prokaryotic Genome Annotation (PROKKA) pipeline in gene prediction (Stoesser et al. 2016; McNally et al. 2016). PROKKA annotates bacterial genomes rapidly using a quad-core desktop computer. It annotates a sequence by relying on external feature prediction tools such as Prodigal (Hyatt 2010) for coding sequences and Aragorn (Laslett and Caback, 2004) for transfer RNA genes. When tested against RAST and another program called xBASE2 (Chaudhuri et al. 2008) to annotate *E. coli* K-12 genome, PROKKA produced the most accurate results in about six minutes (Seemann, 2014).

1.3.10 Recombination detection and analysis

Bacterial recombination can occur after DNA enters a recipient (host) cell via conjugative transfer of plasmids (conjugation), uptake of naked genetic material from lysed cell by another living cell (transformation) or the recruitment of DNA material from a phage to the recipient cell (transduction). Additional HGT mechanisms include nanotubes and extracellular vesicles. Identifying these recombinant segments is important as the most frequently retained blocks after DNA exchange are genes associated with antibiotic resistance (Chewapreecha et al. 2014). Further, recombination facilitates DNA exchange between distantly related species and highly influences the evolution of novel strains. A majority of the most virulent bacterial pathogens belong to monomorphic lineages that show little genetic diversity (Achtman 2008). These clones are thought to have survived population bottlenecks by losing and gaining genetic elements leading to the acquisition of new functions, which may or may not be beneficial to their evolution.

High recombination frequencies may bring benefits but can cause disadvantages due to increased instability as the GC content of these blocks are usually different their sources and hence are prone to degradation (Chewapreecha et al. 2014; Nishida, 2013). Moreover, analysis of imported sequence within closely-related isolates was proven to be more valuable in investigating genomic variation in bacteria (Smith 1992) rather than merely identifying the movement of a segment from a donor to a recipient (Sneath et al. 1975).

The genealogical history of a sample can be represented by a single phylogenetic tree based on a selected DNA substitution model. But with the presence of recombination events, a number of trees can be created based on different positions, which make it impossible to infer the true genealogy of a single strain (Griffiths and Marjoram, 1997). Several statistical approaches have been utilized to assess heterogeneity caused by recombination or HGT. Recombination to mutation (*r/m*) ratios vary across bacterial species and are associated with higher SNP rates (Vos and Didelot, 2009). Here, the three most commonly used tools used to assess HGT events in ST131 are outlined:

- BRATNextGen

Bayesian Recombination Tracker Next Generation (BRAT-NextGen) requires a MATLAB compiler and runs on a Bayesian change-point clustering model that identifies a tip with at given genomic region that is more distinct from the other isolates that may have evolved from mutation (Martinnen et al., 2012). BRATNextGen does not model HGT between the samples, and usually predicts HGT that occurred recently in the external branches of an ML tree. The recombination model configuration in the software is initialized by the taxa that were formed over a fixed-width genomic interval to maximise variability on the branches so that separate imported segments are identified from the same ancestral origin. The clustering procedure estimates the transition probabilities from a non-recombining to non-recombining (*1-p*) or recombining (*1-p*$_0$) α clusters, such that $p_0 = P(n+1=non\text{-}recombining \mid n=non\text{-}recombining)$ from non-recombinant state. The results are summarized in a Proportion of Shared Ancestry (PSA) tree that highlights clusters with common recombination events. Statistical significance of the detected imported segments is assessed by permutation testing of SNPs.

- ClonalFrameML

ClonalFrame Maximum Likelihood (ClonalframeML) performs recombination inference in a maximum likelihood (ML) framework (Didelot and Wilson, 2015). It links with ML trees constructed with phylogenetics tool RAxML (Randomized Axelerated Maximum Likelihood; Stamakis, A. 2006). It initiates using the RAxML phylogeny followed by the reconstruction of ancestral sequences at internal nodes of the clonal genealogy, and any missing base calls in the observed sequences (Pupko et al. 2000). To obtain ML estimates of the recombination parameters and the branch lengths of the clonal genealogy, ClonalframeML uses the Baum-Welch Expectation-Maximisation (EM) algorithm, and the per site importation events are determined using a Viterbi algorithm. Bootstrapping is used to quantify uncertainty in the parameters. ClonalframeML assumes the same values for the number of recombining segments (R/θ), the mean length of imported DNA (δ) and the mean divergence of imported tracts (ν) for all branches, and that the length of branch i, in terms of the expected number of mutations. Since it employs an ML tree, it measures the lengths

of branches and the recombination rate in units of expected numbers of mutations. The enhanced detail obtained from the computed phylogenetic branch lengths allows for accurate quantification of genetic diversity and dispersion along each internal and external branch (Swenson, 2009). The *r/m* ratio is computed as: $\frac{r}{m} = \frac{R}{\theta} x \delta x \upsilon$. ClonalframeML uses a HMM where each nucleotide was subject to recombination (or not) on the branch connecting the two genomes; the nucleotides that were unaffected by recombination are called unimported (U) and those that are subject to recombination are termed as imported (I).

- Gubbins

Gubbins uses a spatial scanning statistic to detect highly variable loci suggestive of recombination and constructs a ML phylogenetic tree based on non-variable sites of the bacterial genome (Croucher et al., 2015). This iterative algorithm is most suitable for the reconstruction of recent evolutionary history as it integrates all polymorphic site information in a given data set before identifying any horizontal gene transfer event. The polymorphic sites were first detected in the alignment file using (for this thesis) a GTR substitution model with a Gamma distribution (GTR+G) of site variation. Clustering of substitutions were identified by performing a non-parametric scan on each branch. Possible recombination events were determined as a set of sliding windows with elevated densities of base differences.

1.3.11 Resolving Bacterial Population structure

Population genetics accounts for the diversity of natural populations and formulate theories that cause that variability (Smith, 1989). A population is a group of organisms that share a common geographical niche and possesses the ability to interbreed or exchange genetic material either sexually or asexually. Distinguishing a pathogenic organism from another is paramount to clearly explain the epidemiology of infectious diseases (Wailan et al. 2018). In addition, assessment of the genetic relatedness of these organisms shed light to their population structure. The high rates of recombination events in bacterial populations often contribute to the complexity

of their structure. Indeed, the bacterial population structures can only be understood by accurately quantifying recombination (Spratt, 2004).

Reconstructing phylogenies is a major hurdle in resolving population structure. This is due to the loss of phylogenetic signals in deep branches and the occurrence of frequent horizontal gene transfers and hidden paralogies. In a phylogenetic tree model for the evolution of a fixed population without acting selective forces, a single sample is represented as the lineage, and the lineages coalesce back in time at constant rate until only one lineage remains (Lawson, 2015).

A Supertree approach developed by (Daubin, Gouy and Perrière 2002) tackles these limitations by taking into account molecular phylogenetic information of hundreds of genes and provides a way to cumulate all of the phylogenetic signal while considering its statistical significance. By building a phylogenetic tree of 45 strains, they were able to identify the core genes by phylogenetically analysing the congruence of tree topologies and use these gene sets to infer a consensus tree. Although, this approach provides excellent support for a number of bacterial lineages, some internal branches remain unresolved and some clustering of strains may only be due to systematic reconstruction artefacts (Daubin, Gouy and Perrière 2002).

Several other techniques were developed to resolve bacterial population structure. One approach uses Bayesian statistical models implemented in BAPS (Bayesian Analysis of genetic Population Structure; Corander et al. 2008) and hierBAPS. BAPS initially fits genetic mixture and admixture models using a fixed number of populations, followed by comparing *a priori* specified biological hypotheses about the population structure and finally analyses the admixture using a genetic linkage model. HierBAPS models variation in DNA sequences and employs hierarchical clustering of DNA sequence data to uncover nested genetic population structures (Cheng et al. 2013).

The FastBAPS (Fast Hierarchical Bayesian Analysis of Population Structure) tool developed by (Tonkin-Hill et al. 2018) extends BAPS and hierBAPS. FastBAPS rapidly assigns an approximate fit to a Dirichlet Process Mixture model (DPM) to cluster

multi-locus genotype data that are 10-100 times larger than previously examined in other clustering methods (Tonkin-Hill et al. 2018).

Structure (Pritchard et al., 2000) uses multi-locus genotype data such as SNPs and microsatellites as inputs to investigate the structure of bacterial populations. Structure infers whether distinct groups or populations are present or absent, assigns isolates to the identified populations, identifies migrants and admixed isolates and provides an estimation of allele frequencies in a population in the presence of migrants or admixed isolates.

Phylogenetic networks demonstrate definitive scenarios of evolutionary reticulation exhibited by HGT or recombination (Kloepper and Huson, 2008). SplitsTree infers phylogenetic networks using different input data like sequence alignment, a distance matrix or a group of phylogenetic trees (Huson and Bryant, 2006). It analyses hybridization or recombination networks using split decomposition (Bandelt and Dress, 1992) or NeighborNet (Bryant and Moulton, 2003) algorithms. Inference of population structure can also be done using dense haplotype data. This method is implemented in fineSTRUCTURE (Lawson et al. 2012), which uses a "chromosome painting" approach to characterize shared ancestry and considers that a stretch of DNA is transferred from one generation to another in chromosomes.

These traditional genetic clustering algorithms, however, are not suitable for sub-typing of low-variant (LV) bacterial populations over small timescales such as less than three years (Wailan et al. 2018). Further, the population clusters formed using these strategies require >10 SNV to achieve an acceptable confidence and are based on the assumption that loci are independent of each other. This problem is overcome by the R package rPinecone, which identifies sub-lineages within LV bacterial populations. To accurately distinguish sub-clusters from each other in a given population, rPinecone assesses the phylogenetic relationship between bacterial strains by computing for the root-to-tip and the SNV distances from ancestral nodes (Wailan et al. 2018).

1.3.12 Pangenome analysis

The pangenome was first coined by (Tettelin et al. 2005), and is the entire genomic repertoire of a given phylogenetic clade. Another definition is that the pangenome is the set of all genes present in the genomes of a group of organisms (Lapierre and Gogarten 2009). Advances in genome sequencing allows for the classification of the pangenome into two parts: the core (usually chromosomal genes common to all strains) and the accessory (plasmid and mobile genetic elements, MGEs) genome (Rouli et al. 2015). McNally et al. (2016) emphasize the importance of analysing the pangenome as it gives a 'super-resolution' view into the evolution of bacterial population. Indeed, interrogating pangenomic datasets can provide a comprehensive genetic landscape and detailed insights into the genetic structure of prokaryotic genomes as well as identify their lineage- and niche-specific markers of evolution and adaptation (Page et al. 2015, Zhang and Sievert 2014, Kim et al. 2017).

A number of approaches exist in performing pangenome analyses. One is Roary, which rapidly builds large-scale pan genomes to identify core and accessory genes in a single species. Roary uses one annotated assembly per sample (GFF/GFF3 format). The coding regions are then extracted and converted into protein sequences and iteratively pre-clustered using CD-HIT (Fu et al. 2012) to obtain a significantly reduced number of protein sequence set. Next, BLASTP aligns the sequences at a certain % identity set by the user (i.e. 95% by default). Similar groups of paralogs are separated into groups of true orthologs before a graph is constructed based on the relationships of the clusters and in the order of occurrence in the input sequences. This ordering step provides context for each gene. The samples finally are grouped according to the presence of gene in the accessory genome taking into account the contribution of isolates to the graph weighted by cluster size. Roary can construct the pangenome of thousands of bacterial samples on a standard computer without compromising on the accuracy of results (Page et al. 2015). Piggy extends Roary by detecting highly divergent ("switched") intergenic regions (IGRs) upstream of genes (Thorpe et al. 2018).

Similarly, PGAweb is a web interface for prokaryotic pangenome analysis: orthologous clustering, pan-genome profiling, sequence variation and evolution analysis, and functional classification, which helpful for featuring genomic structural dynamics and sequence diversity (Chen et al. 2018B).

The algorithms implemented in by PopPUNK (Population Partitioning Using Nucleotide K-mers) examine the core and accessory genome variation by estimating the relative distances between pairs of isolates in large collections. PopPUNK does not align genomes but rather employs annotation data to analyse and cluster populations. The method is done using variable-length k-mer comparisons to differentiate between isolates' shared and non-shared sequence and gene content at the level of k-mers (Lees et al. 2018).

In summary, quantitatively and qualitatively describing the high genetic variation in bacterial populations is complex. It requires identification of the relative contributions of the evolutionary processes that cause genetic variation (Feil and Spratt 2001), particularly recombination, and processes in the population like selection and genetic drift. While many of the tools described in this section were used in our subsequent analyses, some were excluded and were applied mostly due to difficulty in package installation, low computational efficiency in large datasets, or redundancy due to the similarity in tool approaches.

1.4 Research Questions

This PhD project aimed to address the questions below by employing strategies developed for Population pylogenomics and Comparative genomics.

1.4.1 Population Phylogenomics

- What does the genome wide variation tell us about the demographic history of ST131 strains?
- How has the gene composition of the ST131 genomes changed over time?
- How has the acquisition of genes (e.g. common drug resistance genes) via clonal expansion or horizontal gene transfer contributed to the adaptation of ST131 to new niches?

1.4.2 Comparative Genomics

- Identify chromosomal/structural variations (SVs) that took place between isolate genomes.
- Detect (and classify) the type of copy number variations.
- Determine whether these changes are present in other isolates in the wider collection.
- Identify genes/features shared by the strains in a genome set.

1.5 Pilot study: Population Structure, Evolution and Recombination of N=100 E. coli ST131 Isolates from Long-term Care Facilities in Ireland/UK

Several tools were used prior to the commencement of this project to identify the major clades of ST131 with distinct genetic characteristics (Price et al. 2013; Petty et al. 2014). This scheme classifying ST131 strains as from either clade A, B or C has been conventionally used (Price et al. 2013; Petty et al. 2014) and we will explore this in the succeeding chapters. For this pilot study, we applied several phylogenomic methods to resolve the structure of the 100 ST131.

A number of methods including traditional classification schemes such as the multilocus sequence typing (MLST), whole-genome clustering systems and phylogenetics have been applied to understand the evolution of ST131. Although phylogenetic reconstruction may help in resolving ST131's evolutionary history, it could be challenging as the clone, much like other bacteria, undergoes several instances of horizontal gene transfer and homologous recombination.

The samples used in this section were taken from Catherine Ludden's unpublished doctoral thesis (Ludden 2014) and was further genomically characterized to identify the structure of the population. Her research focused on investigating acquisition of ARM among ESBL-producing species from Enterobacteriaceae and aimed in determining the baseline prevalence of colonization, monitoring the colonization status at quarterly intervals, assessing risk factors associated with colonization and finally, characterising antimicrobial susceptibility of the certain antibiotic resistant bacteria in LTCFs.

We also adapted and applied a classification system based on the *bla*CTX-M allelic content of strains to distinguish them from each other. Examining the *bla*CTX-M allele type in the representative strains of the major clades in the phylogeny of 100 ST131 indicated that plasmid-bound *bla*CTX-M-14 may have been the predominant *bla*CTX-M allele in Irish and UK ST131 isolates until it was partially displaced by a lineage with *bla*CTX-M-15, which was eventually integrated into their chromosomes. Identifying the genetic elements responsible for the *bla*CTX-M displacement and

integration will further clarify this initial observation and explored in detail in Chapters 3 and 4.

1.5.1 Genomic data

A collection of 100 ST131 isolates includes 90 from nursing homes in Ireland: 70 of them collected from 2005-2011. Swabbing of the residents was performed at 3-month intervals for a year. The Irish ST131 samples were initially planned to be collected from a total of 88 residents but only 51 of them were recruited and agreed to participate in the study conducted for the thesis of Catherine Ludden (Ludden 2014).

Seventy (70) of the N=100 ST131 strains collected were found to have *bla*CTX-M-1, 20 had *bla*CTX-M-9, 70 had TEM+, 4 with SHV+ and 20 with OXA-1+. The other 10 were UK samples taken from the study of Clark et al. 2012. The strain NA114 was used as the reference genome in the analyses of these 100 strains.

1.5.2 Inferring the genealogical history

Non-recombinant SNPs of 100 Irish/UK ST131 strains were aligned to each other for phylogenetic reconstruction. An ML phylogenetic tree was generated for these genomes using RAxML (Randomized Axelerated Maximum Likelihood; Stamakis, 2014) based on General Time Reversible and Gamma distribution (GTR+G) model visualized with Figtree v.1.4.3 (Rambaut, A., 2016). A phylogenomic network was also drawn from the same non-recombining SNPs using uncorrected p-distances and visualized with Splitstree v4.14.2 (Huson and Bryant, 2006).

1.5.3 Resolving the population structure

Two main approaches were used to understand the population structure of these 100 ST131 strains: [1] analysing of 8,687 genome-wide and 1,412 non-recombining SNPs using the change of the K likelihoods with Structure v.2.3.4 (Pritchard, Stephens & Donnelly, 2000). And [2] by determining the *bla*CTX-M allele of representative strains in a phylogenetic tree generated with RAxML and visualized using Figtree v. 1.4.3 (Rambaut et al. 2016) and label the clusters where each sample belong to according to Ben Zakour et al. 2016's clade classification.

1.5.4 Results from the Pilot Study

1.5.4.1 Classifying genetic clades according to their CTX-M allelic content clarifies the population structure of N=100 ST131 collection

The second-order rate of change of the K likelihoods ΔK from analysing 8,687 genome-wide showed that the most likely number of genetic populations in the 100-strain population were K=2 with cluster A and cluster B/C/C/AB/AC/O & K=5 with clusters A, B, C, AB and AC (Figure 1.5.1). Analysis of 1,412 non-recombining SNPs using the ΔK method resulted to K=2 with cluster A and cluster B/C/C/AB/AC/O & K=7 with A, B, C, AB and AC and two groups (O and O) assigned to divergent sample UTI226 (Figure 1.5.1). The reconstructed phylogenomic networks for both genome-wide (Figure 1.5.2A) and non-recombinant SNPs (Figure 1.5.2B) show the 5 major clades A, B, C, AB and AC with corresponding number of strains in each cluster.

The ML phylogeny generated with RAxML based on the whole genomes of N=100 ST131 (Figure 1.5.3) strains revealed three major clades with representative strains of varying *bla*CTX-M allelic background. The first and second main clusters contain plasmid-bound *bla*CTX-M-15 (represented by strain EC958) and *bla*CTX-M-15 (represented by strain MU027565L), respectively. The third group is composed of representative strains (MJ005670W, MU022181B and MU004181Y) with chromosomal *bla*CTX-M-15.

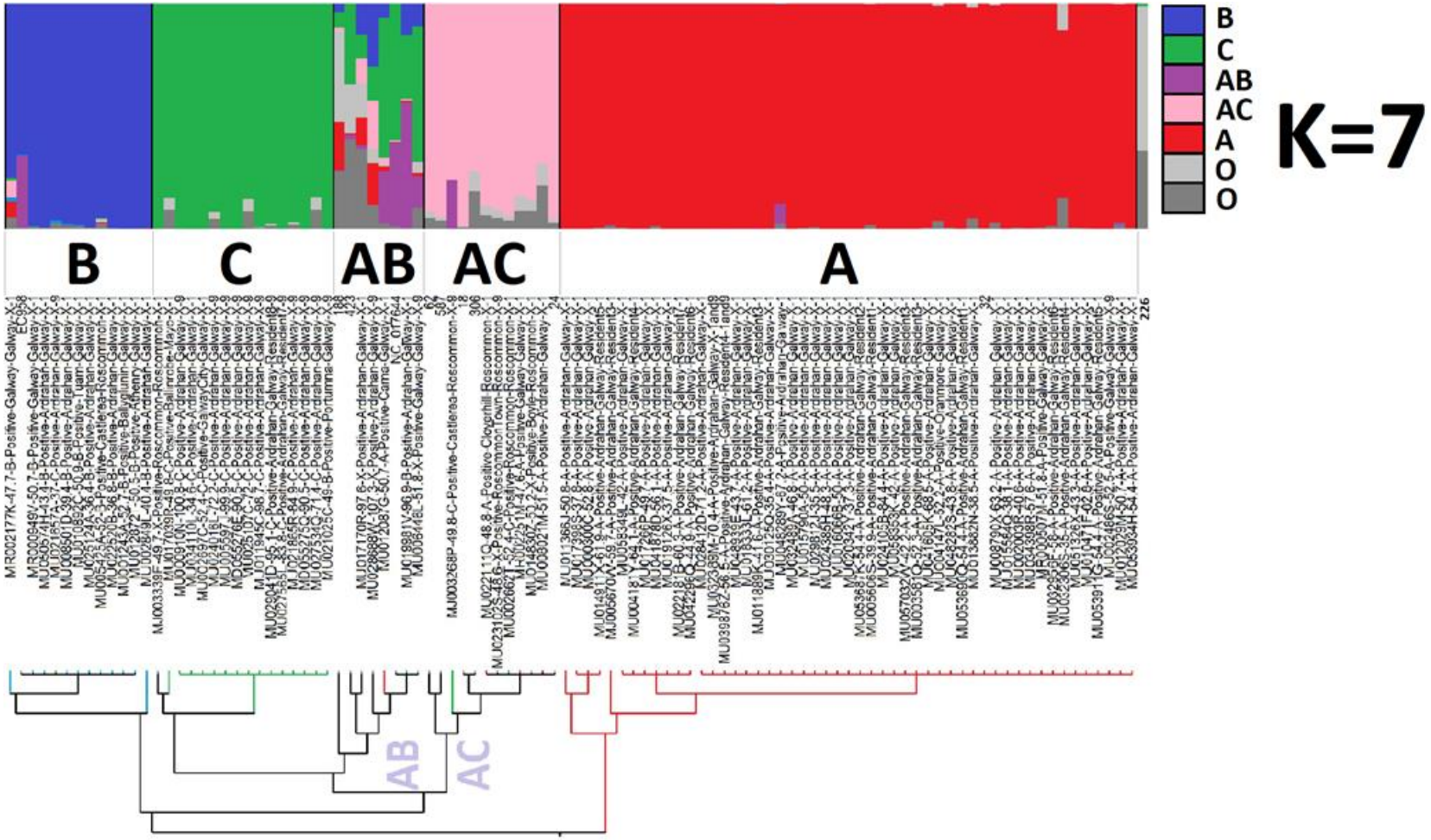


Figure 1.5.1. Model-based classification of groups using Structure bar plots for non-recombining SNPs for K=7 showed five main populations (A, B, C, AB and AC) and two groups assigned to divergent sample UTI226.

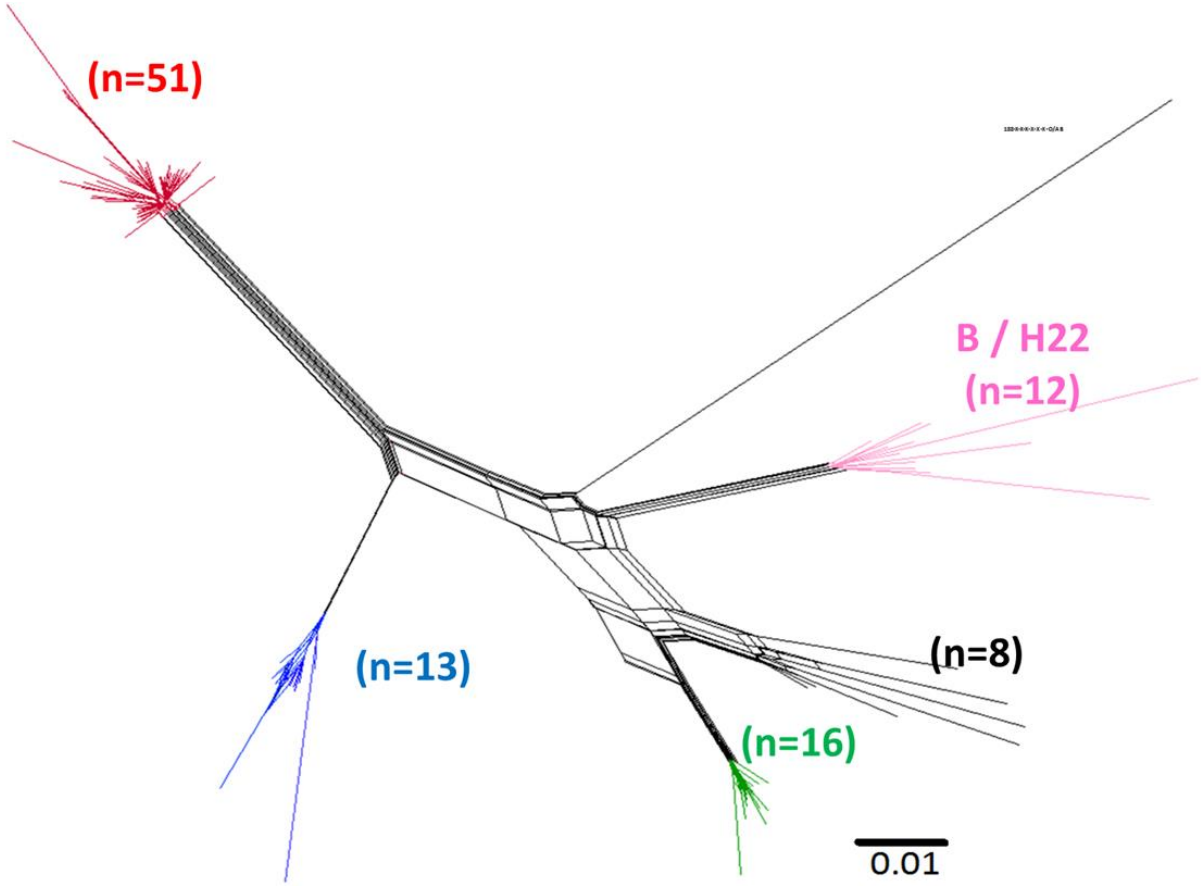


Figure 1.5.2A. Phylogenomic network of 100 E. coli genome-wide SNPs constructed using maximum likelihood method and General Time Reversible model in MEGA 6.0 and visualized with Splitstree v4.14.2. The branch lengths are measured in the number of substitutions per site using Uncorrected P distances. The ST131 subgroups classified using change of K likelihoods method of Structure v.2.3.4 are in red, blue, green, pink and black.

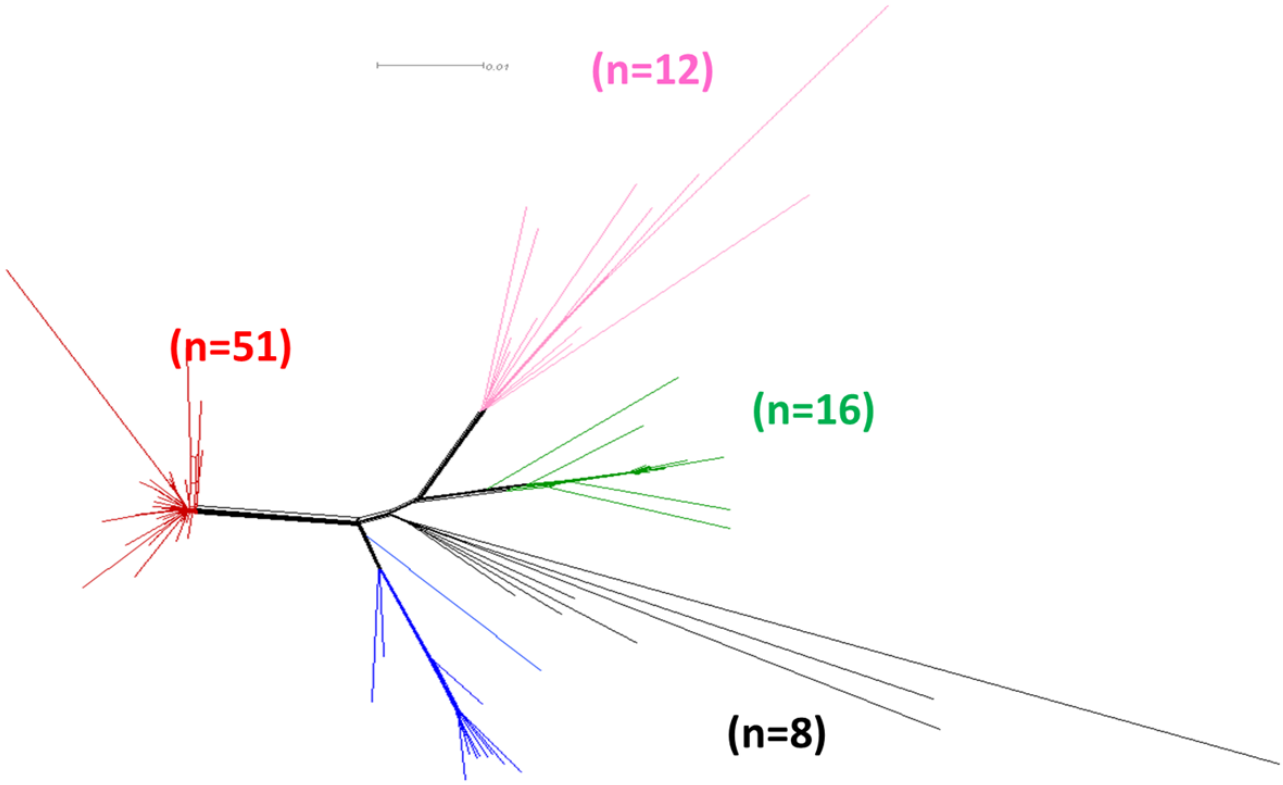


Figure 1.5.2B. Phylogenomic network of 100 E. coli genomes constructed from non-recombining SNPs (B) using maximum likelihood method and General Time Reversible model in MEGA 6.0 and visualized with Splitstree v4.14.2. The branch lengths are measured in the number of substitutions per site with Uncorrected P distances. The C2/ H30Rx subgroups are in red, blue, green, pink and black.

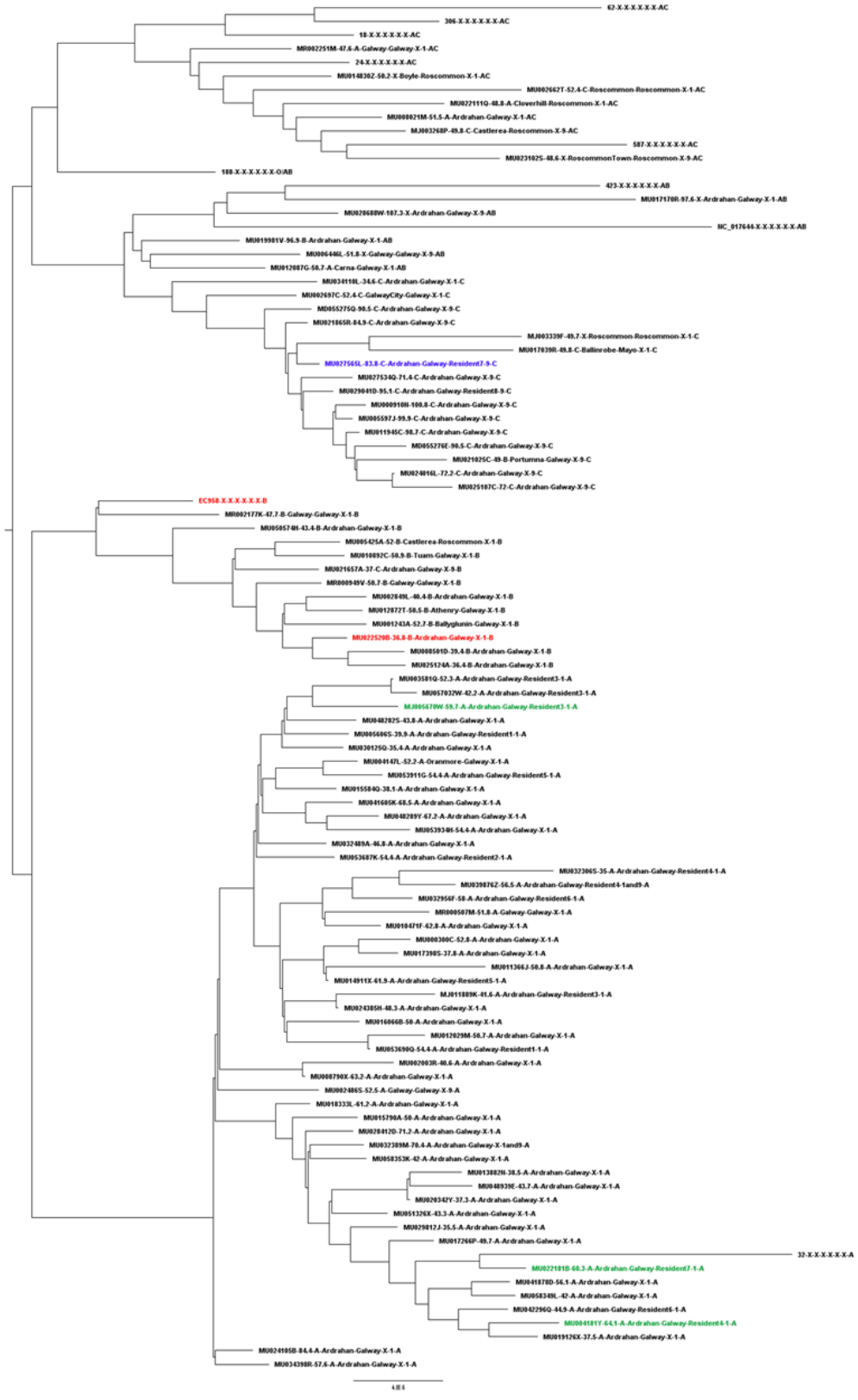


Figure 1.5.3. Genome-wide maximum likelihood phylogeny of N=100 ST131 Irish and UK samples generated using RaxML (GTR+G substitution model) and visualized with Figtree v.1.4.3. Tips highlighted in red had plasmid-bound *bla*CTX-M-15; taxon in blue contains *bla*CTX-M-14 in its plasmid and the samples in green font contain chromosomal *bla*CTX-M-15.

1.6 References

Achtman, M., 2008. Evolution, population structure, and phylogeography of genetically monomorphic bacterial pathogens. Annu. Rev. Microbiol., 62, pp.53-70.

Adams-Sapper, S., Diep, B.A., Perdreau-Remington, F. and Riley, L.W., 2012. Clonal composition and community clustering of drug-susceptible and resistant Escherichia coli isolates from blood stream infections. Antimicrobial agents and chemotherapy, pp.AAC-01025.

Alic, A.S., Ruzafa, D., Dopazo, J., Blanquer, I., 2016. Objective review of de novo stand-alone error correction methods for NGS data. Wiley Interdisciplinary Reviews: Computational Molecular Science 6, 111–146. https://doi.org/10.1002/wcms.1239

Alikhan NF, Zhou Z,Sergeant MJ, Achtman M. A genomic overview of the population structure of Salmonella. 2018 PLoS Genet 14 (4): e1007261 (https://enterobase.warwick.ac.uk).

Alkan, C., Coe, B.P. and Eichler, E.E., 2011. Genome structural variation discovery and genotyping. Nature reviews. Genetics, 12(5), p.363.

Allen, J.E., Pertea, M. and Salzberg, S.L., 2004. Computational gene prediction using multiple sources of evidence. Genome Research, 14(1), pp.142-148.

Antipov, D., Hartwick, N., Shen, M., Raiko, M., Lapidus, A., Pevzner, P.A., 2016. plasmidSPAdes: assembling plasmids from whole genome sequencing data. Bioinformatics btw493. https://doi.org/10.1093/bioinformatics/btw493

Aronesty, E., 2013. Comparison of sequencing utility programs. The Open Bioinformatics Journal, 7(1).

Arredondo-Alonso, S., Rogers, M.R.C., Braat, J.C., Verschuuren, T.D., Top, J., Corander, J., Willems, R.J.L., Schurch, A.C., 2018. mlplasmids: a user-friendly tool to predict

plasmid- and chromosome-derived sequences for single species. bioRxiv 329045. https://doi.org/10.1101/329045

Assefa, S., Keane, T.M., Otto, T.D., Newbold, C., Berriman, M., 2009. ABACAS: algorithm-based automatic contiguation of assembled sequences. Bioinformatics 25, 1968–1969. https://doi.org/10.1093/bioinformatics/btp347

Audano, P.A., Ravishankar, S., Vannberg, F.O., 2017. Mapping-free variant calling using haplotype reconstruction from k-mer frequencies 7.

Avasthi, T.S., Kumar, N., Baddam, R., Hussain, A., Nandanwar, N., Jadhav, S. and Ahmed, N., 2011. Genome of multidrug-resistant uropathogenic Escherichia coli strain NA114 from India. Journal of bacteriology, 193(16), pp.4272-4273.

Aziz, R.K., Bartels, D., Best, A.A., DeJongh, M., Disz, T., Edwards, R.A., Formsma, K., Gerdes, S., Glass, E.M., Kubal, M. and Meyer, F., 2008. The RAST Server: rapid annotations using subsystems technology. BMC genomics, 9(1), p.75.

Bandelt, H.J., Dress, A.W., 1992. Split decomposition: a new and useful approach to phylogenetic analysis of distance data. Mol. Phylogenet. Evol. 1, 242–252.

Bandyopadhyay, S., Kelley, R., Krogan, N.J., Ideker, T., 2008. Functional Maps of Protein Complexes from Quantitative Genetic Interaction Data. PLoS Comput Biol 4. https://doi.org/10.1371/journal.pcbi.1000065

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D. and Pyshkin, A.V., 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. Journal of computational biology, 19(5), pp.455-477.

Baraniak, A., Fiett, J., Sulikowska, A., Hryniewicz, W. and Gniadkowski, M., 2002. Countrywide spread of CTX-M-3 extended-spectrum β-lactamase-producing

microorganisms of the family Enterobacteriaceae in Poland. Antimicrobial Agents and Chemotherapy, 46(1), pp.151-159.

Ben Zakour, N.L.B., Alsheikh-Hussain, A.S., Ashcroft, M.M., Nhu, N.T.K., Roberts, L.W., Stanton-Cook, M., Schembri, M.A. and Beatson, S.A., 2016. Sequential acquisition of virulence and fluoroquinolone resistance has shaped the evolution of Escherichia coli ST131. MBio, 7(2), pp.e00347-16.

Bertels, F., Silander, O.K., Pachkov, M., Rainey, P.B. and van Nimwegen, E., 2014. Automated reconstruction of whole-genome phylogenies from short-sequence reads. Molecular biology and evolution, 31(5), pp.1077-1088.

Boisvert, S., Raymond, F., Godzaridis, É., Laviolette, F. and Corbeil, J., 2012. Ray Meta: scalable de novo metagenome assembly and profiling. Genome biology, 13(12), p.R122.

Brettin, T., Davis, J.J., Disz, T., Edwards, R.A., Gerdes, S., Olsen, G.J., Olson, R., Overbeek, R., Parrello, B., Pusch, G.D., Shukla, M., Thomason, J.A., Stevens, R., Vonstein, V., Wattam, A.R., Xia, F., 2015. RASTtk: a modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. Sci Rep 5, 8365. https://doi.org/10.1038/srep08365

Bryant, D., Moulton, V., 2004. Neighbor-net: an agglomerative method for the construction of phylogenetic networks. Mol. Biol. Evol. 21, 255–265. https://doi.org/10.1093/molbev/msh018.

Bryson K, Loux V, Bossy R, Nicolas P, Chaillou S, van de Guchte M, Penaud S, Maguin E, Hoebeke M, Bessieres P. 2006. AGMIAL: implementing an annotation strategy for prokaryote genomes as a distributed system. Nucleic Acids Res, 34(12):3533-3545.

Cantón, R. and Coque, T.M., 2006. The CTX-M β-lactamase pandemic. Current opinion in microbiology, 9(5), pp.466-475.

Canton, R., Gonzalez-Alba, J.M., Galán, J.C., 2012. CTX-M Enzymes: Origin and Diffusion. Front. Microbiol. 3. https://doi.org/10.3389/fmicb.2012.00110

Carattoli, A., Zankari, E., García-Fernández, A., Larsen, M.V., Lund, O., Villa, L., Aarestrup, F.M., Hasman, H., 2014. In Silico Detection and Typing of Plasmids using PlasmidFinder and Plasmid Multilocus Sequence Typing. Antimicrobial Agents and Chemotherapy 58, 3895–3903. https://doi.org/10.1128/AAC.02412-14

Chaudhuri, R.R., Loman, N.J., Snyder, L.A.S., Bailey, C.M., Stekel, D.J., Pallen, M.J., 2008. xBASE2: a comprehensive resource for comparative bacterial genomics. Nucleic Acids Res. 36, D543-546. https://doi.org/10.1093/nar/gkm928

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

Chen, X., Zhang, Y., Zhang, Z., Zhao, Y., Sun, C., Yang, M., Wang, J., Liu, Q., Zhang, B., Chen, M., Yu, J., Wu, J., Jin, Z., Xiao, J., 2018b n. PGAweb: A Web Server for Bacterial Pan-Genome Analysis. Front. Microbiol. 9. https://doi.org/10.3389/fmicb.2018.01910

Cheng, L., Connor, T.R., Sirén, J., Aanensen, D.M., Corander, J., 2013. Hierarchical and Spatially Explicit Clustering of DNA Sequences with BAPS Software. Mol Biol Evol 30, 1224–1228. https://doi.org/10.1093/molbev/mst028

Chevreux, B., Pfisterer, T., Drescher, B., Driesel, A.J., Müller, W.E., Wetter, T. and Suhai, S., 2004. Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. Genome research, 14(6), pp.1147-1159.

Chewapreecha, C. (2014) 'Dense genomic sampling identifies highways of pneumococcal recombination', 46(3), pp. 305-9.

Clark, G., Paszkiewicz, K., Hale, J., Weston, V., Constantinidou, C., Penn, C., Achtman, M. and McNally, A., 2012. Genomic analysis uncovers a phenotypically diverse but genetically homogeneous Escherichia coli ST131 clone circulating in unrelated urinary tract infections. Journal of antimicrobial chemotherapy, 67(4), pp.868-877.

Compeau, P.E.C., Pevzner, P.A., Tesler, G., 2011. Why are de Bruijn graphs useful for genome assembly? Nat Biotechnol 29, 987–991. https://doi.org/10.1038/nbt.2023

Corander, J., Marttinen, P., Sirén, J., Tang, J., 2008. Enhanced Bayesian modelling in BAPS software for learning genetic structures of populations. BMC Bioinformatics 9, 539. https://doi.org/10.1186/1471-2105-9-539

Coughlan, T., Leder Mackley, K., Brown, M., Martindale, S., Schlögl, S., Mallaband, B., Arnott, J., Hoonhout, J., Szostak, D., Brewer, R. and Poole, E., 2013. Current issues and future directions in methods for studying technology in the home. PsychNology Journal, 11(2), pp.159-184.

Daubin, V., Gouy, M., Perrière, G., 2002. A Phylogenomic Approach to Bacterial Phylogeny: Evidence of a Core of Genes Sharing a Common History. Genome Res. 12, 1080–1090. https://doi.org/10.1101/gr.187002

de Goffau, M.C., Lager, S., Salter, S.J., Wagner, J., Kronbichler, A., Charnock-Jones, D.S., Peacock, S.J., Smith, G.C.S., Parkhill, J., 2018. Recognizing the reagent microbiome. Nature Microbiology 3, 851–853. https://doi.org/10.1038/s41564-018-0202-y

de Toro, M., Garcillán-Barcia, M.P. and de la Cruz, F., 2015. Plasmid diversity and adaptation analyzed by massive sequencing of Escherichia coli plasmids. In Plasmids: Biology and Impact in Biotechnology and Discovery (pp. 219-235). American Society of Microbiology.

Deamer, D., Akeson, M., Branton, D., 2016. Three decades of nanopore sequencing. Nature Biotechnology 34, 518–524. https://doi.org/10.1038/nbt.3423

DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., Angel, G. del, Rivas, M.A., Hanna, M., McKenna, A., Fennell, T.J., Kernytsky, A.M., Sivachenko, A.Y., Cibulskis, K., Gabriel, S.B., Altshuler, D., Daly, M.J., 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature Genetics 43, 491–498. https://doi.org/10.1038/ng.806

de Toro M, Garcillaon-Barcia MP, de la Cruz F. 2014. Plasmid diversity and adaptation analyzed by massive sequencing of Escherichia coli plasmids. Microbiol Spectr2:PLAS-0031-2014. https://doi.org/10.1128/microbiolspec.PLAS-0031-2014.

DiGuistini, S., Liao, N.Y., Platt, D., Robertson, G., Seidel, M., Chan, S.K., Docking, T.R., Birol, I., Holt, R.A., Hirst, M. and Mardis, E., 2009. De novo genome sequence assembly of a filamentous fungus using Sanger, 454 and Illumina sequence data. Genome biology, 10(9), p. R94.

Downing, T., 2015. Tackling drug resistant infection outbreaks of global pandemic Escherichia coli ST131 using evolutionary and epidemiological genomics. Microorganisms, 3(2), pp.236-267.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Ewels, P., Magnusson, M., Lundin, S. and Käller, M., 2016. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics, 32(19), pp.3047-3048.

Ewels, P., Magnusson, M., Lundin, S., Käller, M., 2016. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics 32, 3047–3048. https://doi.org/10.1093/bioinformatics/btw354

Ewing, B., Hillier, L., Wendl, M.C. and Green, P., 1998. Base-calling of automated sequencer traces usingPhred. I. Accuracy assessment. Genome research, 8(3), pp.175-185.

Eisenstein M. Oxford Nanopore announcement sets sequencing sector abuzz. 2012. Nat Biotechnol, 30:295–6.

FastQC: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Feil, E.J., Spratt, B.G., 2001. Recombination and the population structures of bacterial pathogens. Annu. Rev. Microbiol. 55, 561–590. https://doi.org/10.1146/annurev.micro.55.1.561

Flicek, P. and Birney, E., 2009. Sense from sequence reads: methods for alignment and assembly. Nature methods, 6, pp.S6-S12.

Floyd, R.V., Upton, M., Hultgren, S.J., Wray, S., Burdyga, T.V. and Winstanley, C., 2012. Escherichia coli–Mediated Impairment of Ureteric Contractility Is Uropathogenic E. coli Specific. The Journal of infectious diseases, 206(10), pp.1589-1596.

Forde, B.M., Ben Zakour, N.L., Stanton-Cook, M., Phan, M.-D., Totsika, M., Peters, K.M., Chan, K.G., Schembri, M.A., Upton, M., Beatson, S.A., 2014. The complete genome sequence of Escherichia coli EC958: a high quality reference sequence for the globally disseminated multidrug resistant E. coli O25b:H4-ST131 clone. PLoS ONE 9, e104400. https://doi.org/10.1371/journal.pone.0104400

Frost, L.S., Leplae, R., Summers, A.O. and Toussaint, A., 2005. Mobile genetic elements: the agents of open source evolution. Nature reviews. Microbiology, 3(9), p.722.

Fu, L., Niu, B., Zhu, Z., Wu, S., Li, W., 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28, 3150–3152. https://doi.org/10.1093/bioinformatics/bts565

Gardner, S.N. and Hall, B.G., 2013. When whole-genome alignments just won't work: kSNP v2 software for alignment-free SNP discovery and phylogenetics of hundreds of microbial genomes. PloS one, 8(12), p. e81760.

Gibson, D.G., 2011. Enzymatic assembly of overlapping DNA fragments. Meth. Enzymol. 498, 349–361. https://doi.org/10.1016/B978-0-12-385120-8.00015-2

Godwin, B.C., He, W., Helgesen, S., Ho, Chun Heen, Ho, Chun He, Irzyk, G.P., Jando, S.C., Alenquer, M.L.I., Jarvie, T.P., Jirage, K.B., Kim, J.-B., Knight, J.R., Lanza, J.R., Leamon, J.H., Goel, N., Singh, S. and Aseri, T.C., 2013. A review of soft computing techniques for gene prediction. ISRN Genomics, 2013.

Goodwin, S. Gurtowski J1, Ethe-Sayers S1, Deshpande P1, Schatz MC1, McCombie WR. 2015. Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome. Genome Research 25, 1750–1756.

Gordon, D., Green, P., 2013. Consed: a graphical editor for next-generation sequencing. Bioinformatics 29, 2936–2937. https://doi.org/10.1093/bioinformatics/btt515

Gurevich, A., Saveliev, V., Vyahhi, N., Tesler, G., 2013. QUAST: quality assessment tool for genome assemblies. Bioinformatics 29, 1072–1075. https://doi.org/10.1093/bioinformatics/btt086

Hanage, W.P., Fraser, C., Tang, J., Connor, T.R. and Corander, J., 2009. Hyper-recombination, diversity, and antibiotic resistance in pneumococcus. Science, 324(5933), pp.1454-1457.Griffiths, R. C. and Marjoram, P. (1996) 'Ancestral inference from samples of DNA sequences with recombination', J Comput Biol, 3(4), pp. 479-502.

Hargreaves, M.L., Shaw, K.M., Dobbins, G., Vagnone, P.M.S., Harper, J.E., Boxrud, D., Lynfield, R., Aziz, M., Price, L.B., Silverstein, K.A. and Danzeisen, J.L., 2015. Clonal dissemination of Enterobacter cloacae harboring blaKPC-3 in the upper midwestern United States. Antimicrobial agents and chemotherapy, 59(12), pp.7723-7734.

Head, S.R., Komori, H.K., LaMere, S.A., Whisenant, T., Van Nieuwerburgh, F., Salomon, D.R., Ordoukhanian, P., 2014. Library construction for next-generation sequencing:

Overviews and challenges. Biotechniques 56, 61-passim. https://doi.org/10.2144/000114133

Heydari, M., Miclotte, G., Demeester, P., Van de Peer, Y., Fostier, J., 2017. Evaluation of the impact of Illumina error correction tools on de novo genome assembly. BMC Bioinformatics 18. https://doi.org/10.1186/s12859-017-1784-8

Hinnebusch, J. and Tilly, K., 1993. Linear plasmids and chromosomes in bacteria. Molecular microbiology, 10(5), pp.917-922.

Hooper, D.C., 2001. Emerging mechanisms of fluoroquinolone resistance. Emerging infectious diseases, 7(2), p.337.

Hunt, M., Kikuchi, T., Sanders, M., Newbold, C., Berriman, M., Otto, T.D., 2013. REAPR: a universal tool for genome assembly evaluation. Genome Biol 14, R47. https://doi.org/10.1186/gb-2013-14-5-r47

Hunt, M., Newbold, C., Berriman, M. and Otto, T.D., 2014. A comprehensive evaluation of assembly scaffolding tools. Genome biology, 15(3), p.R42.

Hunt, M., Silva, N.D., Otto, T.D., Parkhill, J., Keane, J.A., Harris, S.R., 2015. Circlator: automated circularization of genome assemblies using long sequencing reads. Genome Biology 16. https://doi.org/10.1186/s13059-015-0849-0

Huson DH and Bryant D, 2006. Application of phylogenetic networks in evolutionary studies. Molecular Biology and Evolution, Volume 23, Issue 2, February 2006, Pages 254–267, https://doi.org/10.1093/molbev/msj030.

Hyatt, D., Chen, G.-L., Locascio, P.F., Land, M.L., Larimer, F.W., Hauser, L.J., 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11, 119. https://doi.org/10.1186/1471-2105-11-119

Iqbal, Z., Turner, I., McVean, G., 2013. High-throughput microbial population genomics using the Cortex variation assembler. Bioinformatics 29, 275–276. https://doi.org/10.1093/bioinformatics/bts673

Jadhav, S., Hussain, A., Devi, S., Kumar, A., Parveen, S., Gandham, N., Wieler, L.H., Ewers, C. and Ahmed, N., 2011. Virulence characteristics and genetic affinities of multiple drug resistant uropathogenic Escherichia coli from a semi urban locality in India. PloS one, 6(3), p.e18063.

Jayakumar, V., Sakakibara, Y., n.d. Comprehensive evaluation of non-hybrid genome assembly tools for third-generation PacBio long-read sequence data. Brief Bioinform. https://doi.org/10.1093/bib/bbx147

Johnson, J.R., Johnston, B., Clabots, C., Kuskowski, M.A. and Castanheira, M., 2010. Escherichia coli sequence type ST131 as the major cause of serious multidrug-resistant E. coli infections in the United States. Clinical infectious diseases, 51(3), pp.286-294.

Johnson, J.R., Tchesnokova, V., Johnston, B., Clabots, C., Roberts, P.L., Billig, M., Riddell, K., Rogers, P., Qin, X., Butler-Wu, S. and Price, L.B., 2013. Abrupt emergence of a single dominant multidrug-resistant strain of Escherichia coli. The Journal of infectious diseases, 207(6), pp.919-928.

Johnson, N.B., Hayes, L.D., Brown, K., Hoo, E.C. and Ethier, K.A., 2014. CDC National Health Report: leading causes of morbidity and mortality and associated behavioral risk and protective factors—United States, 2005–2013.

Juhas, M., van der Meer, J.R., Gaillard, M., Harding, R.M., Hood, D.W. and Crook, D.W., 2009. Genomic islands: tools of bacterial horizontal gene transfer and evolution. FEMS microbiology reviews, 33(2), pp.376-393.

Kim, Y., Koh, I., Lim, M.Y., Chung, W.-H., Rho, M., 2017. Pan-genome analysis of Bacillus for microbiome profiling. Scientific Reports 7, 10984. https://doi.org/10.1038/s41598-017-11385-9

Kloepper, T.H., Huson, D.H., 2008. Drawing explicit phylogenetic networks and their integration into SplitsTree. BMC Evol. Biol. 8, 22. https://doi.org/10.1186/1471-2148-8-22

Langmead, B., Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods 9, 357–359. https://doi.org/10.1038/nmeth.1923

Langmead, B., Trapnell, C., Pop, M., Salzberg, S.L., 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology 10, R25. https://doi.org/10.1186/gb-2009-10-3-r25

Lanza, V.F., de Toro, M., Garcillán-Barcia, M.P., Mora, A., Blanco, J., Coque, T.M. and de la Cruz, F., 2014. Plasmid flux in Escherichia coli ST131 sublineages, analyzed by plasmid constellation network (PLACNET), a new method for plasmid reconstruction from whole genome sequences. PLoS genetics, 10(12), p. e1004766.

Lapierre, P., Gogarten, J.P., 2009. Estimating the size of the bacterial pan-genome. Trends in Genetics 25, 107–110. https://doi.org/10.1016/j.tig.2008.12.004

Laslett, D., Canback, B., 2004. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. Nucleic Acids Res. 32, 11–16. https://doi.org/10.1093/nar/gkh152

Lawson, D.J., 2015. Populations in Statistical Genetic Modelling and Inference, in: Kreager, P., Winney, B., Ulijaszek, S., Capelli, C. (Eds.), Population in the Human Sciences. Oxford University Press, pp. 108–130. https://doi.org/10.1093/acprof:oso/9780199688203.003.0004

Lee, W.-P., Stromberg, M.P., Ward, A., Stewart, C., Garrison, E.P., Marth, G.T., 2014. MOSAIK: A Hash-Based Algorithm for Accurate Next-Generation Sequencing Short-Read Mapping. PLOS ONE 9, e90581. https://doi.org/10.1371/journal.pone.0090581

Lees, J.A., Harris, S.R., Tonkin-Hill, G., Gladstone, R.A., Lo, S.W., Weiser, J.N., Corander, J., Bentley, S.D., Croucher, N.J., 2018. Fast and flexible bacterial genomic epidemiology with PopPUNK. https://doi.org/10.1101/360917

Lefkowitz, S.M., Lei, M., Li, J., Lohman, K.L., Lu, H., Makhijani, V.B., McDade, K.E., McKenna, M.P., Myers, E.W., Nickerson, E., Nobile, J.R., Plant, R., Puc, B.P., Ronan, M.T., Roth, G.T., Sarkis, G.J., Simons, J.F., Simpson, J.W., Srinivasan, M., Tartaro, K.R., Tomasz, A., Vogt, K.A., Volkmer, G.A., Wang, S.H., Wang, Y., Weiner, M.P., Yu, P., Begley, R.F., Rothberg, J.M., 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature 437, 376–380. https://doi.org/10.1038/nature03959

Levene, M.J., Korlach, J., Turner, S.W., Foquet, M., Craighead, H.G., Webb, W.W., 2003. Zero-Mode Waveguides for Single-Molecule Analysis at High Concentrations. Science 299, 682–686. https://doi.org/10.1126/science.1079700

Li, H., 2016. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. Bioinformatics 32, 2103–2110. https://doi.org/10.1093/bioinformatics/btw152

Li, H., Durbin, R., 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25, 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, Z., Chen, Y., Mu, D., Yuan, J., Shi, Y., Zhang, H., Gan, J., Li, N., Hu, X., Liu, B. and Yang, B., 2012. Comparison of the two major classes of assembly algorithms: overlap–

layout–consensus and de-bruijn-graph. Briefings in functional genomics, 11(1), pp.25-37.

Ludden, C. 2014. The role of long-term care facilities in the dissemination of antimicrobial resistance. Unpublished doctoral dissertation. National University of Ireland Galway, Ireland.

Ludden, C., Cormican, M., Vellinga, A., Johnson, J.R., Austin, B. and Morris, D., 2015. Colonisation with ESBL-producing and carbapenemase-producing Enterobacteriaceae, vancomycin-resistant enterococci, and meticillin-resistant Staphylococcus aureus in a long-term care facility over one year. BMC infectious diseases, 15(1), p.168.

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., He, G., Chen, Y., Pan, Q., Liu, Yunjie, Tang, J., Wu, G., Zhang, H., Shi, Y., Liu, Yong, Yu, C., Wang, B., Lu, Y., Han, C., Cheung, D.W., Yiu, S.-M., Peng, S., Xiaoqian, Z., Liu, G., Liao, X., Li, Y., Yang, H., Wang, Jian, Lam, T.-W., Wang, Jun, 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience 1, 18. https://doi.org/10.1186/2047-217X-1-18

Magoc, T., Pabinger, S., Canzar, S., Liu, X., Su, Q., Puiu, D., Tallon, L.J. and Salzberg, S.L., 2013. GAGE-B: an evaluation of genome assemblers for bacterial organisms. Bioinformatics, 29(14), pp.1718-1725.

Margulies, M., M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, J. Berka, M. S. Braverman, Y. J. Chen, Z. Chen, S. B. Dewell, L. Du, J. M. Fierro, X. V. Gomes, B. C. Godwin, W. He, S. Helgesen, C. H. Ho, G. P. Irzyk, S. C. Jando, M. L. Alenquer, T. P. Jarvie, K. B. Jirage, J. B. Kim, J. R. Knight, J. R. Lanza, J. H. Leamon, S. M. Lefkowitz, M. Lei, J. Li, K. L. Lohman, H. Lu, V. B. Makhijani, K. E. McDade, M. P. McKenna, E. W. Myers, E. Nickerson, J. R. Nobile, R. Plant, B. P. Puc, M. T. Ronan, G. T. Roth, G. J. Sarkis, J. F. Simons, J. W. Simpson, M. Srinivasan, K. R. Tartaro, A. Tomasz, K. A. Vogt, G. A. Volkmer, S. H. Wang, Y. Wang, M. P. Weiner, P. Yu, R. F. Begley, and J. M. Rothberg. 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature 437:376-380.

Marttinen, P., Hanage, W.P., Croucher, N.J., Connor, T.R., Harris, S.R., Bentley, S.D. and Corander, J., 2011. Detection of recombination events in bacterial genomes from large population samples. Nucleic acids research, 40(1), pp.e6-e6.

Matsumura, Y., Yamamoto, M., Higuchi, T., Komori, T., Tsuboi, F., Hayashi, A., Sugimoto, Y., Hotta, G., Matsushima, A., Nagao, M. and Takakura, S., 2012. Prevalence of plasmid-mediated AmpC β-lactamase-producing Escherichia coli and spread of the ST131 clone among extended-spectrum β-lactamase-producing E. coli in Japan. International journal of antimicrobial agents, 40(2), pp.158-162.

McNally, A., Oren, Y., Kelly, D., Pascoe, B., Dunn, S., Sreecharan, T., Vehkala, M., Välimäki, N., Prentice, M.B., Ashour, A. and Avram, O., 2016. Combined analysis of variation in core, accessory and regulatory genome regions provides a super-resolution view into the evolution of bacterial populations. PLoS genetics, 12(9), p. e1006280.

Medvedev, P., Stanciu, M. and Brudno, M., 2009. Computational methods for discovering structural variation with next-generation sequencing. Nature methods, 6, pp.S13-S20.

Miller, J.R., Delcher, A.L., Koren, S., Venter, E., Walenz, B.P., Brownley, A., Johnson, J., Li, K., Mobarry, C. and Sutton, G., 2008. Aggressive assembly of pyrosequencing reads with mates. Bioinformatics, 24(24), pp.2818-2824.

Mikheyev, A.S. & Tin, M.M. 2014. A first look at the Oxford Nanopore MinION sequencer. Mol. Ecol. Resour. 14, 1097–1102.

Mir, K.U. and Southern, E.M., 2000. Sequence variation in genes and genomic DNA: methods for large-scale analysis. Annual review of genomics and human genetics, 1(1), pp.329-360.
monomorphic bacterial pathogens. Annu Rev Microbiol, 62, pp. 53-70.

Myers, E.W., 2005. The fragment assembly string graph. Bioinformatics, 21(suppl_2), pp.ii79-ii85.

Namiki, T., Hachiya, T., Tanaka, H. and Sakakibara, Y., 2012. MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. Nucleic acids research, 40(20), pp.e155-e155.

Nicolas-Chanoine, M.-H., Bertrand, X., Madec, J.-Y., 2014. Escherichia coli ST131, an Intriguing Clonal Group. Clin Microbiol Rev 27, 543–574. https://doi.org/10.1128/CMR.00125-13

Nicolas-Chanoine, M.H., Blanco, J., Leflon-Guibout, V., Demarty, R., Alonso, M.P., Caniça, M.M., Park, Y.J., Lavigne, J.P., Pitout, J. and Johnson, J.R., 2008. Intercontinental emergence of Escherichia coli clone O25: H4-ST131 producing CTX-M-15. Journal of Antimicrobial Chemotherapy, 61(2), pp.273-281.

Nicolas-Chanoine, M.H., Robert, J., Vigan, M., Laouénan, C., Brisse, S., Mentré, F. and Jarlier, V., 2013. Different factors associated with CTX-M-producing ST131 and non-ST131 Escherichia coli clinical isolates. PLoS One, 8(9), p.e72191.

Nielsen, R., Paul, J.S., Albrechtsen, A. and Song, Y.S., 2011. Genotype and SNP calling from next-generation sequencing data. Nature Reviews Genetics, 12(6), pp.443-451.
Nikolenko, S.I., Korobeynikov, A.I., Alekseyev, M.A., 2013. BayesHammer: Bayesian clustering for error correction in single-cell sequencing. BMC Genomics 14, S7. https://doi.org/10.1186/1471-2164-14-S1-S7

Ning, Z., Cox, A.J., Mullikin, J.C., 2001. SSAHA: A Fast Search Method for Large DNA
Nishida, H., 2012. Evolution of genome base composition and genome size in bacteria. Frontiers in microbiology, 3.

Nishito, Y., Osana, Y., Hachiya, T., Popendorf, K., Toyoda, A., Fujiyama, A., Itaya, M. and Sakakibara, Y., 2010. Whole genome assembly of a natto production strain Bacillus subtilis natto from very short read data. BMC genomics, 11(1), p.243.

Olesen, B., Hansen, D.S., Nilsson, F., Frimodt-Møller, J., Leihof, R.F., Struve, C., Scheutz, F., Johnston, B., Krogfelt, K.A. and Johnson, J.R., 2013. Prevalence and characteristics of the epidemic multi-resistant Escherichia coli ST131 clonal group among extended-spectrum β-lactamase (ESBL)-producing E. coli in Copenhagen. Journal of clinical microbiology, pp.JCM-00346.

Page, A.J., Cummins, C.A., Hunt, M., Wong, V.K., Reuter, S., Holden, M.T.G., Fookes, M., Falush, D., Keane, J.A., Parkhill, J., 2015. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics 31, 3691–3693. https://doi.org/10.1093/bioinformatics/btv421

Page, A.J., De Silva, N., Hunt, M., Quail, M.A., Parkhill, J., Harris, S.R., Otto, T.D. and Keane, J.A., 2016. Robust high-throughput prokaryote de novo assembly and improvement pipeline for Illumina data. Microbial genomics, 2(8).

Pell, J., Hintze, A., Canino-Koning, R., Howe, A., Tiedje, J.M. and Brown, C.T., 2012. Scaling metagenome sequence assembly with probabilistic de Bruijn graphs. Proceedings of the National Academy of Sciences, 109(33), pp.13272-13277.

Peng, Y., Leung, H.C.M., Yiu, S.M., Chin, F.Y.L., 2012. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics 28, 1420–1428. https://doi.org/10.1093/bioinformatics/bts174

Petty, Nicola K., et al. "Global dissemination of a multidrug resistant Escherichia coli clone." Proceedings of the National Academy of Sciences111.15 (2014): 5694-5699.

Phan, M.D., Forde, B.M., Peters, K.M., Sarkar, S., Hancock, S., Stanton-Cook, M., Zakour, N.L.B., Upton, M., Beatson, S.A. and Schembri, M.A., 2015. Molecular characterization of a multidrug resistance IncF plasmid from the globally disseminated Escherichia coli ST131 clone. PLoS One, 10(4), p.e0122369.

Pitout, J.D. and DeVinney, R., 2017. Escherichia coli ST131: a multidrug-resistant clone primed for global domination. F1000Research, 6.

Pitout, J.D. and Laupland, K.B., 2008. Extended-spectrum β-lactamase-producing Enterobacteriaceae: an emerging public-health concern. The Lancet infectious diseases, 8(3), pp.159-166.

Pollard, M.O., Gurdasani, D., Mentzer, A.J., Porter, T., Sandhu, M.S., 2018. Long reads: their purpose and place. Hum Mol Genet 27, R234–R241. https://doi.org/10.1093/hmg/ddy177

Ponstingl, H., Ning, Z., Ponstingl, H., Ning, Z., 2010. SMALT – A new mapper for DNA sequencing reads . F1000Research 1. https://doi.org/10.7490/f1000research.327.1

Pop, M., 2004. Shotgun sequence assembly. Advances in computers, 60, pp.193-248.

Price, L.B., Johnson, J.R., Aziz, M., Clabots, C., Johnston, B., Tchesnokova, V., Nordstrom, L., Billig, M., Chattopadhyay, S., Stegger, M. and Andersen, P.S., 2013. The epidemic of extended-spectrum-β-lactamase-producing Escherichia coli ST131 is driven by a single highly pathogenic subclone, H30-Rx. MBio, 4(6), pp.e00377-13.

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.

Pritchard, J.K., Stephens, M., Donnelly, P., 2000. Inference of population structure using multilocus genotype data. Genetics 155, 945–959.

Rahman, A. and Pachter, L., 2013. CGAL: computing genome assembly likelihoods. Genome biology, 14(1), p.R8.

Pupko T, Pe'er I, Shamir R, Graur D. 2000. A fast algorithm for joint reconstruction of ancestral amino acid sequences. Mol Biol Evol 17: 890–896.

Rimmer, A., Phan, H., Mathieson, I., Iqbal, Z., Twigg, S.R.F., Consortium, W., Wilkie, A.O.M., McVean, G., Lunter, G., 2014. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. Nature Genetics 46, 912–918. https://doi.org/10.1038/ng.3036

Rogers, B.A., Sidjabat, H.E., Paterson, D.L., 2011. Escherichia coli O25b-ST131: a pandemic, multiresistant, community-associated strain. Journal of Antimicrobial Chemotherapy 66, 1–14. https://doi.org/10.1093/jac/dkq415

Ronen, R., Boucher, C., Chitsaz, H., Pevzner, P., 2012. SEQuel: improving the accuracy of genome assemblies. Bioinformatics 28, i188–i196. https://doi.org/10.1093/bioinformatics/bts219

Rouli, L., Merhej, V., Fournier, P.-E., Raoult, D., 2015. The bacterial pangenome as a new tool for analysing pathogenic bacteria. New Microbes and New Infections 7, 72–85. https://doi.org/10.1016/j.nmni.2015.06.005

Ruiz, J., 2003. Mechanisms of resistance to quinolones: target alterations, decreased accumulation and DNA gyrase protection. Journal of Antimicrobial Chemotherapy, 51(5), pp.1109-1117.

Salipante, S.J., Roach, D.J., Kitzman, J.O., Snyder, M.W., Stackhouse, B., Butler-Wu, S.M., Lee, C., Cookson, B.T. and Shendure, J., 2015. Large-scale genomic sequencing of extraintestinal pathogenic Escherichia coli strains. Genome research, 25(1), pp.119-128.

Sandmann, S., Graaf, A.O. de, Karimi, M., Reijden, B.A. van der, Hellström-Lindberg, E., Jansen, J.H., Dugas, M., 2017. Evaluating Variant Calling Tools for Non-Matched Next-Generation Sequencing Data. Scientific Reports 7, 43169. https://doi.org/10.1038/srep43169

Sanger, F. & Coulson, A. R. 1975. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J. Mol. Biol. 94, 441±448.

Sanger, F., Nicklen, S., Coulson, A. 1977. DNAsequencing by chain terminating inhibitors. Proc. Natl. Acad. Sci. USA 74, 5463-5467.

Schbath, S., Martin, V., Zytnicki, M., Fayolle, J., Loux, V., Gibrat, J.-F., 2012. Mapping Reads on a Genomic Sequence: An Algorithmic Overview and a Practical Comparative Analysis. J Comput Biol 19, 796–813. https://doi.org/10.1089/cmb.2012.0022

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

Segerman, B., 2012. The genetic integrity of bacterial species: the core genome and the accessory genome, two different stories. Frontiers in cellular and infection microbiology, 2.

Selander, R.K., Caugant, D.A., Ochman, H., Musser, J.M., Gilmour, M.N. and Whittam, T.S., 1986. Methods of multilocus enzyme electrophoresis for bacterial population genetics and systematics. Applied and environmental microbiology, 51(5), p.873.

Shintani, M., Sanchez, Z.K., Kimbara, K., 2015. Genomics of microbial plasmids: classification and identification based on replication and transfer systems and host taxonomy. Front Microbiol 6, 242. https://doi.org/10.3389/fmicb.2015.00242

Sidjabat, H.E., Townell, N., Nimmo, G.R., George, N.M., Robson, J., Vohra, R., Davis, L., Heney, C. and Paterson, D.L., 2015. Dominance of IMP-4-producing Enterobacter cloacae among carbapenemase-producing Enterobacteriaceae in Australia. Antimicrobial agents and chemotherapy, 59(7), pp.4059-4066.

Simpson, J.T. and Durbin, R., 2012. Efficient de novo assembly of large genomes using compressed data structures. Genome research, 22(3), pp.549-556.

Simpson, J.T., Wong, K., Jackman, S.D., Schein, J.E., Jones, S.J. and Birol, I., 2009. ABySS: a parallel assembler for short read sequence data. Genome research, 19(6), pp.1117-1123.

Smith, J.M., 1992. Analyzing the mosaic structure of genes. Journal of molecular evolution, 34(2), pp.126-129.

Sneath, P.H., 1975. Cladistic representation of reticulate evolution. Systematic Zoology, 24(3), pp.360-368.

Spratt, B.G., 2004. Exploring the concept of clonality in bacteria. Methods Mol. Biol. 266,

Stamatakis, A., 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics, 22(21), pp.2688-2690.OBJ

Stoesser, N., Batty, E.M., Eyre, D.W., Morgan, M., Wyllie, D.H., Del Ojo Elias, C., Johnson, J.R., Walker, A.S., Peto, T.E.A. and Crook, D.W., 2013. Predicting antimicrobial susceptibilities for Escherichia coli and Klebsiella pneumoniae isolates using whole genomic sequence data. Journal of Antimicrobial Chemotherapy, 68(10), pp.2234-2244.

Stoesser, N., Sheppard, A.E., Pankhurst, L., De Maio, N., Moore, C.E., Sebra, R., Turner, P., Anson, L.W., Kasarskis, A., Batty, E.M. and Kos, V., 2016. Evolutionary history of the global emergence of the Escherichia coli epidemic clone ST131. MBio, 7(2), pp.e02162-15.

Swain, M.T., Tsai, I.J., Assefa, S.A., Newbold, C., Berriman, M., Otto, T.D., 2012. A Post-assembly genome-improvement toolkit (PAGIT) to obtain annotated genomes from contigs. Nat Protoc 7, 1260–1284. https://doi.org/10.1038/nprot.2012.068

Tettelin, H., Riley, D., Cattuto, C., Medini, D., 2008. Comparative genomics: the bacterial pan-genome. Current Opinion in Microbiology 11, 472–477. https://doi.org/10.1016/j.mib.2008.09.006

Thompson, J. F. & Milos, P. M. 2011. The properties and applications of single-molecule DNA sequencing. Genome Biology 12, 217.

Thorpe, H.A., Bayliss, S.C., Sheppard, S.K., Feil, E.J., 2018. Piggy: a rapid, large-scale pan-genome analysis tool for intergenic regions in bacteria. Gigascience 7. https://doi.org/10.1093/gigascience/giy015

Tonkin-Hill, G., Lees, J.A., Bentley, S.D., Frost, S.D.W., Corander, J., 2018. Fast Hierarchical Bayesian Analysis of Population Structure. bioRxiv 454355. https://doi.org/10.1101/454355

Totsika, M., Beatson, S.A., Sarkar, S., Phan, M.D., Petty, N.K., Bachmann, N., Szubert, M., Sidjabat, H.E., Paterson, D.L., Upton, M. and Schembri, M.A., 2011. Insights into a multidrug resistant Escherichia coli pathogen of the globally disseminated ST131 lineage: genome analysis and virulence mechanisms. PloS one, 6(10), p.e26578.

Totsika, M., Gomes Moriel, D., Idris, A., A Rogers, B., J Wurpel, D., Phan, M.D., L Paterson, D. and A Schembri, M., 2012. Uropathogenic Escherichia coli mediated urinary tract infection. Current drug targets, 13(11), pp.1386-1399.

Totsika, M., Kostakioti, M., Hannan, T.J., Upton, M., Beatson, S.A., Janetka, J.W., Hultgren, S.J., Schembri, M.A., 2013. A FimH inhibitor prevents acute bladder infection and treats chronic cystitis caused by multidrug-resistant uropathogenic Escherichia coli ST131. J. Infect. Dis. 208, 921–928. https://doi.org/10.1093/infdis/jit245.

Toh H, Oshima K, Toyoda A, Ogura Y, Ooka T, Sasamoto H, Park SH, Iyoda S, Kurokawa K, Morita H, Itoh K, Taylor TD, Hayashi T, Hattori M. Complete genome sequence of the wild-type commensal Escherichia coli strain SE15, belonging to phylogenetic group B2. J Bacteriol. 2010 192(4):1165-6. doi: 10.1128/JB.01543-09

Travers, K. J., Chin, C.-S., Rank, D. R., Eid, J. S. & Turner, S. W. 2010. A flexible and efficient template format for circular consensus sequencing and SNP detection. Nucleic Acids Research 38, e159–e159.

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

Vezzi, F., Narzisi, G. and Mishra, B., 2012. Feature-by-feature–evaluating de novo sequence assembly. PloS one, 7(2), p.e31002.

Vos, M. and Didelot, X., 2009. A comparison of homologous recombination rates in bacteria and archaea. The ISME journal, 3(2), p.199.

Vos, M., Hesselman, M.C., te Beek, T.A., van Passel, M.W.J., Eyre-Walker, A., 2015. Rates of Lateral Gene Transfer in Prokaryotes: High but Why? Trends in Microbiology 23, 598–605. https://doi.org/10.1016/j.tim.2015.07.006

Wailan, A.M., Coll, F., Heinz, E., Tonkin-Hill, G., Corrander, J., Feasey, N.A., Thomson, N.R., 2018. rPinecone: Define sub-lineages of a clonal expansion via a phylogenetic tree. bioRxiv 404624. https://doi.org/10.1101/404624

Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C.A., Zeng, Q., Wortman, J., Young, S.K., Earl, A.M., 2014. Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. PLOS ONE 9, e112963. https://doi.org/10.1371/journal.pone.0112963

Wang, J., Liu, G., Munge, B., Lin, L., Zhu, Q., 2004. DNA-based amplified bioelectronic detection and coding of proteins. Angew. Chem. Int. Ed. Engl. 43, 2158–2161. https://doi.org/10.1002/anie.200453832

Warnes, S.L., Highmore, C.J. and Keevil, C.W., 2012. Horizontal transfer of antibiotic resistance genes on abiotic touch surfaces: implications for public health. MBio, 3(6), pp.e00489-12.

Watson, M., 2018. Mind the gaps - ignoring errors in long read assemblies critically affects protein prediction. bioRxiv 285049. https://doi.org/10.1101/285049

Wick, R.R., Judd, L.M., Gorrie, C.L., Holt, K.E., 2017. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. PLOS Computational Biology 13, e1005595. https://doi.org/10.1371/journal.pcbi.1005595

Wilm, A., Aw, P.P.K., Bertrand, D., Yeo, G.H.T., Ong, S.H., Wong, C.H., Khor, C.C., Petric, R., Hibberd, M.L., Nagarajan, N., 2012. LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. Nucleic Acids Res 40, 11189–11201. https://doi.org/10.1093/nar/gks918

Woodford, N., Carattoli, A., Karisik, E., Underwood, A., Ellington, M.J. and Livermore, D.M., 2009. Complete nucleotide sequences of plasmids pEK204, pEK499, and pEK516, encoding CTX-M enzymes in three major Escherichia coli lineages from the United Kingdom, all belonging to the international O25: H4-ST131 clone. Antimicrobial agents and chemotherapy, 53(10), pp.4472-4482.

Woodford, N., Turton, J.F. and Livermore, D.M., 2011. Multiresistant Gram-negative bacteria: the role of high-risk clones in the dissemination of antibiotic resistance. FEMS microbiology reviews, 35(5), pp.736-755.

Zerbino, D.R., Birney, E., 2008. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18, 821–829. https://doi.org/10.1101/gr.074492.107

Zhang, Y., Sievert, S.M., 2014. Pan-genome analyses identify lineage- and niche-specific markers of evolution and adaptation in Epsilonproteobacteria. Front. Microbiol. 5. https://doi.org/10.3389/fmicb.2014.00110

Zhong, Y.M., Liu, W.E., Liang, X.H., Li, Y.M., Jian, Z.J. and Hawkey, P.M., 2015. Emergence and spread of O16-ST131 and O25b-ST131 clones among faecal CTX-M-producing Escherichia coli in healthy individuals in Hunan Province, China. Journal of Antimicrobial Chemotherapy, 70(8), pp.2223-2227.

Zimin, A.V., Marçais, G., Puiu, D., Roberts, M., Salzberg, S.L. and Yorke, J.A., 2013. The MaSuRCA genome assembler. Bioinformatics, 29(21), pp.2669-2677.

Chapter 2: Recombination Analysis of N=100 E. coli ST131 Isolates from Long-term Care Facilities in Ireland/UK

Abstract

Increasing rates of morbidity and mortality cased by complications with bacteremia or sepsis affect hundreds of millions of human populations annually. Extra-intestinal pathogenic *Escherichia coli* (ExPEC) commonly cause bloodstream and urinary tract infections. Antimicrobial resistance (AMR) has also developed among ExPEC, most particularly in the globally disseminated *E. coli* sequence type 131 clone. Several methods including phylogenetic reconstruction may help in resolving ST131's evolutionary history. However, it could be challenging as the clone, much like other bacteria, undergoes several instances of horizontal gene transfer and homologous recombination. This chapter extends the pilot study performed in Chapter 1 and focused on developing the methods for recombination analysis in ST131 population. Here, we detected and analysed recombination events in N=100 ST131 isolates by comparing and contrasting three recombination detection software: BRATNextGen, Gubbins and ClonalframeML. We evaluated the performance of each software in estimating the relative recombination rates and other parameters that measure evolutionary processes in each strain.

2.1 Introduction

Increasing rates of morbidity and mortality caused by complications of bacteremia or sepsis affect hundreds of millions of human populations annually (Russo and Johnson 2003). Gram-negative bacterium associated with bloodstream and urinary tract infections are commonly caused by Extra-intestinal pathogenic *Escherichia coli* (ExPEC) (Pitout 2012). Antimicrobial resistance (AMR) in these ExPEC pathogens is also prevalent making matters worse. Resistance to first to last-line antibiotics has steadily increased due to the production of extended-spectrum β-lactamases (ESBL) by *E. coli* isolates particularly in a specific globally disseminated clone, *E. coli* sequence type 131 (ST131) (Jones et al. 1994; Turnidge J, 1996; Hummers-Pradier et al. 2005; Coque and Canton, 2008; Nicolas-Chanoine et al. 2008; Coque et al. 2008; Johnson et al. 2010). Indeed, ST131 is now known to be the most important cause of community-acquired and nosocomial infections. Deep genome sequencing and analysis are required to understand the origin, evolution and spread of ST131 (Downing 2015) and ultimately curb the infections caused by this clone.

There are three mechanisms that brought about homologous recombination in bacteria: transduction, where a virus (bacteriophage) transfers DNA segment/s from the donor to the recipient, transformation where the donor DNA is freely taken up by the recipient from the environment, and conjugation where donor and recipient come into direct contact via the bacterial pilus (Vos and Didelot, 2009). Accurate inference of phylogenetic relationships within the ST131 clone therefore requires correctly detecting and accounting for recombination events (Didelot and Wilson 2015).

In this chapter, we determined the population structure and inferred the genealogical history of N=100 ST131 collection using various approaches for phylogenomic analysis. This ST131 population includes 90 strains from long-term care facilities (LTCFs) and hospitals from Ireland and 10 community/hospital samples from the UK. We further detected and analysed recombination events in this population by comparing and contrasting three recombination detection software: BRATNextGen (Bayesian Recombination Tracker Next Generation; Marttinen et al. 2011), Gubbins (Genealogies Unbiased By recomBinations In Nucleotide Sequences; Croucher et al. 2014) and

ClonalframeML (Didelot and Wilson 2015) that classify SNPs as recombining or not. We sought to determine whether these tools yield consistent and accurate results, or which one will yield the most informative set of recombination parameter values. We evaluated the performance of each software in estimating the relative recombination rates and other parameters that measure evolutionary processes in each strain. This chapter extends the pilot study performed in Chapter 1 and focused on developing the methods for recombination analysis in ST131 population.

2.2 Materials and Methods

2.2.1 Data sources

Data used in this chapter were taken from the genomic characterization of those from the pilot study (Ludden 2014) described in Chapter 1. A collection of N=100 ST131 samples was used in the recombination analysis; n=90 of these ST131 were ESBL-producing isolates. The initial study was conducted to identify the baseline prevalence of colonization, monitor the colonization status at quarterly intervals, assess the risk factors associated with colonization and finally, characterise profiles for antimicrobial susceptibility of certain antibiotic resistant bacteria in LTCFs. The other n=10 were taken from Clark et al. (2012).

2.2.2 Implementing BRATNextGen, ClonalframeML and Gubbins

BRATNextGen (Marttinen et al. 2011) was implemented on a desktop as follows: the PSA tree was drawn from the aligned FASTA file containing 4,039 polymorphic sites. This step is followed by clustering, learning the recombination value and setting the hyperparameter, α indicates the amount of expected variation in a particular cluster and highly influences the number of clusters that will be generated in the process.

Statistical significance for this run was estimated by creating 100 replicates.
The tabular outputs of BRATNextGen contain the length of each detected recombinant segment with their positions in the genome, the distance (d) between the root of the tree

to any given leaf (representing the depth of phylogenetic heterogeneity in the branch) and the number of HGT events and their origins.

Gubbins (Croucher et al. 2014) was initially ran in ICHEC server using modified settings: run_gubbins.py ALL.test.fasta -s RaxML_bootstrap.ALLtest.out.tre –u where the starting tree (-s) and timestamp (-u) were specified. Results of the run were then recorded in a set of output files. A python script in the package was used to generate a figure for visualizing the distribution of SNPs brought about by recombination I versus those that have arisen from point mutations (m). Recombinant tracts that were found in the same genomic location (with the same start-end coordinated) are considered "common" among the results of the recombination analysis platforms tested here.

Assessment of the historical and recent recombination events using ClonalframeML (Didelot and Wilson 2015) was done by running the standard and per-branch models (respectively) using the commands below. The parameter kappa (transition/transversion ratio scaled by base frequencies) was initially calculated with RaxML.

The transition:traversion ratio (indicated by the parameter kappa) was initially estimated to be 1.93 for all 4,039 SNPs (ALL) following phylogenetic reconstruction by the latest version of RAxML (tree output: RAxM_bootstrap.ALLtest.out.tre). The option –em true directs the program to estimate the recombination parameters using a Baum-Welch expectation maximization (EM) algorithm such that the parameters are shared by all branches. This option is replaced by –embranch true for getting the estimates of the recombination events in the outer branches; the -embranch_dispersion value is set to 0.1 to 1.0 to check if any difference in the results will be observed. This option indicates the constraint on the changes of recombination parameters among the tree branches and is scaled from 0-1, with 0 being the most constrained (least dispersed). Correlation coefficients and Cohen's d were computed to determine the relationship between and the effect of the dispersion values and/to both the delta and the recombination frequencies, respectively. The –emsim 100 estimates uncertainty in the EM algorithm requests for 100 pseudobootstrap replicates. The results of the run are then recorded in a log file (i.e ALL.log.txt). Detected recombinant blocks in each internal and external branch were

drawn by running the R script in the package. The script was modified to change the background colour to white instead of the standard skyblue.

2.3 Results

2.3.1 ClonalframeML had higher sensitivity in detecting recombinant SNPs than BRATNextGen and Gubbins

The patterns of base substitutions caused by recombination and point mutations in the genomes of 100 ST131 isolates were analysed using BRATNextGen, ClonalframeML and Gubbins. These tools were tested for accurate and consistent values for recombination parameters. Supplementary Table S2.1A and B present the key features of each algorithm in estimating genetic variations in a sample population; while all three can detect the donated regions to the core genome of an isolate.

Graphical representations of recombinant regions detected by the three software are shown in Figure 2.1. BRATNextGen generated a PSA tree which highlights clusters with common recombination events and show a total of 154 recombinogenic tracts in the isolates. The colours of the detected segments indicate the cluster in which the segment is most prevalent. Two major clusters were formed after the run. The cyan-coloured tracts, for example, are shared by more taxa in the upper cluster while the red tracts are more abundant on the lower clade; grey bars are the missing SNPs in the analysis (Figure 2.1). ClonalframeML identified 222 recombinant segments throughout the reconstructed phylogeny of the population: 152 are products of recent recombination events and 70 were introduced in the ancestral branches. These tracts are shown in red while true events were shown in blue (Figure 2.2). Using the ML tree by RAxML v7.2.8, Gubbins generated a reconstructed phylogeny showing panels relating to 195 predicted laterally transferred genetic segments (165 in the external branches while 30 in the internal ones). Each column relates to a base in the reference genome while each row corresponds to a branch in the ML tree. Predicted recombinations that occur on an internal branch shared by multiple taxa through common ancestry are shown in red blocks. The blue blocks, on the other hand, represent recombination events on terminal branches and are unique to individual samples (Figure 2.3).

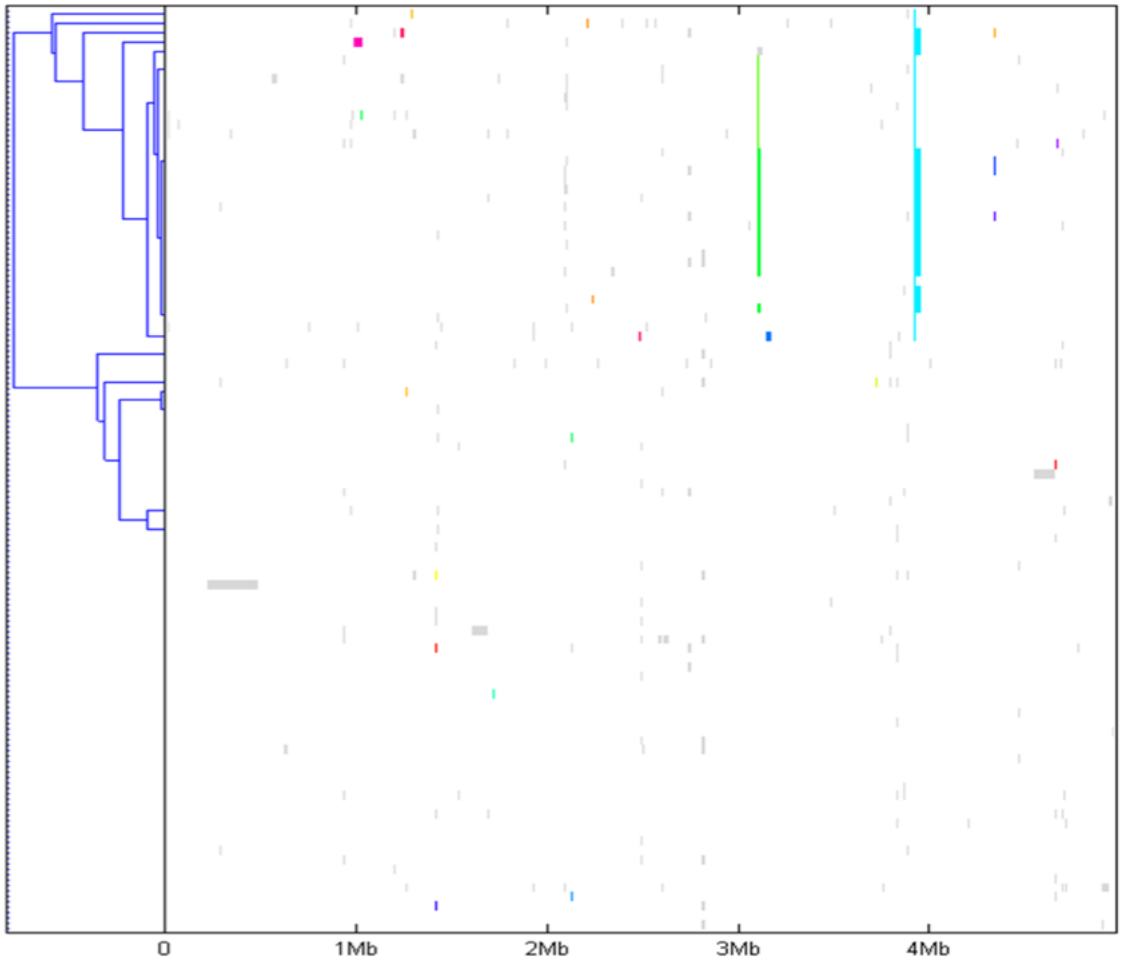


Figure 2.1. Results for recombination analysis of the 100 E. coli ST131 samples using BRATNextGEn. The PSA treeis shown on the left. The tree is cut at threshold of 0.1 to produce 7 clusters. On the right, the horizontal colored bars show the recombination events for each isolate. The colors of the detected segments indicate the cluster in which the segment is most prevalent. Grey bars show missing SNPs.

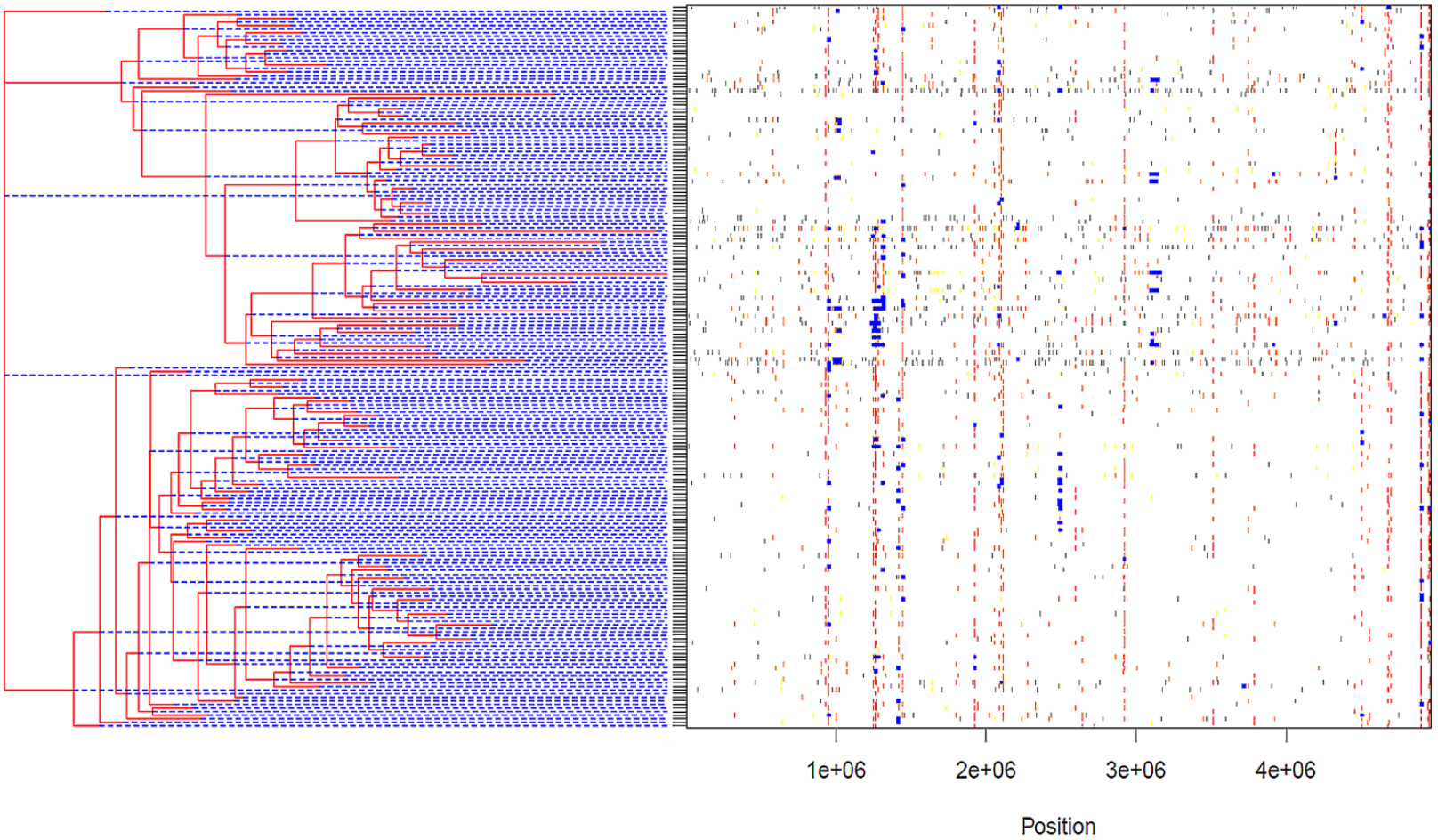


Figure 2.2. Lateral gene transfer assessed by ClonalframeML using the maximum likelihood phylogeny generated from the whole genome alignment of 100 isolates. The panels represent the pattern of predicted recombinations from the analyses using RAxML phylogenetic tree reconstruction. Each column relates to a base in the reference genome; each row represents an isolate in the phylogeny. True events are shown in blue and segments detected by ClonalFrameML are shown in red. Grey bars show missing SNPs.

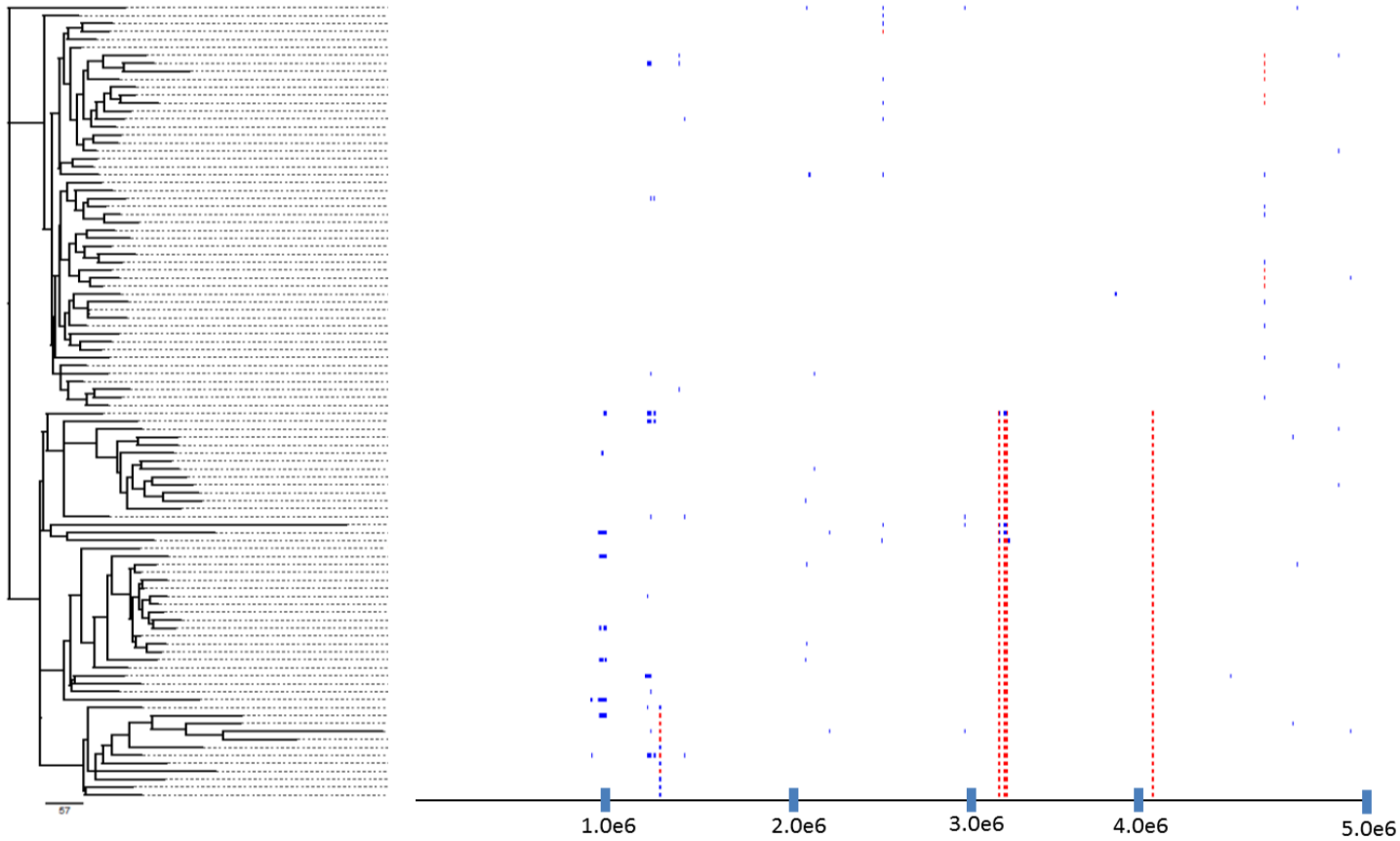


Figure 2.3. Representation of recombination events along the genome for each branch of the reconstructed phylogeny produced by Gubbins. The scale bar underneath the phylogenies represent a phylogenetic distance of 57 point mutations. Red blocks indicate predicted combinations occurring on an internal branch, which are therefore shared by multiple isolates through common descent. Blue blocks represent recombinations that occur on terminal branches, which are unique to individual isolates.

BRATNextGen does not compute recombination frequencies, so we compared the recombination to mutation ratios (r/m) estimated by Gubbins (Supplementary Table S2.1A and B) and ClonalframeML (Supplementary Table S2.2A and B). Although results for both agree that recombination rather than mutation occurred more often (r/m>1), Gubbins had a slightly more conservative estimate of the donation frequency (~1.18) in contrast to the value calculated by ClonalframeML (~1.87; Table 2.1).

Sample Count	**R/Θ**		**δ (bp)**		**ν**		**r/m**	
	Gubbins	CfML	Gubbins	CfML	Gubbins	CfML	Gubbins	CfML
100	0.028	0.273	16506.72	816.6	0.003	0.014	1.177	1.874

Table 2.1. Summary of recombination parameters computed by Gubbins and ClonalframeML (CfML). Shown are the quantified recombinogenic SNPs (R), the recombination to mutation ratio (r/m), the mean DNA import length (δ), the mean divergence of recombinant segment (ν) and the number of SNPs caused by point mutation (Θ) in both recent (external branches) and ancestral (internal branches).

The reverse was true for determining the average length of imported segments: Gubbins estimated the mean length in the 100 isolates of 16.5 Kb, while it was only 816.6 bp with ClonalframeML. This computation was confounded by some samples with relatively large import sizes detected by Gubbins (i.e. MR002251M-47.6-A-Galway-Galway-X-1-AC: ~63.9kb) and a few with none, whereas ClonalframeML had a better resolution of exact recombinant tract boundaries. MU053687K-54.4-A-Ardrahan-Galway-Resident2-1-A, for example, had no donated segment according to Gubbins but had an imported ~79.84 bp tract based on ClonalframeML. Although we found that the number of tracts was higher when the –embranch dispersion value was increased, there was no conclusive correlation between the mean recombinant tract length and recombination per branch.

We were not able to detect recombinogenic regions using BRATNextGen (Figure 2.4A) hence we excluded the results obtained using this program and based our recombination analysis of these 100 ST131 using ClonalframeML and Gubbins in the external and internal branches (Figure 2.4B). Both programs detected 21 segments on the external branches and two on the internal ones, and four were shared by both internal and external branches.

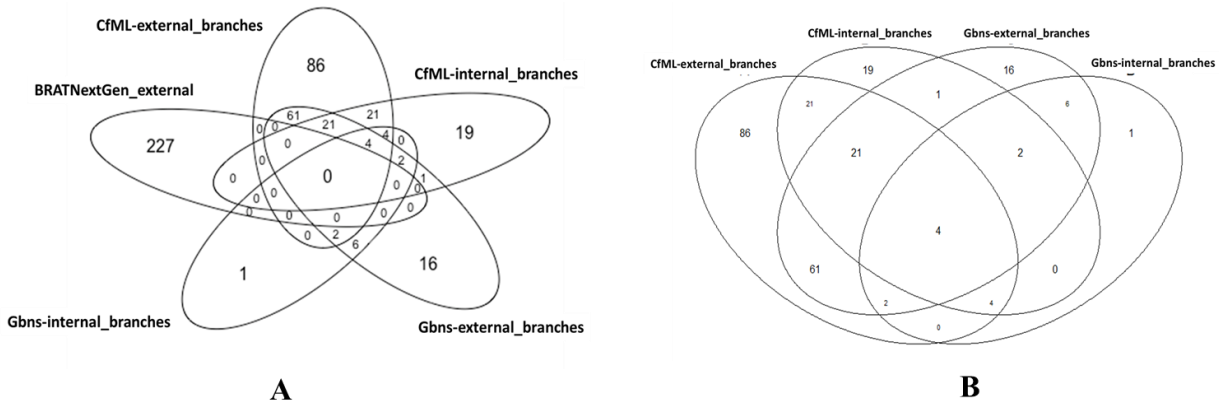


Figure 2.4. Number of recombinant tracts detected by BRATNextGen, ClonalframeML (CfML) and Gubbins (Gbns) in the internal and external branches of Irish/UK N=100 ST131 population. No tracts that were found in the same genomic location (common tracts) were detected between BRATNextGen, ClonalframeML and Gubbins (A) hence BRATNextGen was excluded in the analysis (B).

2.4 Discussion

The frequency and impact of recombination events complicates bacterial population structure interpretation. This was the case in the 100 ST131 here, with a recombination to mutation ratio greater than 1.0. The three recombination detection programs we tested on the 100 ST131 suggested that recombination more than mutation may be a key factor elevating the genetic heterogeneity of this collection. Additionally, all three tools were found to be computationally viable in a large bacterial genome dataset (unlike the original ClonalFrame). We used Gubbins and ClonalframeML more in subsequent studies because of ClonalframeML's higher sensitivity in detecting recombinogenic SNPs, and Gubbins's widespread use by the community in bacterial genome analysis pipelines. BRATNextGen can be utilized for identifying whether the detected segments originated from the data set at hand or were introduced from an unknown external source but was not applied further in thesis.

This chapter showed that phylogenetic methods such as RAxML can be effective in dissecting complex bacterial population structure using genomic data, but that model-based methods like Structure may not be. This may be because of the lack of core genome variation in a closely related collection, and discrete nature of recombination signals that only affected small portions of the core genome overall but were at a much higher concentration in the accessory genome.

This study was not explored further because it used the NA114 genome published by Avasthi et al. (2011) as a reference, but it emerged during our work that this reference was fundamentally flawed due to improper assembly methods (Forde et al. 2014). This genome had been used to identify the SNP set we examined in 2012 but was only 4.9 Mb in length mainly due to contig misplacement and the lack of appropriate gaps at contig edges, and so was missing 200 Kb compared to other references (Petty et al. 2013, Forde et al. 2014); this likely affected studies that used the NA114 genome as a reference such as Price et al. (2013). Consequently, an alternate reference genome was used in Chapters 3-6.

2.5 Supplementary Tables

The supplementary tables of this chapter are publicly available on Figshare: https://figshare.com/s/6e8ff1247c95a69683bc.

Supplementary Table S2.1. Recombination to mutation ratio (r/m), the mean DNA import length (δ) computed by Gubbins in both recent (external branches (A)) and ancestral (internal branches (B)).

Supplementary Table S2.2. Recombination to mutation ratio (r/m), the mean DNA import length (δ) and the mean divergence of recombinant segment (ν) computed by ClonalframeML in both recent (external branches (A)) and ancestral (internal branches (B)).

2.6 References

Avasthi TS, Kumar N, Baddam R, Hussain A, Nandanwar N, Jadhav S, Ahmed N. Genome of multidrug-resistant uropathogenic Escherichia coli strain NA114 from India. J Bacteriol. 2011 193(16):4272-3. doi: 10.1128/JB.05413-11.

Clark, G., Paszkiewicz, K., Hale, J., Weston, V., Constantinidou, C., Penn, C., Achtman, M. and McNally, A., 2012. Genomic analysis uncovers a phenotypically diverse but genetically homogeneous Escherichia coli ST131 clone circulating in unrelated urinary tract infections. Journal of antimicrobial chemotherapy, 67(4), pp.868-877.

Coque TM, Baquero F, Canton R. 2008. Increasing prevalence of ESBL-producing Enterobacteriaceae in Europe. Euro Surveill. 13(47):pii=19044.

Coque TM, Novais A, Carattoli A, Poirel L, Pitout J, Peixe L, Baquero F, Canton R, Nordmann P. 2008. Dissemination of clonally related Escherichia coli strains expressing extended-spectrum β-lactamase CTX-M-15. Emerg. Infect. Dis. 14:195–200. 10.3201/eid1402.07035010.

Croucher N. J., Page A. J., Connor T. R., Delaney A. J., Keane J. A., Bentley S. D., Parkhill J., Harris S.R. "Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins". doi:10.1093/nar/gku1196, Nucleic Acids Research, 2014.

Didelot X, Wilson DJ (2015) ClonalFrameML: Efficient Inference of Recombination in Whole Bacterial Genomes. PLoS Comput Biol 11(2): e1004041. https://doi.org/10.1371/journal.pcbi.1004041.

Downing, T., 2015. Tackling drug resistant infection outbreaks of global pandemic Escherichia coli ST131 using evolutionary and epidemiological genomics. Microorganisms, 3(2), pp.236-267.

Hummers-Pradier E, Koch M, Ohse AM, Heizmann WR, Kochen MM. 2005. Antibiotic resistance of urinary pathogens in female general practice patients. Scand. J. Infect. Dis. 37:256–261. 10.1080/00365540410021009.

Huson DH and Bryant D, 2006. Application of phylogenetic networks in evolutionary studies. Molecular Biology and Evolution, Volume 23, Issue 2, February 2006, Pages 254–267, https://doi.org/10.1093/molbev/msj030.

Johnson JR, Brian J, Connie C, Kuskowski MA, Mariana C. 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. 10.1086/653932.

Jones RN, Kehrberg EN, Erwin ME, Anderson SC. 1994. Prevalence of important pathogens and antimicrobial activity of parenteral drugs at numerous medical centers in the United States, I. Study on the threat of emerging resistances: real or perceived? Fluoroquinolone Resistance Surveillance Group Diagn. Microbiol. Infect. Dis. 19:203–215.

Ludden, C. 2014. The role of long-term care facilities in the dissemination of antimicrobial resistance. Unpublished doctoral dissertation. National University of Ireland Galway, Ireland.

Marttinen, P., Hanage, W.P., Croucher, N.J., Connor, T.R., Harris, S.R., Bentley, S.D. and Corander, J., 2011. Detection of recombination events in bacterial genomes from large population samples. Nucleic acids research, 40(1), pp.e6-e6.

Nicolas-Chanoine MH, Blanco J, Leflon-Guibout V, Demarty R, Alonso MP, Canica MM, Park Y-J, Lavigne J-P, Pitout J, Johnson JR. 2008. Intercontinental emergence of Escherichia coli clone O25:H4-ST131 producing CTX-M-15. J. Antimicrob. Chemother. 61:273–281 doi:10.1093/jac/dkm464.

Petty, Nicola K., et al. "Global dissemination of a multidrug resistant Escherichia coli clone." Proceedings of the National Academy of Sciences111.15 (2014): 5694-5699.

Pitout JD: Extraintestinal Pathogenic Escherichia coli: A Combination of Virulence with Antibiotic Resistance. Front Microbiol. 2012;3:9. 10.3389/fmicb.2012.00009.

Price, L.B., Johnson, J.R., Aziz, M., Clabots, C., Johnston, B., Tchesnokova, V., Nordstrom, L., Billig, M., Chattopadhyay, S., Stegger, M. and Andersen, P.S., 2013. The epidemic of extended-spectrum-β-lactamase-producing Escherichia coli ST131 is driven by a single highly pathogenic subclone, H30-Rx. MBio, 4(6), pp.e00377-13.

Rambaut A, Lam TT, Max Carvalho L, Pybus OG. 2016. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). Virus Evolution 2.

Russo TA, Johnson JR. 2003. Medical and economic impact of extraintestinal infections due to Escherichia coli: an overlook epidemic. Microbes Infect. 5:449–456. 10.1016/S1286-4579(03)00049-2.

Stamatakis A. Raxml-vi-hpc: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics. 2006;22:2688–2690.

Turnidge J. 1995. Epidemiology of quinolone resistance. Eastern hemisphere. Drugs 49:43–47.

Vos M, Didelot X (2009) A comparison of homologous recombination rates in bacteria and archaea. ISME J 3: 199–208. pmid:18830278.

Chapter 3: Genomic surveillance of *E. coli* ST131 reveals the evolutionary history of epidemic antimicrobial resistant clones

Abstract

Escherichia coli sequence type 131 (ST131) is a pandemic clonal group that is evolving rapidly with increasing trends in antimicrobial resistance. Here, we investigated an outbreak of extended spectrum β-lactamase (ESBL) producing *E. coli* ST131 in a long-term care facility (LTCF) in Ireland (n=90) and combined this data with global (n=704) ST131 genomes to reconstruct the evolutionary history and further understand changes in population structure and recombination patterns over time. Three major ST131 clades circulating worldwide named A, B and C were identified here, of which the latter was the largest cluster accounting for 686/794 (86%) isolates and was associated with ciprofloxacin resistance, the presence of ESBL genes, diverse plasmids and transposable elements. Clade C had various evolutionary events influenced by plasmid flux, recombination and local rearrangements. C subclades also had distinctive changes in plasmid content and ESBL gene variants (*bla*$_{\text{CTX-M-14}}$ vs *bla*$_{\text{CTX-M-15}}$) including a chromosomal insertion of *bla*$_{\text{CTX-M-15}}$ at the *mppA* gene identified in an Irish LTCF lineage using long-read sequencing. IS*Ecp1* transposed the *bla*$_{\text{CTX-M-15}}$ gene from an IncFIA plasmid and subsequently this C2 subtype clonally expanded, causing an outbreak in this LTCF. We conclude that within the pandemic ST131 was been diversifying over a 7- year period with a mutation rate of 4.14×10^{-7} SNPs/site annually. There was evidence that extensive rearrangement of ESBL genes in plasmids in chromosome occurred which contributed to the spread of diverse clones worldwide and a local outbreak in a LTCF in Ireland which spanned four years.

Publication: for submission to *mBio* in 2019 with Ludden C, Jamrozy D, Zhou Z, Pickard D, Horner C, Morris D, Parkhill J, Peacock SJ, Achtman M, Dougan G, Downing T, Cormican M.

3.1 Introduction

Escherichia coli is the leading cause of urinary tract infections and bloodstream infections (BSIs) (Tumbarello et al. 2010, Burns et al. 2012), with the number of E. coli BSIs continuing to increase in Europe and the United States since the early 2000s (Public Health England 2017, Gagliotti et al. 2011, Poolman and Wacker 2016, Thaden et al. 2016, ECDC 2017). This has been associated with the emergence and dissemination of antibiotic-resistant E. coli producing extended-spectrum β-lactamases (ESBL-E. coli) conferring resistance to many beta-lactam antibiotics, including cephalosporins (Thaden et al. 2016, ECDC 2017). Infections caused by ESBL-E. coli are associated with higher morbidity and mortality, longer hospital stays and higher healthcare costs compared to infections with antibiotic-susceptible E. coli (Tumbarello et al. 2010, Schwaber and Carmeli 2007, Rottier et al. 2012, Roberts et al. 2009).

The global spread of ESBL-E. coli is largely attributed to the dissemination of E. coli strains carrying the blaCTX-M-15 gene, especially E. coli O25b:H4-ST131. Genomic analyses estimated that ST131 emerged in North America over 30 years ago, coinciding with the first use of fluoroquinolone (FQ) in 1986 (Stoesser et al. 2016, Ben Zakour et al. 2016). Previously, three major lineages of ST131 were identified that differed mainly based on their fimH alleles: A (mainly fimH41), B (mainly fimH22) and C (mainly fimH30) (Matsumura et al. 2017). Clade C has predominated since the 2000s, corresponding with the rapid dissemination of the blaCTX-M-15 allele (Matsumura et al. 2017, Canton et al. 2012). Clade B also contains the subclade B0 which differs phylogenetically from the remaining B isolates by carrying fimH27 and is considered ancestral to Clade C (Matsumura et al. 2017, Canton et al. 2012, Kallonen et al. 2017). Clade C consists of three subclades termed C0, C1 and C2. Clade C0 has been reported as ancestral and is composed of FQ-susceptible isolates. In contrast, clades C1 (also known as H30R) and C2 (also known as H30Rx) are characterised by a double mutation at the gyrA and parC genes conferring high-level resistance to FQ (Stoesser et al. 2016, Matsumura et al. 2017, Price et al. 2013). Clade C2 is sub-divided from C1 based on specific SNPs at fimH30 as previously described and is associated with the blaCTX-M-15 gene (Price et al. 2013, Zerbino and Birney 2008).

ST131 has principally been associated with the hospital setting, though in recent years it has also been reported at high prevalence in the community (Vidal-Navarro et al. 2010, Rogers et al. 2011, Tchesnokova et al. 2018). There is increasing evidence that ST131 is common in the elderly and that long-term care facilities (LTCFs) are important reservoirs for ESBL-producing ST131. Reported rates of multidrug-resistant (MDR) E. coli ST131 carriage in residents of LTCFs include 55% in Ireland, 36% in the UK and 24% in the United States (Ludden et al. 2015, Brodrick et al. 2017, Burgess et al. 2015). It is projected that the proportion of the European Union population aged ≥ 65 years and ≥ 80 years will increase to 29% and 11.5 % by 2060, respectively (Suetens 2012). This will likely lead to a rise in the number of people residing in LTCFs, potentially expanding the reservoir of ESBL-producing ST131. Infection control measures targeting E. coli have focused primarily on hospitals, and there is still a limited understanding of E. coli transmission dynamics within LTCFs, and between hospitals and LTCFs (Brodrick et al. 2017, Burke et al. 2012). To develop effective strategies for containment and prevention of infections, it is necessary to improve our ability to detect transmission events and to monitor the emergence of new clones. Here, we used short and long read genome sequencing to investigate an ESBL-E. coli ST131 outbreak in a LTCF in Ireland. We describe the genetic basis of antibiotic resistance and the evolution of ESBL-E. coli ST131 over a seven-year period. We focused our analyses on ST131 clade C because of its high frequency in this LTCF, and its MDR profile. We analysed the population structure and inferred the evolutionary history of the LTCF isolates in the context of a local hospital and global collections of E. coli ST131 to further our understanding of its epidemiology.

3.2 Methods

Author contributions: This chapter is a collaborative work between several researchers working on bacterial genomics and clinical microbiology. I was involved in all methods, bioinformatic processing, genomic analysis, interpreting results, drafting the paper, editing the paper and visualization of the results. First-authorship is shared between myself and Catherine Ludden, who helped in securing project funding, conceptualization, interpreting results, drafting the paper and editing the paper. Dorota Jamrozy assisted with temporal phylogenetic analyses. Zhemin Zhou helped with initial bioinformatic and genomic analyses. Mark Pickard, Carolyne Horner, Dearbhaile Morris, Mark Achtman, Sharon Peacock and Gordon Dougan were in volved in conceptualization, sample collection and study design. Julian Parkhill and Martin Cormican were in volved in conceptualization, study design, project management and paper writing. Tim Downing helped with bioinformatics, genomics analyses, interpreting results and paper writing. I completed a significant component of the work in this multi-partner chapter and was involved in all aspects.

3.2.1 Irish bacterial isolate collection and short read genome sequencing

A total of 90 E. coli ST131 isolates from Ireland were isolated and sequenced. Among these 90, 69 were sampled from 63 residents during 2005-2011 from a single LTCF with an outbreak of ESBL-producing *E. coli* in 2006 (Pelly et al. 2006) and 21 were clinical isolates from the referral hospital (Galway University Hospital: n=8 hospitalized patients, n=11 residents of other Irish LTCFs, and n=2 community isolates submitted from general practitioners).

Bacterial genomic DNA for the 90 isolates was extracted using the QIAxtractor (Qiagen, Valencia, CA, USA) according to the manufacturer's instructions. Library preparation was conducted according to the Illumina protocol and sequenced (96-plex) on an Illumina HiSeq 2000 platform (Illumina, San Diego, CA, USA) using 100 bp paired-end reads. On average, 5,014,175 (range 3,489,126-8,166,084) raw sequence reads were generated per isolate, with a mean insert size of 260 (range 244-280).

3.2.2 Complementary datasets

For context, DNA read libraries and associated metadata were retrieved for 704 E. coli ST131 isolates, 14 of which were BSIs from four referral hospitals in Ireland and 4/14 isolates were obtained from the referral hospital (Galway University Hospital). The remaining global 690 were isolated between 1967 and 2014 and included 167 (clinical=155, environmental=7, unknown=5) isolates obtained from global collections (Price et al. 2013, Hull et al. 1981), 297 from a UK LTCF (Brodrick et al. 2017), and 226 were associated with BSI in the UK (Kallonen et al. 2017, Brodrick et al. 2017) (Supplementary Table 3.1).

3.2.3 Long read sequencing, assembly and annotation

DNA was extracted using the phenol-chloroform method (Hull et al. 1981) and sequenced using a PacBio RSII Instrument (Pacific Biosciences, Menlo Park, CA, USA) for five isolates (ERR191646, ERR191657, ERR191663, ERR191724 and ERR191697). Sequence reads were assembled using HGAP v3 (Chin et al. 2013) of the SMRT analysis software v2.3.0 (https://github.com/PacificBiosciences/SMRT-Analysis), circularized using Circlator v1.1.3 (Hunt et al. 2015) and Minimus 2 (Sommer et al. 2007), and polished using the PacBio RS_Resequencing protocol and Quiver v1 (https://github.com/PacificBiosciences/SMRT-Analysis). This assembled the plasmids for each of the isolates used as references for short read mapping. NCTC13441's HDF5 files were converted to FASTQ with 308,854 reads using pbh5 tools (smrtanalysis v2.3.0p4). These reads were screened for PacBio adapter sequence using Cutadapt v1.9.1 and corrected using BayesHammer from SPAdes v3.0.0 with a seed k-mer of 127, yielding a total of 41,813 reads.

3.2.4 Genome assembly, read mapping, AMR gene identification and plasmid typing of the 794

De novo assembly of short read data for the 794 libraries was performed using VelvetOptimiser v2.2.5 (Gladman and Seemann 20018) and Velvet v1.2 (Zerbino and Birney 2008). An assembly improvement step was applied to the assembly with the best N50, whose contigs were scaffolded using SSPACE (Boetzer et al. 2011) and contig gaps reduced using GapFiller (Boestzer et al. 2012). The assembly pipeline generated an

average total length of 5,166,846 bp (range 4,697,700-5,460,279 bp) from 97 contigs (range 31-486) with an average contig length of 59,340 bp (range 11,186-1,661,401 bp) and an N50 of 227,849 bp (range 30,788-763,538 bp) (Supplementary Table 3.5). Assemblies annotated using Prokka v1.5 (Seemann 2014) and a genus-specific database from RefSeq (Pruitt et al. 2012).

The 794 short read libraries were mapped to NCTC13441 genome (accession ERS530440) (Brodrick et al. 2017), PacBio assemblies and reference plasmids using SMALT v7.6 (http://www.sanger.ac.uk/resources/software/smalt/). The genomic locations of the blaCTX-M genes and nearby MGEs were examined by aligning the short and long read assemblies using BLAST to the blaCTX-M-positive TU isoforms, including one with a split mppA gene containing the TU (Supplementary Figure 3.3). The two observed mppA isoforms were recorded as T for truncated (separated 327 bp and 1290 bp segments) or I for intact (Supplementary Table 3.1). SNP screening at mppA across the 794 showed limited variation: just one doubleton and four singleton SNPs.

AMR genes in the 794 were identified by alignment with the 2,158 gene homolog subset of the Comprehensive Antibiotic Resistance Database (CARD) v1.1.5. Plasmid incompatibility group and replicon types were identified (Supplementary Table 3.6) by comparing the genomes against the PlasmidFinder database (accessed date 16/03/17) (Carattoli et al. 2014) with a 95% identity threshold.

3.2.5 Quality control, genome assembly and read mapping of 54 Irish read libraries

Adapter sequences in the libraries of the 54 Irish Clade C reads were trimmed with Trimmomatic v0.36 (Bolger et al. 2014) using a Phred score threshold of 30 (Q30), a ten bp sliding window and a minimum read length of 50 bp. On average, these had 2,400,763 reads initially, of which 7.8% were removed by trimming. These were corrected using BayesHammer in SPAdes v3.9. The effects of removing low-quality bases and reads was quantified using FastQC v0.11.5 with MultiQC v1.3, which showed base-correction removed an additional 14.3% of reads on average, leaving a mean of 1,898,990 per

library. This showed levels of base quality and potential contaminants were consistent across the libraries.

Read libraries of the Irish 54 were assembled into contigs using SPAdes v3.9 with a k-mer of 77 (Bankevich 2012). This optimal k-mer maximised the N50 value determined by Quast v5.0 (Gurevich et al. 2013). The contigs were ordered and scaffolded based on the NCTC13441 reference chromosome, plasmid and annotation using ProgressiveMauve (Darling et al. 2010), producing an average scaffold N50 of 177,758±12,199 (mean±SD) bp with a mean assembly length of 5,434,674±153,210 bp and an average of 234 contigs per library.

A total of 59,536 bases at low complexity repeats, homopolymers, sites within 1 Kb of chromosome edges, bases within 100 bp of a contig edge, or at tandem repeats were masked from the NCTC13441 reference chromosome using Tantan v0.13 (www.cbrc.jp/tantan/), which was indexed using SMALT v7.6 using a k-mer of 19 with a skip of one, as were all reference sequences here. The short read libraries were mapped to reference sequences using SMALT v7.6, and the resulting SAM files were converted to BAM format, sorted and PCR duplicates removed using SAMtools v1.19. The MGE, mppA and blaCTX-M gene structures were examined by alignment as above so that local copy number changes, mapping breakpoints and read pileups could be screened by mapping Illumina reads to the PacBio and contig references. The local gene structure was visualised with R v3.5.2 and the MARA Galileo AMR database (Partridge et al. 2009, Tsafnat et al. 2009).

3.2.6 Phylogenetic analysis of 794 isolates

To construct phylogenies reflecting the genealogical relationships and evolutionary changes, SNPs were identified using Gubbins v2.3.4 (Croucher et al. 2014). The SNPs arising by mutation were used to create a maximum-likelihood midpoint-rooted phylogeny using RAxML v8.0.19 (Stamakis 2014) using a General Time Reversible + gamma (GTR+G) substitution model with 100 bootstraps across 362,009 sites. Phylogenetic trees were visualised with iToL (http://itol.embl.de) (Letunic and Bork 2016) and FigTree v1.4.3 (http://tree.bio.ed.ac.uk/software/figtree/) (Rambaut et al.

2016). For the 54 Irish Clade C collection, a phylogeny was created as above with RAxML with 100 bootstraps, and a network was constructed using uncorrected p-distances with Splitstree v4.14.2 (Huson and Bryant 2005), visualized with FigTree.

The 362,009 core SNP sites were also used as the sparse matrix using the default parameters of a hierarchical Bayesian clustering algorithm implemented in Fastbaps v1.0 (Fast Hierarchical Bayesian Analysis of Population Structure; Tonkin-Hill et al. 2019) 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.

3.2.7 Inference of subclade common ancestry and historical population size changes

To reconstruct time-calibrated phylogeny ST131 we used a core genome alignment of 794 isolates that contained 8,567 SNPs after the exclusion of regions representing MGEs, recombinant tracts and sites with an uncalled genotype across >1% of sequences. Each sequence in the alignment was annotated with the year of isolation. The strength of the molecular clock signal was measured by linear regression of the root-to-tip genetic distance against year of sampling using TempEst (Rambaut et al. 2016), which revealed a correlation coefficient of R2 = 0.4. Bayesian inference of phylogeny was performed with BEAST v2.4.7 (Bouckaert et al. 2014) based on a GTR+G nucleotide substitution model. To optimise computing efficiency in a large dataset, model selection was implemented on a subset of isolates (n=205) that tested two clock rates (strict versus relaxed uncorrelated lognormal) across three population models (constant, exponential and Bayesian skyline). Five replicates for each of the six models were tested. The MCMC chain was run for 50 million generations, sampling every 1,000 states. Log files from the five independent runs per model option were assessed for convergence using Tracer v1.5, and combined after removal of the burn-in (10% of samples) using LogCombiner. The relaxed lognormal clock with Bayesian skyline model was the best fit, consistent with previous work (von Mentzer et al. 2014) and so this was used to model the evolutionary history across all 794 isolates with 15 replicates. The maximum clade credibility (MMC) tree was generated with TreeAnnotator.

3.2.8 Summary workflows

For simplicity, the methods described in the above sections of this chapter are illustrated in the following flow diagrams. Processing of N=90 Irish ST131 samples and its subsets are shown in Figure 3.2.8.1 while the analyses of the total N=794 ST131 libraries ,composed of the N=90 Irish isolates and the n=704 global ST131 from previous studies, are presented in Figure 3.2.8.2.

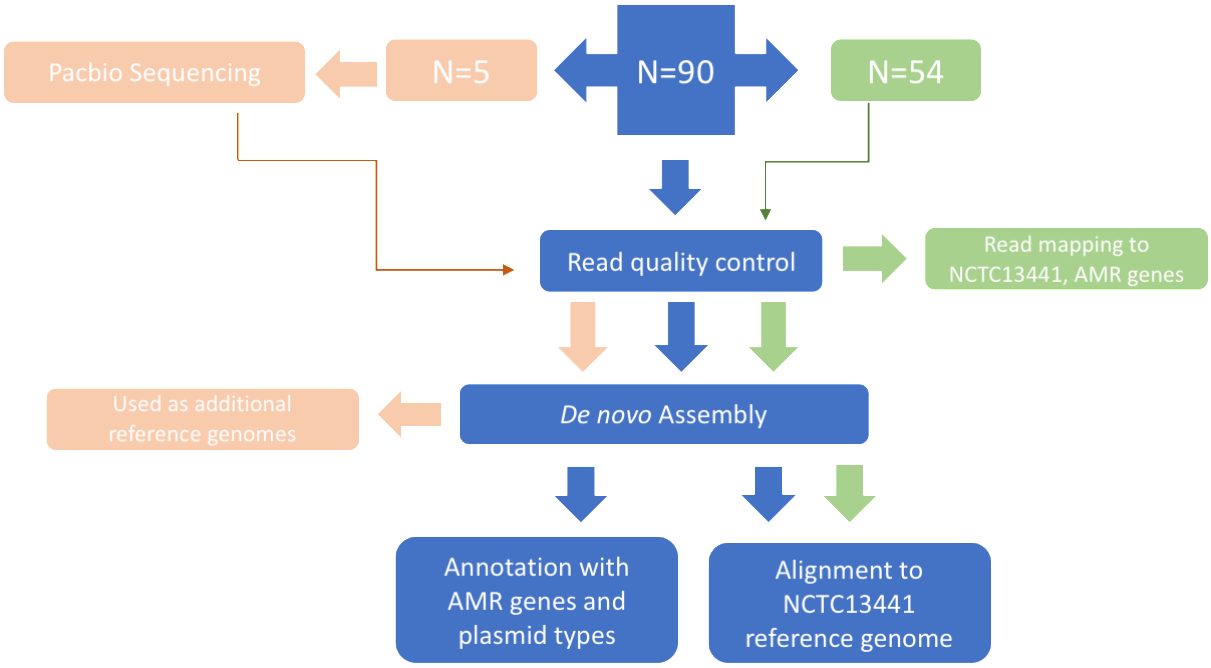


Figure 3.2.8.1. Illumina HiSeq libraries of N=90 ST131 (blue) isolates that caused an outbreak in Irish nursing homes were rid of adapters and base calling errors (read quality control), *de novo* assembled, annotated with AMR genes and plasmid types and finally aligned to NCTC13441 reference genome. Long reads of a subset of this N=90 ST131 with N=5 samples (orange) were sequenced using Pacbio long read sequencing technology to ultimately used as supporting reference genomes, following read quality control. Another subgroup from the N=90, which contains n=54 strains (green) were processed using the same methods for read quality control before mapping their Illumina Hiseq reads to NCTC13441 and reference AMR genes.

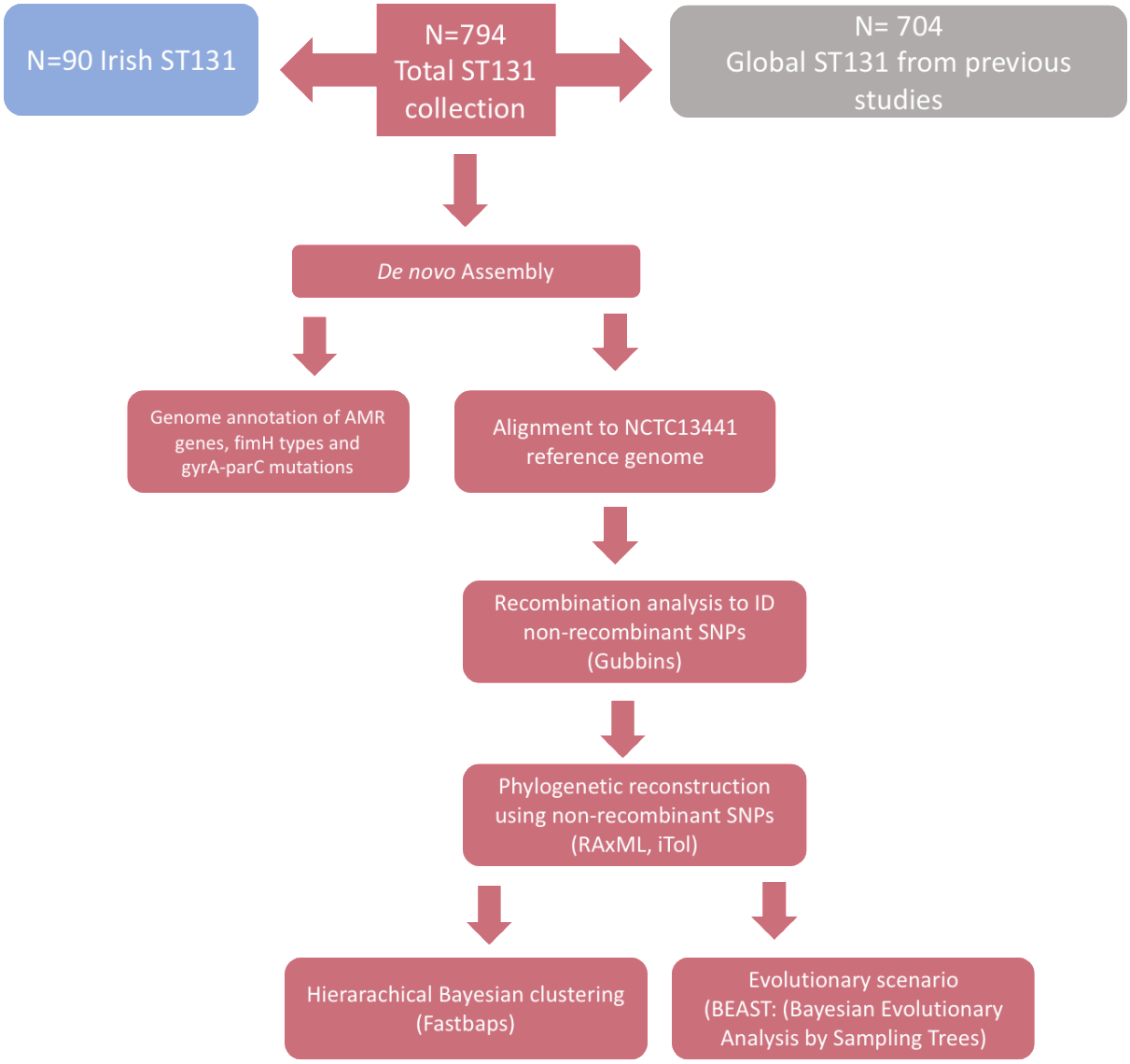


Figure 3.2.8.2. The global context of the N=90 ST131 (blue) from Irish nursing homes was analysed by including n=704 (grey) more ST131 strains in the collection. The total N=794 ST131 (pink) Illumina Hiseq libraries were de novo assembled, annotated with blaCTX-M genes, fimH types, gyrA-parC mutations, and aligned to the reference genome NCTC13441. The alignment was then used to identify the non-recombinant (core) SNPs by running Gubbins. These core SNPs were then employed to generate a maximum likelihood phyogeny using RAxML and drawn using iTol. Unique genetic clusters with the reconstructed phylogenetic tree was identified using a hierarchichal Bayesian clustering implemented in Fastbaps. Finally, the evolutionary secnarios in the N=794 ST131 population were dated using BEAST (Bayesian Evolutionary Analysis by Sampling Trees).

3.3 Results

3.3.1 ESBL gene profiles among an *E. coli* ST131 outbreak in Ireland

In this study, we focused on the genetic profiles of 90 *E. coli* ST131 (local collection) isolated between 2005 and 2011 in Ireland, of which 69 were from one LTCF where an outbreak of ESBL-*E. coli* was first detected in 2006 (Pelly et al., 2006). The other isolates were from other LTCFs (n=9), the referral hospital (n=10) and the community (n=2) (Supplementary Table 3.1). Initial screening of the 90 isolates indicated that 64 were *bla*$_{CTX\text{-}M\text{-}15}$-positive, 17 were *bla*$_{CTX\text{-}M\text{-}14}$-positive, one was *bla*$_{CTX\text{-}M\text{-}27}$-positive, and four were positive for both *bla*$_{CTX\text{-}M\text{-}15}$ and *bla*$_{CTX\text{-}M\text{-}14}$ (Supplementary Table 3.1). Resistance to meropenem and ertapenem was not detected. Ribosomal sequence typing (rST) demonstrated a high incidence of rST1850 (44/90, 49%) (Supplementary Table 3.1), suggesting emergence of a unique local epidemic clone.

3.3.2 ST131 clade C predominates in Ireland and elsewhere

We analysed the 90 isolates from the local collection in the context of a global collection of 704 *E. coli* ST131 genomes that contained four additional isolates from the referral hospital described in the local collection and 10 isolates from other hospitals in Ireland. To better understand the global population structure of *E. coli* ST131, we reconstructed the phylogeny of all 794 isolates based on a core genome alignment containing 12,518 SNPs (Figure 3.1). This recapitulated the three established ST131 clades (A, B and C) (Johnson et al. 2017) and showed that most isolates were from C (n=686, 86.4%) followed by B (n=75, 9.4%) and A (n=33, 4.2%). The clade classification was supported by previously described *fimH* allelic differences (Price et al. 2013): clade A was largely *fimH41* (30 out of 33), clade B *fimH22* (60 out of 70), subclade B0 *fimH27* (4 out of 5) and clade C *fimH30* (679 out of 686) (Table 3.1).

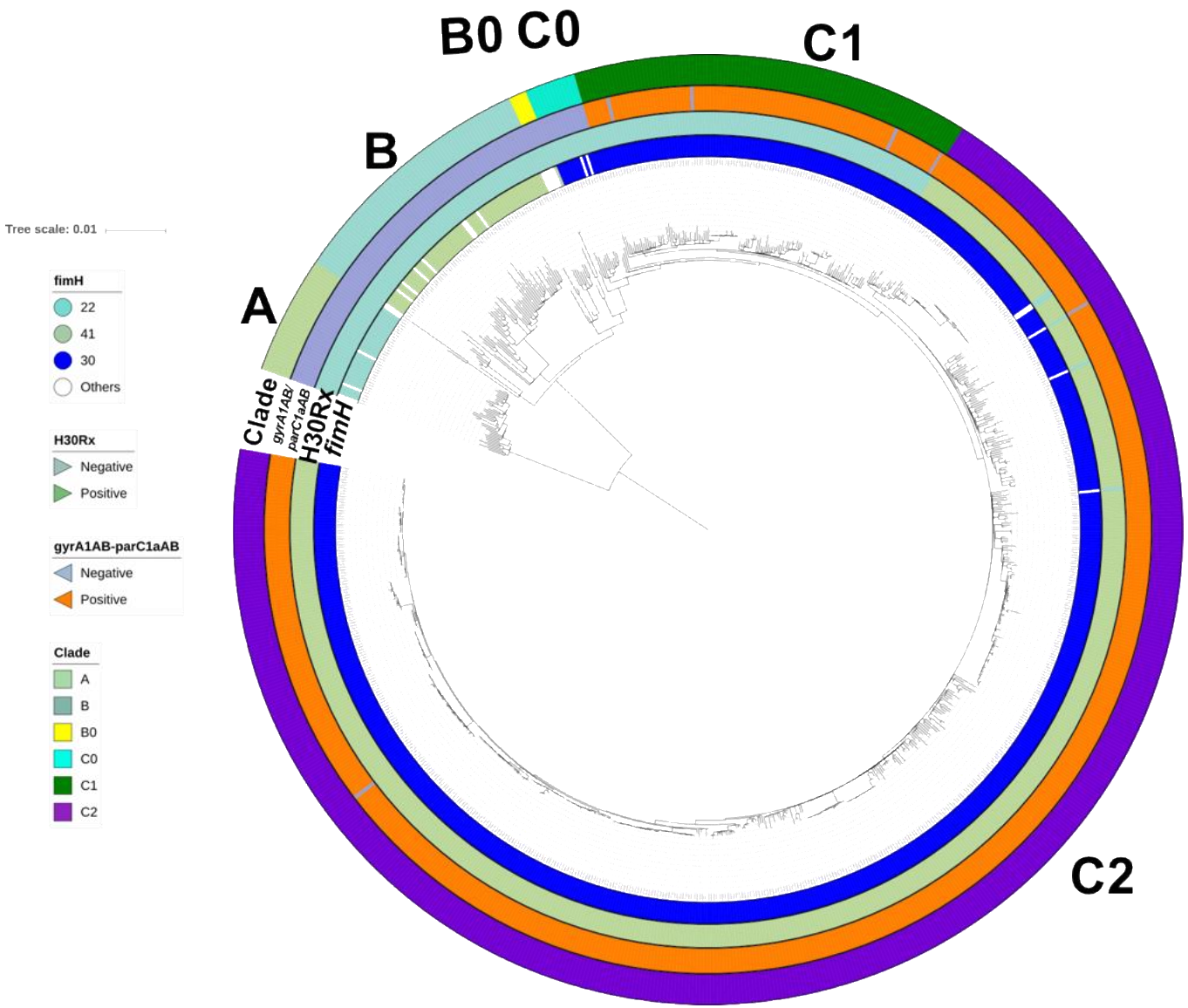


Figure 3.1. Maximum likelihood phylogeny of n=794 global ST131 showed three main clades A (n=33), B (n=70), B0 (n=5) and C (n=686) with three common subclades in C: C0 (n=14), C1 (n=111) and C2 (n=561). The mid-point rooted phylogram was constructed with RAxML from the chromosome-wide SNPs arising by mutation, and visualized with iTol. Allelic profiling of fimH, gyrA-parC, the H30Rx phenotype, and clade classification are represented in colored strips around the phylogenetic tree.

***fimH* allele**	A	B	B0	C0	C1	C2	Total
H41	30						30
H22		60	1				61
H27			4				4
H30 (non-Rx)				12	111		123
H30Rx						55	556
Other	3	10		2	1	4	20
Total	33	70	5	14	112	56	794

Table 3.1. The entire ST131 set (n=794) consisted of three main clades sub-divided into six subclades: A (n=33), B (n=70), B0 (n=5), C0 (n=14), C1 (n=111) and C2 (n=561). The frequencies of the four most common *fimH* allele types are shown: *H41*, *H22*, *H27* and *H30*; the rest are listed as "other". No FQ-resistance mutations were detected in *fimH22/27/41.*

FQ-resistance alleles *gyrA1AB* and *parC1aAB* (Hull et al. 1981) were present in nearly all C1 (96%) and C2 (99.7%) isolates, along with the *fimH30* allele, contrasting with their absence from the clades A, B, and B0 (Supplementary Table 3.1). This indicated that the Clade C ancestor acquired the *fimH30* allele and then differentiated into FQ-S (*H30*S or C0) and FQ-R (*H30*R or C1, *H30*Rx or C2) subclades. A limited number of C1 (n=1) and C2 (n=4) isolates had lost the FQ-R *gyrA1AB-parC1aAB* genotype, consistent with intermittent recombination at these and the *fimH* genes (Stoesser et al. 2016).

Considerable diversity within Clade C was demonstrated by the genetic clusters identified by Fastbaps (Figure 3.2 and Table 3.2): C0 (n=14, Fastbaps clusters 2-5 and 11), C1 (n=111, Fastbaps cluster 10), and C2 (n=560, Fastbaps clusters 7-9). All 104 Irish ST131 from the National Collection (local = 90, additional Irish isolates = 14, see Methods) were from clade C and there were no major differences in the rates of C0, C1 and C2 in the National collection (1%, 23% and 75%, respectively) compared to the global isolate collection (2%, 12%, 70%, respectively).

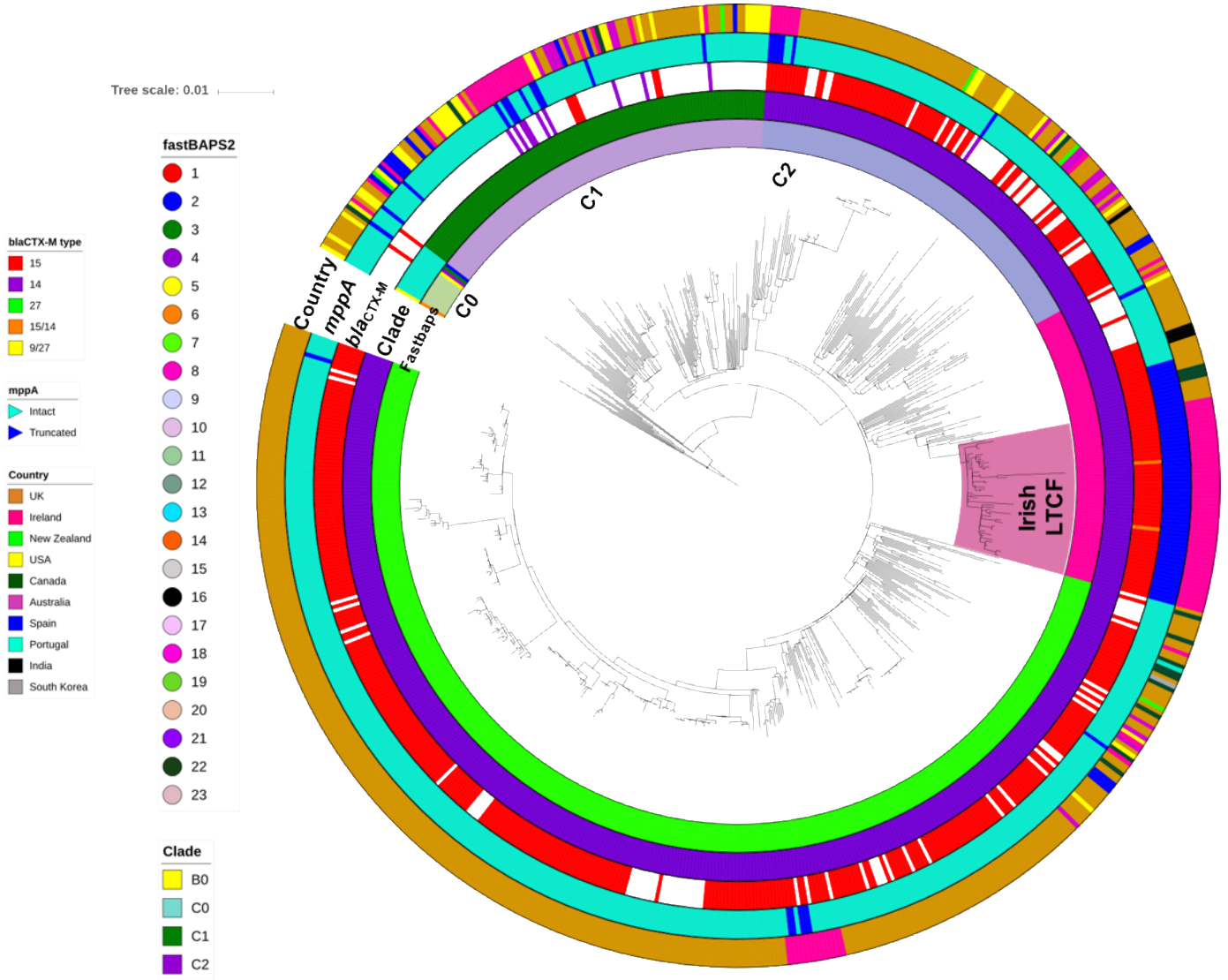


Figure 3.2. Phylogenetic reconstruction of 686 strains from Clade C with B0 as the outgroup. This shows 3 common subclades in C: C0 (n=14), C1 (n=111) and C2 (n=561) where the latter had three distinct subgroups: C2_7 (n=362, Fastbaps cluster 7), C2_8 (n=86, Fastbaps cluster 8) and C2_9 (n=113, Fastbaps cluster 9). Colored strips surrounding the phylogram represent the clade classification, Fastbaps clusters, *bla*$_{CTX-M}$ allelic profile, *mppA* state (intact or truncated) and the country of origin of each strain. The highlighted "Irish LTCF" clade was in C2_8.

3.3.3 Phylogenetic reconstruction of three genetically distinct ST131 subclade C2 groups

Subclade C2 was structured into three Fastbaps clusters: 7 (n=362, named C2_7), 8 (n=86, C2_8) and 9 (n=113, C2_9) (Figure 3.2, Table 3.2). Most of the isolates in the National Collection (n=104) were represented by C2_8 (n=53, 51%), followed by C2_7 (n=17, 16%) and C2_9 (n=8, 8%). Within the global collection most isolates were C2_7 (n=345, 50%), with less in C2_8 (n=33, 5%) and C2_9 (n=105, 15%) (Figure 3.3). This showed C2_7 was more common globally than in Ireland (odds ratio = 3.1, $p<6.5x10^{-6}$), and C2_8 was more widespread in Ireland than elsewhere (odds ratio = 10.6, $p<2.2x10^{-16}$) (Figure 3.3).

This difference was paralleled by the rST results, which showed that rST1503 was highly predictive of C2_7 globally (319 out of 345, 92.5%) and in Ireland (16 out of 17, 94%). Similarly, rST1850 was highly associated with C2_8 in Ireland (n=45, 85%), but less so for the global collection (11 out of 33, 33%, Table 2). This limited resolution suggests rMLST (ribosomal Multilocus Sequence Typing) has insufficient discrimination to accurately reflect the evolutionary history of clonal pathogens like ST131, and that core genome analysis was more informative.

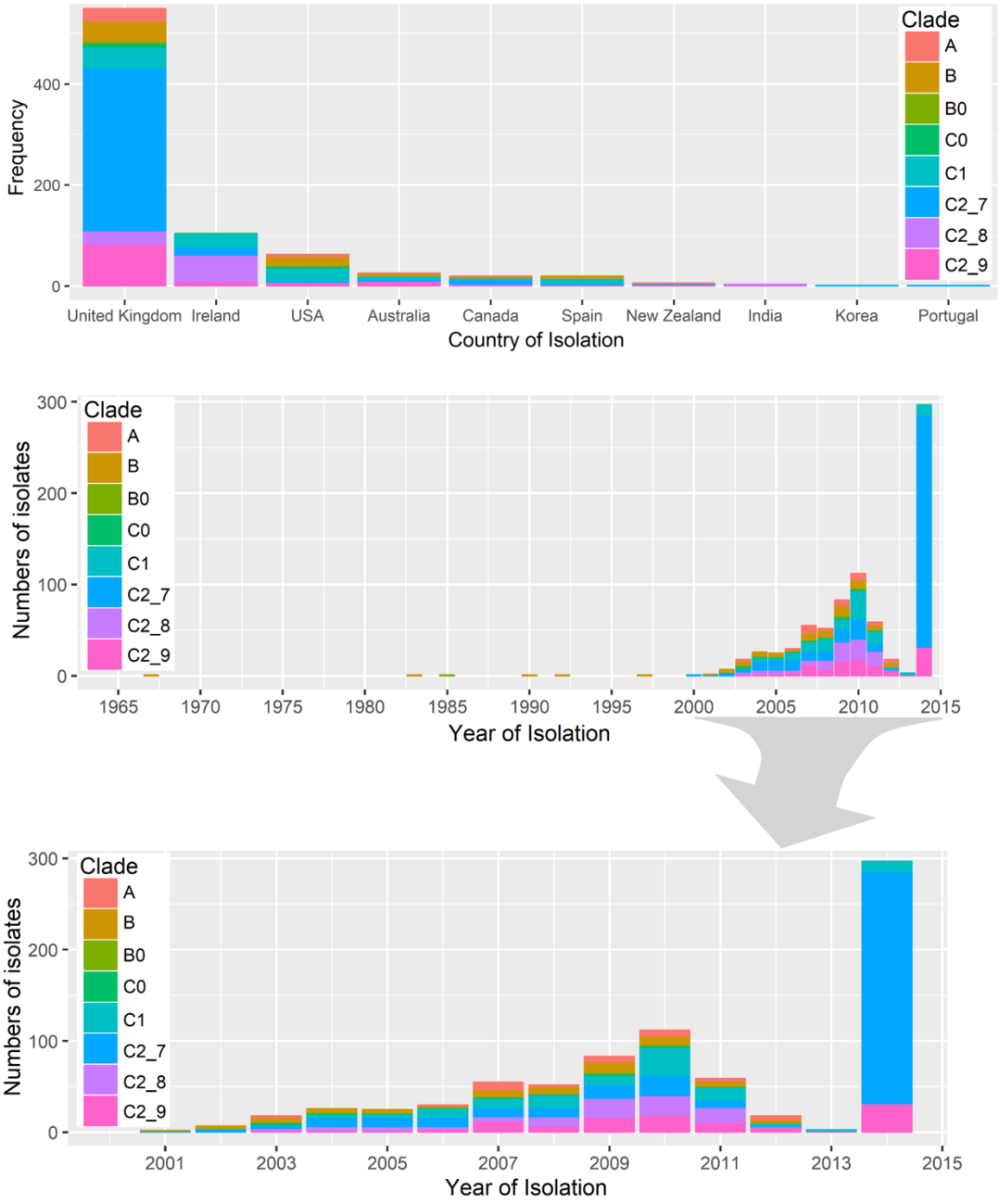


Figure 3.3. ST131 from the eight subclades (n=794) showed differing frequencies across country of origin (top) and year of isolation (middle and bottom). The subclades were A (n=33), B (n=70), B0 (n=5), C0 (n=14), C1 (n=111), C2_7 (n=362), C2_8 (n=86) and C2_9 (n=113). The ST131 were sampled during 1967-2014. The figures were generated using the ggplot2 and ggjoy packages in R v.3.5.2.

	Subclade	C0	C1	C2_7	C2_8	C2_9	Totals
	Fastbaps	2-5,11	10	7	8	9	
National collection	rST1503	1	24	16	1	7	49
	rST1850				45		45
	Other rSTs		1	1	7	1	10
	Total	1	25	17	53	8	104
Global collection	rST1503	11	82	319	21	101	535
	rST1850				11		11
	Other rSTs	2	4	26	1	4	37
	Total	13	86	345	33	105	582
	Totals	14	111	362	86	113	686

Table 3.2. The entire ST131 set (n=794) was largely composed of isolates from clade C (n=686, 86% of total) that was categorised into five subclades by Fastbaps clustering: C0 (n=14, clusters 2-5 and 11), C1 (n=111, cluster 10), C2_7 (n=362, cluster 7), C2_8 (n=86, cluster 8) and C2_9 (n=113, cluster 9). The National (n=104) and global (n=690) ST131 had two main ribosomal sequence types (rSTs): rST1850 associated with the Irish C2_8 LTCF set (85%), and rST1503 that often corresponded to C2_7 (92.5%). Fastbaps clusters 2, 3, 4 and 5 in C0 represented one isolate each – only cluster 3 was *bla*$_{CTX\text{-}M\text{-}15}$-positive.

3.3.4 Long read sequencing uncovers chromosomal transposition of blaCTX-M genes

Five isolates from the Irish collection were selected for long-read sequencing to more accurately determine the location and genomic environment of the *bla*$_{CTX\text{-}M\text{-}14}$ and *bla*$_{CTX\text{-}M\text{-}15}$ genes. Four of five samples selected were *bla*$_{CTX\text{-}M\text{-}15}$ positive and members of Clade C2, of which three belonged to the predominant LTCF subclade (C2_8) and one from the predominant global clade (C2_7). The remaining long-read sequenced isolate was from Clade C1 and was *bla*$_{CTX\text{-}M\text{-}14}$-positive. Each of the PacBio assemblies were used as references for Illumina read mapping for the collection of 794 isolates. The three C2_8 PacBio genomes (ERR191646, ERR191657, ERR191663) demonstrated chromosomal insertion of a 2,971 bp ISEcp*1*-*bla*$_{CTX\text{-}M\text{-}15}$-orf477Δ-*Tn2* transposon unit (TU) (Supplementary Figure 3.1), similar to integration sites described previously (Johnson et al. 2010, Johnson et al. 2017). This TU was transposed into the 1,617 bp *mppA* gene

(encoding murein peptide permease A), which was split into 327 bp and 1290 bp segments (at NCTC13441 genome coordinates 2,522,100-2,523,713 bp). No direct repeats flanking the *bla*$_{CTX\text{-}M\text{-}15}$ element were observed. The *bla*$_{CTX\text{-}M\text{-}15}$ was separated upstream by a 48 bp spacer sequence from a fragmented ISEcp*1* upstream adjacent to IS26, and downstream *bla*$_{CTX\text{-}M\text{-}15}$ was separated by 46 bp spacer from an orf477 segment, which was flanked by an incomplete Tn*2* and IS26 elements at the 3' and 5' ends (Supplementary Table 3.2, Supplementary Figure 3.1), suggestive of one-ended transposition or a deletion following transposition (Johnson et al. 2010, Johnson et al. 2017). The fourth assembly from C2_7 (ERR191697) contained a *bla*$_{CTX\text{-}M\text{-}15}$ gene on an IncFII/FIA plasmid with an incomplete Tn*2* element and a fragmented ISEcp*1* (p_*bla*$_{CTX\text{-}M\text{-}15}$-orf477Δ-Tn*2*) flanked by IS*26* elements (Supplementary Table 3.2, Supplementary Figure 3.1). The fifth assembly was from C1 (ERR191724) and had a *bla*$_{CTX\text{-}M\text{-}14}$-positive pV130-like IncFII plasmid (100% identity) with an intact ISEcp*1* at the 5' end and an incomplete copy of IS*903B* at the 3' end (p_*ISEcp1-bla*$_{CTX}$-M-14-IS*903B*) (Supplementary Table 3.2, Supplementary Figure 3.1).

3.3.5 Genomic context of blaCTX-M-15 the Irish collection highlight genetically diverse C subclades

Our findings indicated that the chromosomal *bla*$_{CTX}$-M-15 TU inserted into the chromosome was a potentially unique characteristic of the Irish LTCF C2_8 isolates, in contrast to the plasmid-associated *bla*$_{CTX\text{-}M\text{-}15}$ in other C2 isolates, and plasmid-associated *bla*$_{CTX\text{-}M\text{-}14}$ in C1 identified by the PacBio sequencing (Supplementary Figure 3.1). This was tested in 54 Clade C isolates from the Irish LTCF by resolving the exact genomic architecture of regions with the *bla*$_{CTX\text{-}M}$ by genome assembly and mapping reads to construct a phylogeny (Supplementary Figure 3.2). Assemblies of the 54 were compared with the PacBio references and NCTC13441 (ERR718783) and the *bla*$_{CTX\text{-}M}$, IS*Ecp1*, Tn*2*, IS*903B*, and *mppA* copy numbers were inferred from read mapping distributions, including verification of reads spanning the genetic elements and TU boundaries (Supplementary Figure 3.3). Of the 54, 38 were *bla*$_{CTX\text{-}M\text{-}15}$-positive (all C2), nine were *bla*$_{CTX\text{-}M\text{-}14}$-positive (all C1), five had no *bla*$_{CTX\text{-}M}$ gene (n=3 from C2, n=2 from C1), and two had both *bla*$_{CTX\text{-}M\text{-}15}$ and *bla*$_{CTX\text{-}M\text{-}14}$ genes (ERR191646 and ERR191657 from C2_8) (Supplementary Figure 3). C2_8 isolates (n=29) had a chromosomal insertion of *bla*$_{CTX}$-M-15 (Supplementary

Figure 3.4), contrasting with C2_7 (n=9) that typically had a fragmented IS*Ecp1* with a plasmid-associated *bla*$_{CTX\text{-}M\text{-}15}$ gene like the C2_8 and C2_7 PacBio reference strains references (Supplementary Figure 3.5). The C2_9 (n=5) isolates had a plasmid-bound *bla*$_{CTX\text{-}M\text{-}15}$ gene adjacent to a 496 bp IS*Ecp1* fragment (p_shortIS*Ecp1*-*bla*$_{CTX\text{-}M\text{-}15}$-orf477Δ-Tn*2*, Supplementary Figure 3.5). Like the PacBio C1 assembly above, the C1 (n=11) isolates had a plasmid-associated IS*Ecp1*-*bla*$_{CTX\text{-}M\text{-}14}$-IS*903B* TU with three IS*Ecp1* copies along with a duplicated *bla*$_{CTX\text{-}M\text{-}14}$ gene, though two were *bla*$_{CTX\text{-}M}$-negative.

Examining the rest of the collection in the same way showed that the *mppA* TU insertion was unique to the 41 Irish LTCF isolates in Clade C2_8 and this mutation was not found among any of the other 63 isolates from Ireland either in LTCF, community or hospitals. This is consistent with a pattern of clonal expansion in the LTCF. Of the 690 global isolates, 11 of the 19 with a disrupted *mppA* gene were *bla*$_{CTX\text{-}M\text{-}15}$-positive and clustered within the clonally expanded C2_8 *mppA*-insertion lineage. The remaining eight were independent events: six had no *bla*$_{CTX\text{-}M}$ gene and one had a *bla*$_{CTX\text{-}M\text{-}19}$ gene. Across all 794, C2 had a high rate of *bla*$_{CTX\text{-}M\text{-}15}$-positives isolates, reiterating the correlation of *bla*$_{CTX\text{-}M\text{-}15}$ with the expansion of C2, with incidences of 84% in C2_7 isolates (303 out of 362), 83% in C2_8 (71 out of 86), and 67% in C2_9 (76 out of 113).

3.3.6 Time of origin of the ST131 clones

The estimated time of the most recent common ancestor (TMRCA) of different phylogenetic groups was investigated with BEAST. We estimated a mutation rate of 4.14×10^{-7} SNPs/site/year (95% highest posterior density [HPD] intervals $3.74\text{-}4.57 \times 10^{-7}$), equivalent to 1.858 mutations/genome/year. A dated phylogeny (Figure 3.4) of all 794 isolates estimated a TMRCA for ST131 of around 1901 (95% HPD intervals 1842-1948). Clade C originated in 1985 (95% HPD 1980-1989). The FQ-R C1/C2 ancestor originated in 1992 (95% HPD 1989-1994), more recently than previous estimates of 1987 (12) and 1986 (Kallonen et al. 2017). Following this event, C1 and C2 diversified in parallel around 1994 (95% HPD 1991-1996, 95% HPD 1992-1995, respectively). C2 is composed of divergent subclades C2_7, C2_8 and C2_9. C2_7 diversified from the C2_8/9 lineage in 1995 (95% HPD 1993-1997). Finally, a group of strains radiated within C2_8 and formed a “displacement clade” (Figure 3.4A).

The TMRCA of the C2_8 clade was estimated at around 2003 (95% HPD 2001-2005) and all isolates in this clade contained a chromosomal *bla*$_{CTX\text{-}M\text{-}15}$ inserted between a truncated *mppA* gene. The "displacement clade" within the C2_8 subcluster comprised of 41 Irish LTCF isolates from the local collection which had a unique TU insertion in the *mppA.* This is in addition to ten other Irish isolates from clinical or community sources (n=51 in total) with a mutant *mppA* but showed a different TU insertion. The 11 *bla*$_{CTX\text{-}M\text{-}15}$-positive isolates that clustered with the Irish C2_8 isolates also had a disrupted *mppA* gene and were from the UK (n=8) and Canada (n=3). Together, these 62 shared a TMRCA of around 1998 (95% HPD 1997-2001), indicating that the *mppA* insertion may have occurred in the ancestral branch dating to 1996-1998 in the UK or North America (Figure 3.4B).

This evidence highlighted a single genetic origin of the ancestral C2_8 lineage in the Irish LTCF (Figure 3.4), though it was rare until 2009 (Table 3.3), potentially presenting opportunities for multiple introductions of C2_8. Prior to 2008, C1_10 was most common, consistent with a pattern of replacement by C2_8 with the mutant *mppA* insertion that clonally expanded. Nine out of 12 isolates from this facility detected between 2005-2007 belonged to C1_10. This was the group of isolates corresponding with the outbreak identified in 2006. Conversely, in the 57 samples isolated from 2008-2011, all were *bla*$_{CTX\text{-}M\text{-}15}$ positive, and 36 were classified in C2_8 (four of which were also both *bla*$_{CTX\text{-}M\text{-}14}$-positive), seven in C2_9 and eight in C2_7. In the global isolates, in contrast to the LTCF, C2_8 accounted for only 5% of isolates, whereas C1_10 and C2_7 accounted for 14% and 49% (respectively) with no evidence of this clonal displacement outside of the LTCF.

Clade	C0	C1_10	C2_7	C2_8	C2_9	Total
2005			1			1
2006	1	7				8
2007		2		1		3
2008		2		2		4
2009				12		12
2010		6	5	17	7	35
2011			2	4		6
Total	1	17	8	36	7	69

Table 3.3. The numbers of isolates from the LTCF in Ireland (n=69) across the ST131 clades showed that C1_10 was most common at the outset of the study, and that C2_8 became more prevalent after 2008, suggesting a possible replacement and clonal expansion of this lineage.

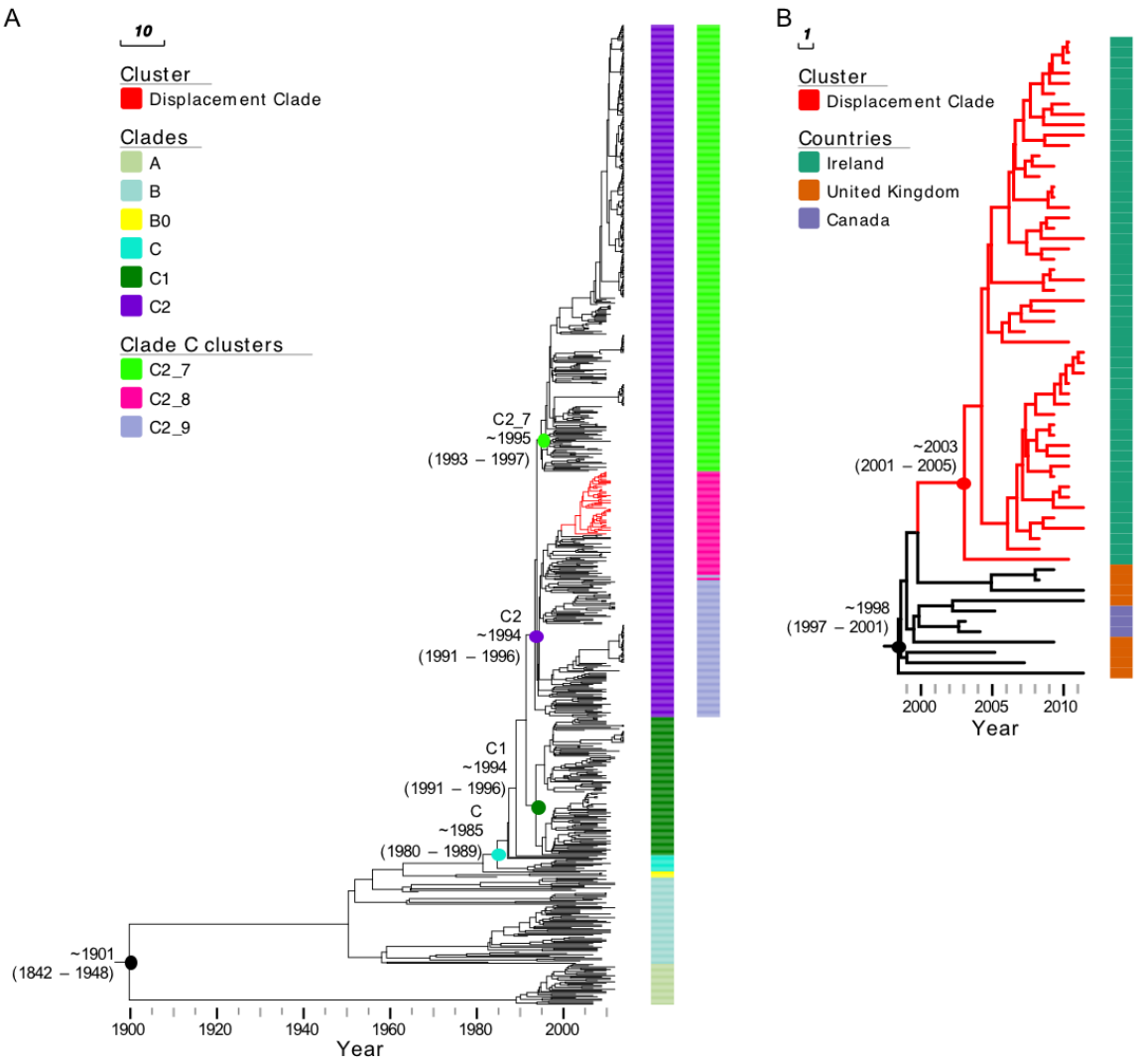


Figure 3.4. Bayesian maximum clade credibility tree of E. coli ST131 isolates. (A) Phylogeny of 794 isolates analysed in this study. The tree is annotated with column representing major phylogenetic clades (Clades) as well as subclades within clade C (clade C clusters). The estimated TMRCA for major clades is shown on the tree. Branches of the cluster representing isolates from the Irish LTCF displacement clone are coloured in red. (B) A higher resolution view of the Irish LTCF displacement clone, annotated with colour strips representing isolate's country of origin.

3.4 Discussion

Here, we traced the genomic background of ESBL-E. coli ST131 isolates collected from residents of a LTCF in Ireland where an outbreak was recognised in 2006. The relationship between the isolates was first identified based on indistinguishable pulsed field gel electrophoresis (PFGE) patterns among 18 patients (Pelly et al. 2006). Since the outbreak was detected in 2006, there has been extensive progress in the higher discriminatory power of genome-sequencing compared to PFGE and other typing tools, such as MLST (Salipante et al. 2015, Rumore et al. 2018, Ludden et al. 2019). To gain a further understanding of the origins of the outbreak and to observe changes in E. coli population structure in LTCF residents, we performed whole genome sequencing of all ST131 ESBL-E. coli isolates submitted from the LTCF over seven years. We compared these to 35 other ST131 isolated in Ireland; 9 from other LTCFs, 2 from the community and 24 from hospitals including 14 from the referral hospital and 10 from 3 other hospitals, in addition to 690 ST131 from global datasets.

We identified distinct genetic clusters within this set of 794 closely related isolates based on core genome phylogenetic signals, and as in previous studies (Stoesser et al. 2016), we identified subclade C2 as the most abundant ST131 group accounting for 71% of the entire collection. Four genetic subgroups were common in the specific LTCF, one from subclade C1 (C1_10) and three from C2 (C2_7, C2_8, C2_9). The resident ST131 lineage (C1_10) in the LTCF in the period 2005-2007 was the cause of the initial outbreak investigation, but surprisingly a newly introduced ST131 variant (C2_8) was much more common by 2009, indicating displacement of blaCTX-M-14-positive C1 isolates and clonal expansion by a genetically distinct blaCTX-M-15-positive C2 lineage within the LTCF. This pattern of clonal displacement has not yet been published for E. coli, but is common in other species such as methicillin-resistant Staphylococcus aureus (MRSA) where it can be driven by inter-hospital transfer of patients (Hsu et al 2015).

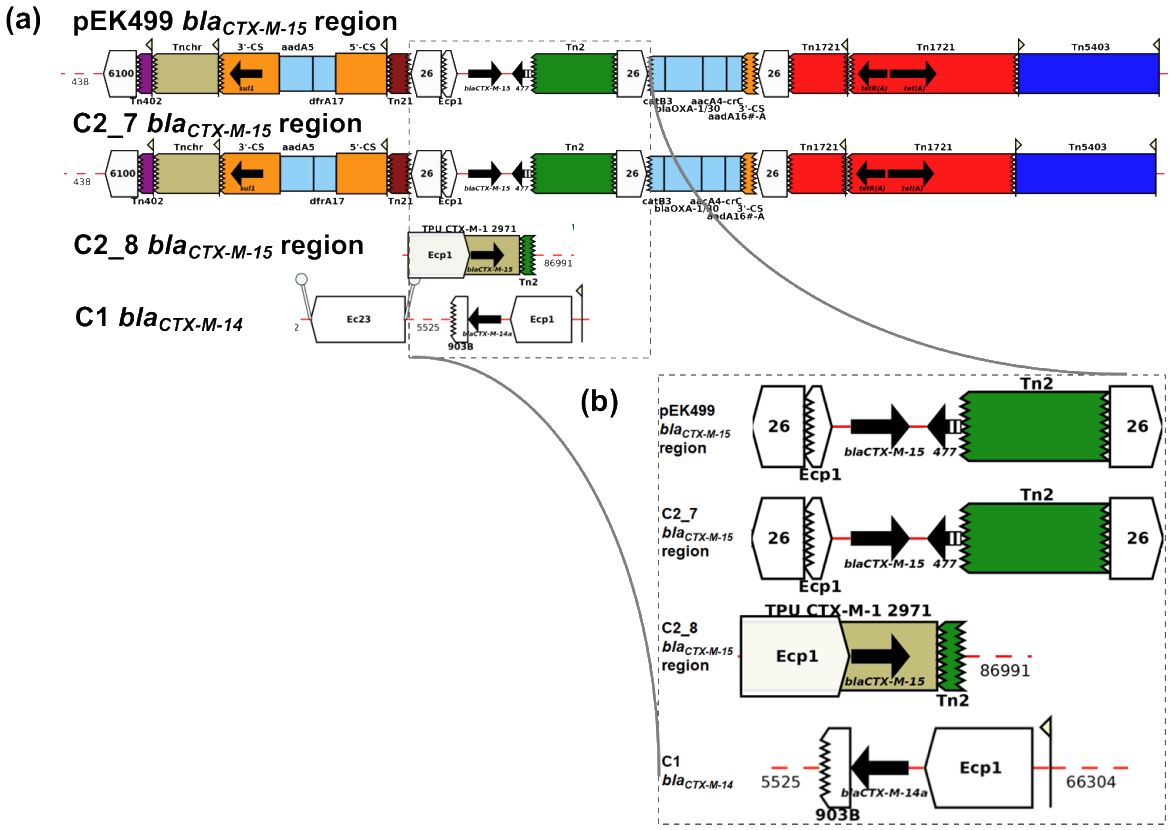
In this study, we analysed the largest global collection of whole genome data on ST131 E. coli and estimated the emergence of ST131 in approximately 1901. The clonal expansion of C2 in 1994 identified here was similar to Kallonen et al (2017) and Zakour et al. (2016), who reported 1990 and 1987, respectively. We dated the C2_8 LTCF lineage to have

emerged in 2001-2005 and we postulate that the clone originated in the UK or North America in 1996-1998. This was consistent with the first observation of blaCTX-M-15-positive cephalosporin-resistant E. coli isolated in 2001 in three locations in Britain and Northern Ireland (Mushtaq et al. 2003, Livermore et al. 2003). However, C2_8 was generally not as successful as C2_7, which emerged around the same time (1995) and disseminated globally. It has been suggested that the evolution of C2 subclades has been shaped by the acquisition of IncFII plasmids encoding blaCTX-M-15 (Stoesser et al. 2016), which was also observed here for C2_7. We extend this by showing that blaCTX-M-15 in C2_8 was mobilized from IncFII plasmids by ISEcp1-mediated transposition to the chromosome at mppA in a TU structured as ISEcp1-blaCTX-M-15-orf477Δ-Tn2. The high copy number and fragmented pattern of ISEcp1, which enabled a chromosomal insertion, was found for different blaCTX-M alleles in E. coli and may be linked to altered expression of the gene on the chromosome relative the plasmid (Matsumura et al. 2017, Canton et al. 2012). Our work shows although ST131 is disseminated globally, evolutionary events have resulted in the clonal expansion of new lineages, such as C2_7 globally and C2_8 locally in one LTCF. This has coincided not only with the horizontal gene transfer of plasmids encoding blaCTX-M-15 or blaCTX-M-14, but also the chromosomal insertions like blaCTX-M-15 in C2_8 followed by vertical transmission, and also blaCTX-M-14 5' of the chromosomal rlmL gene in one C1 isolate (ERR191666).

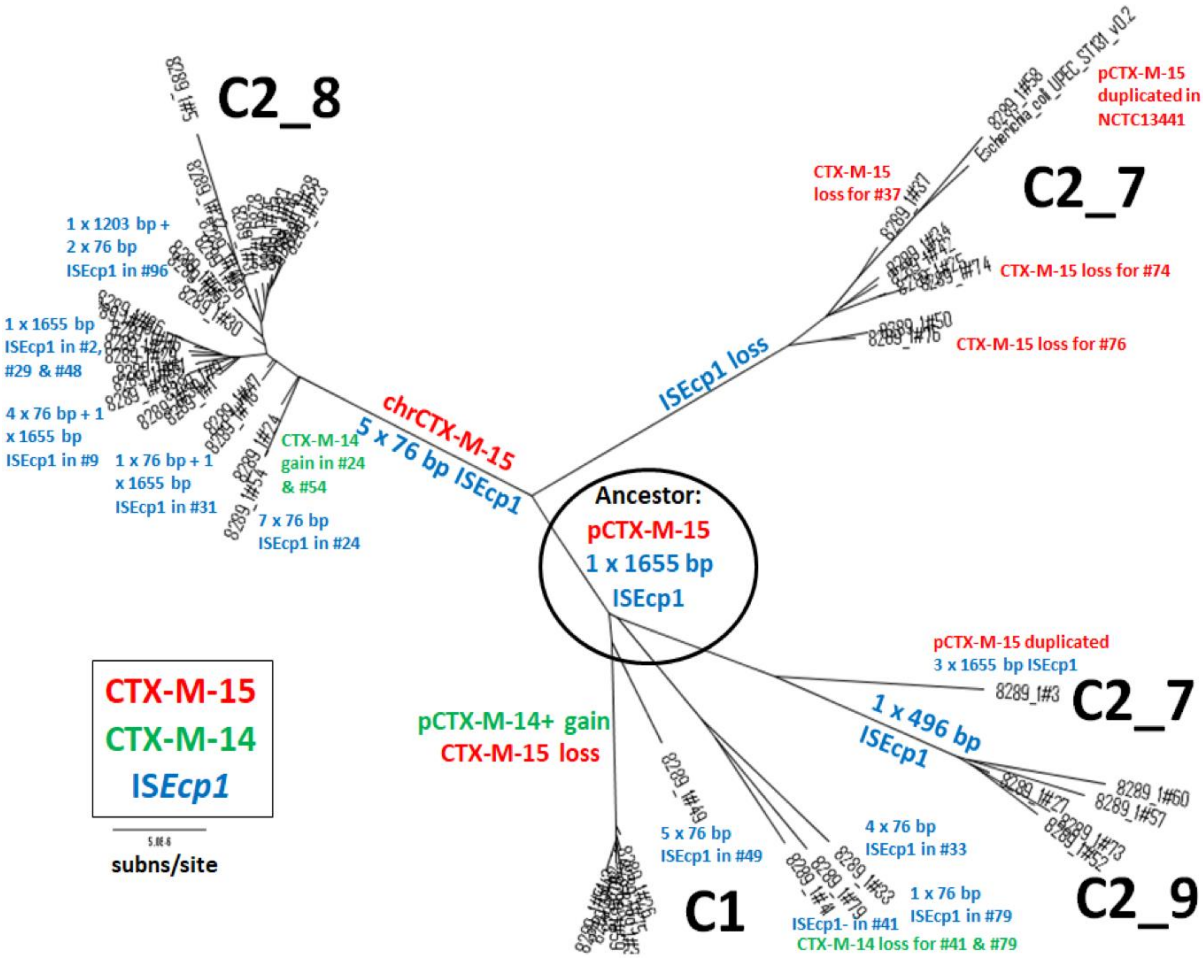
In conclusion, we investigated an outbreak of ESBL E. coli ST131 in a LTCF in Ireland and observed changes in this LTCF different to the global pattern. We found that the outbreak began with a Clade C1 strain encoding blaCTX-M-14 gene on a plasmid, and that this lineage was displaced by a Clade C2 strain with a chromosomally-encoded blaCTX-M-15 gene. Both lineages associated with the LTCF are resistant to broad-spectrum cephalosporins and the selective forces in this specific niche driving lineage displacement are unclear.

3.5 Supplementary Tables and Figures

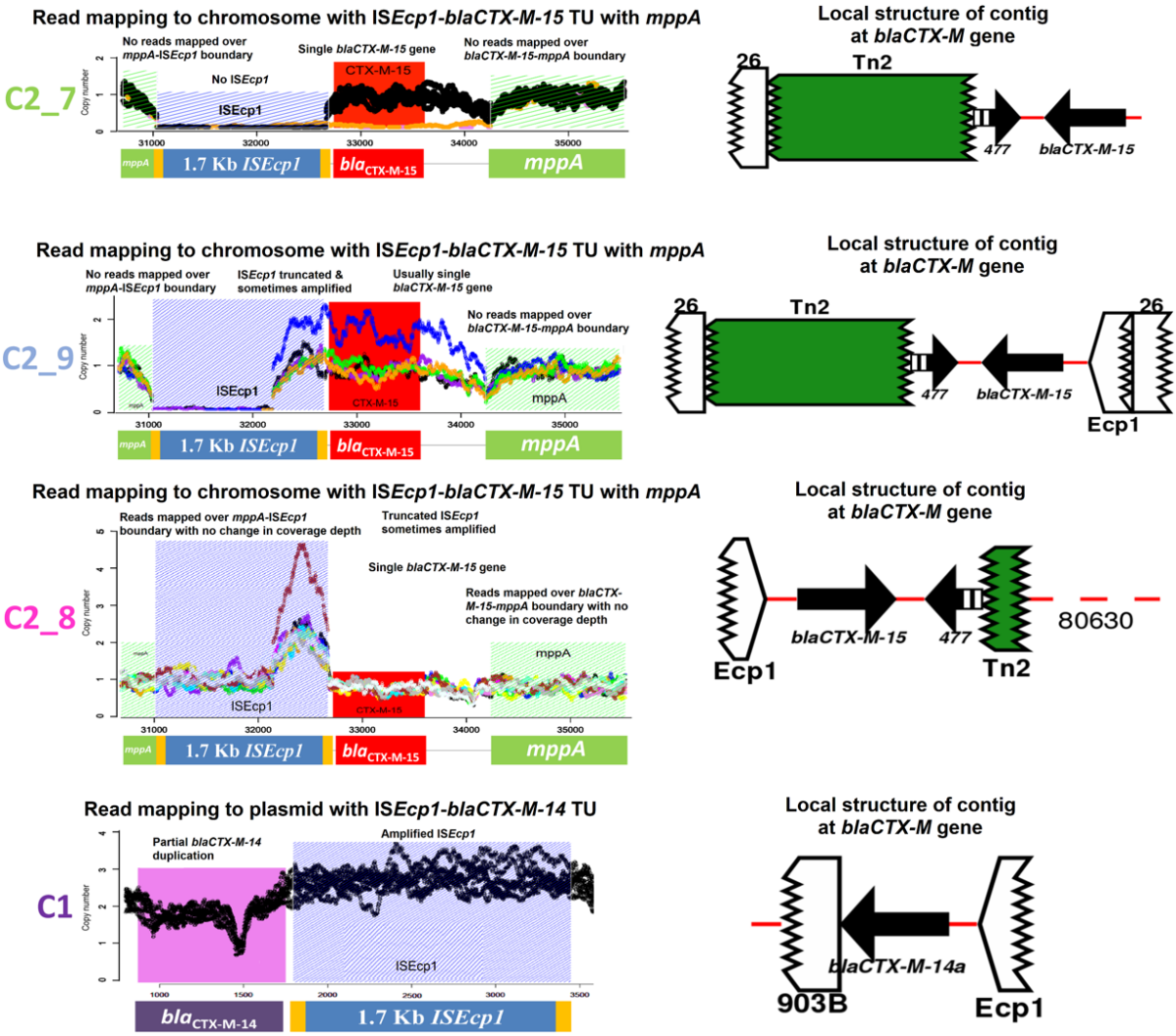
Supplementary tables for Chapter 3 are publicly available on Figshare: https://figshare.com/s/0fcc2204056feffafb39.



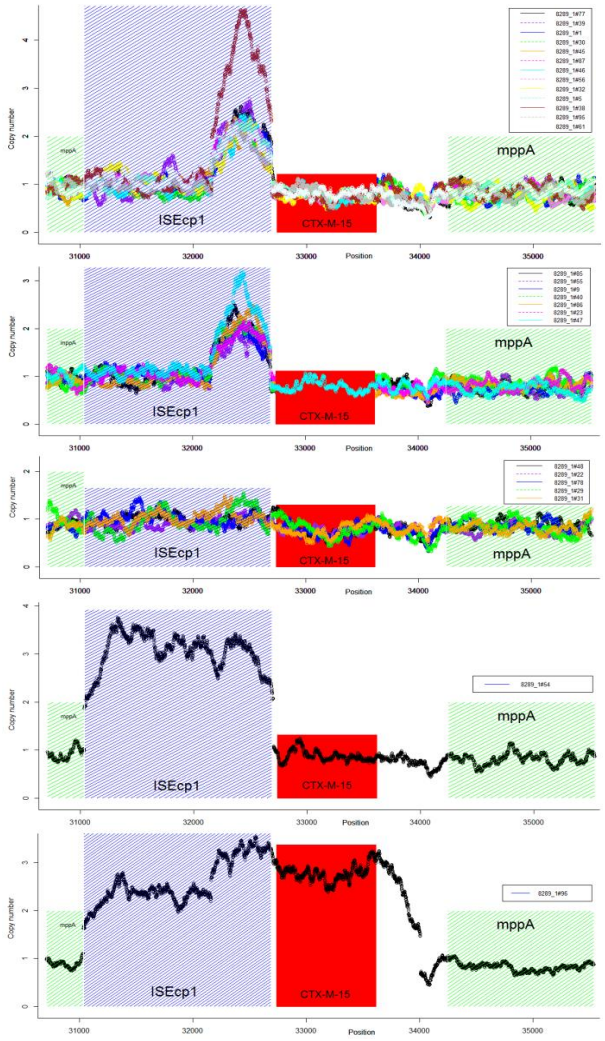
Supplementary figure 3.1. (a) *bla*$_{CTX\text{-}M\text{-}15}$ element on the IncFII/FIA pEK499 reference, (b) *bla*$_{CTX\text{-}M\text{-}15}$ element on the IncFII/FIA subclade C2_7, (c) *bla*$_{CTX\text{-}M\text{-}15}$ chromosomal element in C2_8 isolates. All Clade C2_8 PacBio genomes contained the same *bla*$_{CTX\text{-}M\text{-}15}$ chromosomal element, therefore only one is shown (d) a *bla*$_{CTX\text{-}M\text{-}14}$ element on IncFII subclade C1 plasmid. The region encoding *bla*$_{CTX\text{-}M\text{-}14}$ and *bla*$_{CTX\text{-}M\text{-}15}$ are highlighted in a black box. Arrows indicate the orientation of features, with the forward direction defined as the direction of transcription for genes, towards the main part of the attC site for cassettes, in integrons towards attI for 5' flanking regions, away from the cassette array for 3'-flanking regions, relative to the direction of transcription of the transposase gene for insertion sequences and transposons (Tn) (ie, inverted repeat left to inverted repeat right) and to the direction of the reverse transcriptase for Group II introns. The missing end of a feature is shown by a zig-zag line. The inset shows the area bounded by the dashed line in more detail.



Supplementary figure 3.2. A phylogenomic network of the 54 Irish Clade C samples' chromosomal mutational SNPs built using RAxML and ClonalframeML and drawn with FigTree v.1.4.3. The phylogeny of the n=54 was rooted using the topology of n=794, which C1 as the most divergent lineage, with C2_9 diverging next, followed by C2_8 and C2_7, though here the smaller sample size meant that the ancestral lineage was unclear and so could be approximated by the C1-C2 origin, where the ancestor likely had a plasmid with a 1,655 bp IS*Ecp1* 5' of a *bla*$_{CTX\text{-}M\text{-}15}$ gene. The *bla*$_{CTX\text{-}M\text{-}15}$ gene changes are in red, the *bla*$_{CTX\text{-}M\text{-}14}$ gene mutations are in green, and the IS*Ecp1* differences are in blue. The subclades C1, C2_7, C2_8 and C2_9 are shown, and the scale bar shows five substitutions per Mb.



Supplementary Figure 3.3. Read mapping copy numbers for 33 of the 54 the Irish isolates from C2_7 (n=6 shown), C2_9 (all n=5 shown), C2_8 (n=15 shown) and C1 (n=7 shown) across IS*Ecp1* elements (blue), the *bla*$_{CTX\text{-}M\text{-}15}$ gene (red), the *bla*$_{CTX\text{-}M\text{-}14}$ gene (pink) or the *mppA* gene (green). C2_8 all had consistent coverage of the chromosomally inserted TU isoform IS*Ecp1*-*bla*$_{CTX\text{-}M\text{-}15}$-shortTn*2* spanning *mppA* and typically had with IS*Ecp1* fragments of 1,203 bp, 529 bp, 76 bp and 76 bp. The C2_9 isolates had a 496 bp IS*Ecp1* element and a *bla*$_{CTX\text{-}M\text{-}15}$ gene. Most C2_7 isolates had no IS*Ecp1* and one *bla*$_{CTX\text{-}M\text{-}15}$ gene. C1 had the TU isoform p_IS*Ecp1*-*bla*$_{CTX\text{-}M\text{-}14}$-IS*903B* with a duplicated IS*Ecp1* element and duplicated *bla*$_{CTX\text{-}M\text{-}14}$ gene. Reads from non-C2_8 libraries mapped at *mppA* in the TU isoform, but with gaps indicating no contiguous mapping to the *bla*$_{CTX\text{-}M\text{-}15}$ gene.



Strain	ISEcp1-1203bp			ISEcp1-529bp			ISEcp1-76bp			CTM15	
#77	2515747	2516949	101	2517050	2517578	102	2517680	2517755	50	2517805	2518681
#39	2463781	2464983	101	2465084	2465612	102	2465714	2465789	50	2465839	2466715
#1	2517655	2518857	101	2518958	2519486	102	2519588	2519663	50	2519713	2520589
#30	2522218	2523420	101	2523521	2524049	102	2524151	2524226	50	2524276	2525152
#45	2507955	2509157	101	2509258	2509786	102	2509888	2509963	50	2510013	2510889
#87	2513858	2515060	101	2515161	2515689	102	2515791	2515866	50	2515916	2516792
#46	2486010	2487212	101	2487313	2487841	102	2487943	2488018	50	2488068	2488944
#56	2493709	2494911	101	2495012	2495540	102	2495642	2495717	50	2495767	2496643
#32	2492145	2493347	101	2493448	2493976	102	2494078	2494153	50	2494203	2495079
#5	2483954	2485156	101	2485257	2485785	102	2485887	2485962	57725	2543687	2544563
#38	2514489	2515691	101	2515792	2516320	27242	2543562	2543637	50	2543687	2544563
#95	2513089	2514291	101	2514392	2514920	32815	2547735	2547810	50	2547860	2548736
#61	2515094	2516296	101	2516397	2516925	6835	2523760	2523835	50	2523885	2524761

Strain	ISEcp1-1203bp			ISEcp1-76bp			CTM15	
#85	2514846	2516048	101	2516149	2516224	50	2516274	2517150
#55	2512460	2513662	101	2513763	2513838	50	2513888	2514764
#9	2516536	2517738	101	2517839	2517914	50	2517964	2518840
#40	2515338	2516540	101	2516641	2516716	50	2516766	2517642
#86	2514655	2515857	101	2515958	2516033	50	2516083	2516959
#23	2513116	2514318	101	2514419	2514494	50	2514544	2515420
#47	2507474	2508676	101	2508777	2508852	50	2508902	2509778

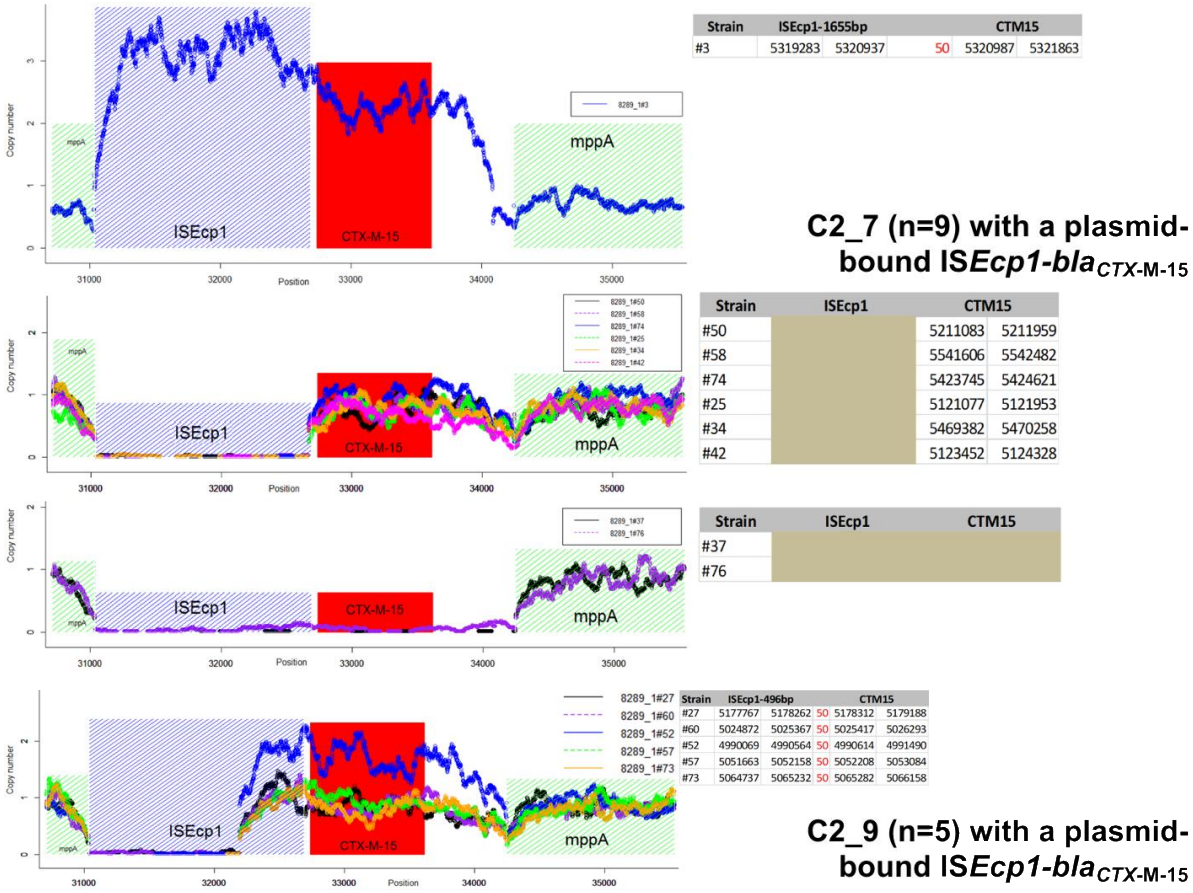
Strain	ISEcp1-1655bp			CTM15	
#48	2504370	2506024	50	2506074	2506950
#22	2495270	2496924	50	2496974	2497850
#78	2524663	2526317	50	2526367	2527243
#29	2499422	2501076	50	2501126	2502002
#31	2504151	2505805	50	2505855	2506731

Strain	ISEcp1-76bp			ISEcp1-74bp			CTM15	
#54	2514328	2514403	101	2514504	2514577	50	2514627	2515503

Strain	ISEcp1-1203bp			ISEcp1-529bp			CTM15	
#96	2513914	2515116	2584923	5100039	5100567	50	5100617	5101493

Supplementary Figure 3.4. C2_8 (n=27) with a chromosomal ISEcp1-blaCTX-M-15 at the mppA gene. Read mapping copy number for n=27 C2_8 isolates showing that all had 1+ IS*Ecp1* elements (blue) 5' of a *bla*$_{CTX\text{-}M\text{-}15}$ gene (red) with consistent coverage spanning the *mppA* gene (green) including reads spanning each elements' breakpoints, with the TU isoform chr_shortIS*Ecp1*-*bla*$_{CTX\text{-}M\text{-}15}$-shortTn*2*. Most (n=13, top panel) had IS*Ecp1* fragments of 1,203 bp, 529 bp, 76 bp and 76 bp with one *bla*$_{CTX\text{-}M\text{-}15}$ gene, though within this some had one or two extra 76 bp IS*Ecp1* fragments on their plasmids, including 8289_1#24 (ERR191657 in Table 3, not shown here) and 8289_1#53 (not shown here). 8289_1#24 had a *bla*$_{CTX\text{-}M\text{-}14}$ gene, and had 74-77 bp IS*Ecp1* fragments on its plasmid, as well as 76 bp, 529 bp and 1203 bp IS*Ecp1* segments on its chromosome. 8289_1#5 (ERR191638) also had its *bla*$_{CTX\text{-}M\text{-}15}$ gene was 57,725 bp distant from the IS*Ecp1* fragments. 8289_1#38 (ERR191671), 8289_1#61 (ERR191694) and 8289_1#95 (ERR191728) had a 76 bp IS*Ecp1* fragment adjacent to the chromosomal *bla*$_{CTX\text{-}M\text{-}15}$ gene

at 27,742 bp (8289_1#38), 6,835 bp (8289_1#61) and 32,815 (8289_1#95) from the other IS*Ecp1* copies, consistent with recombination between IS*Ecp1* segments. A minority (n=7, second panel) were like this previous group, but without the 529 bp IS*Ecp1* fragment. Another set (n=5, third panel) had a full 1,655 bp IS*Ecp1* element with no fragmentation. One isolate (8289_1#54, fourth panel) had three full 1,655 bp IS*Ecp1* elements on both its chromosome and plasmid, and fragments of 76 bp and 74 bp adjacent to the *bla*$_{CTX\text{-}M\text{-}15}$ gene. One isolate (ERR1917 29, 8289_1#96, fifth panel) had three *bla*$_{CTX\text{-}M\text{-}15}$ gene copies and a first IS*Ecp1* fragment 529 bp where the TU was inverted and duplicated, and separate from a 1,203 bp IS*Ecp1* at mppA, suggesting that recombination between the chromosomal and plasmid IS*Ecp1* IRs may have transferred the *bla*$_{CTX\text{-}M\text{-}15}$ gene back to the plasmid. The table (right) shows the IS*Ecp1* assembly coordinates, spacer DNA lengths, and *bla*$_{CTX\text{-}M\text{-}14}$ assembly coordinates.



Supplementary Figure 3.5. Read mapping copy number for n=9 C2_7 and n=5 C2_9 isolates showing one from C2_7 (8289_1#3, ERR191636, top) had a three IS*Ecp1* copies (blue) and a duplicated *bla*$_{CTX\text{-}M\text{-}15}$ gene (red), but no contiguity if reads mapping across to *mppA* (green) unlike C2_8. These had the TU isoform p_*bla*$_{CTX\text{-}M\text{-}15}$-orf477Δ-Tn*2*. The majority of C2_7 (n=6, second diagram) had no IS*Ecp1* and one *bla*$_{CTX\text{-}M\text{-}15}$ gene. Two C2_7 isolates (third diagram) had no IS*Ecp1* and no *bla*$_{CTX\text{-}M\text{-}14}$ gene. All n=5 C2_9 isolates (bottom diagram) had a 496 bp IS*Ecp1* element (blue) 5' of a *bla*$_{CTX\text{-}M\text{-}15}$ gene (red), but no contiguity if reads mapping across to *mppA* (green) unlike C2_8. One (8289_1#52, blue) had a partial amplification of this TU. The table (right) shows the IS*Ecp1* assembly coordinates, 50 bp spacer length and *bla*$_{CTX\text{-}M\text{-}14}$ assembly coordinates. ERR191636 from C2_7 had no homology to pV130 or pEK499, and the assemblies of eight other isolates from C2_7 (n=3), C2_8 (n=4) and C2_9 (n=1) had no detected pV130 or pEK499 homology.

3.6 References

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. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. Journal of computational biology : a journal of computational molecular cell biology 19:455-477.

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.

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

Boetzer M, Pirovano W. 2012. Toward almost closed genomes with GapFiller. Genome Biology 13:R56.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114-2120.

Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu C-H, Xie D, Suchard MA, Rambaut A, Drummond AJ. 2014. BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. PLOS Computational Biology 10:e1003537.

Brodrick HJ, Raven KE, Kallonen T, Jamrozy D, Blane B, Brown NM, Martin V, Török ME, Parkhill J, Peacock SJ. 2017. Longitudinal genomic surveillance of multidrug-resistant Escherichia coli carriage in a long-term care facility in the United Kingdom. Genome Medicine 9:70.

Burgess MJ, Johnson JR, Porter SB, Johnston B, Clabots C, Lahr BD, Uhl JR, Banerjee R. 2015. Long-term care facilities are reservoirs for antimicrobial-resistant sequence type 131 Escherichia coli. Open Forum Infectious Diseases 2:ofv011.

Burke L, Humphreys H, Fitzgerald-Hughes D. 2012. The revolving door between hospital and community: extended-spectrum beta-lactamase-producing Escherichia coli in Dublin. Journal of Hospital Infection 81:192-198.

Burns K, Foley M, Donlon S. 2012. Point prevalence survey of hospital acquired infections & antimicrobial use in european acute care hospitals: May 2012 Republic of Ireland National Report. HPSC Ireland, Dublin.

Canton R, Gonzalez-Alba JM, Galán JC. 2012. CTX-M Enzymes: Origin and Diffusion. Frontiers in Microbiology 3.

Carattoli A, Bertini A, Villa L, Falbo V, Hopkins KL, Threlfall EJ. 2005. Identification of plasmids by PCR-based replicon typing. Journal of Microbiological Methods 63:219-228.

Chin C-S, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A, Copeland A, Huddleston J, Eichler EE, Turner SW, Korlach J. 2013. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat Meth 10:563-569.

Croucher N. J., Page A. J., Connor T. R., Delaney A. J., Keane J. A., Bentley S. D., Parkhill J., Harris S.R. 2014. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. doi:10.1093/nar/gku1196, Nucleic Acids Research, 2014.

Darling AE, Mau B, Perna NT. 2010. progressiveMauve: Multiple Genome Alignment with Gene Gain, Loss and Rearrangement. PLOS ONE 5:e11147.

Decano AG, Ludden C, Feltwell T, Judge K, Parkhill J, Downing T. 2019. Complete assembly of Escherichia coli sequence type 131 genomes using long reads demonstrates antibiotic resistance gene variation within diverse plasmid and chromosomal contexts. mSphere 4:e00130-19.

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

Gagliotti C, Balode A, Baquero F, Degener J, Grundmann H, Gür D, Jarlier V, Kahlmeter G, Monen J, Monnet DL, Rossolini GM, Suetens C, Weist K, Heuer O, AMR). tE-NPDSCPf. 2011. Escherichia coli and Staphylococcus aureus: bad news and good news from the European Antimicrobial Resistance Surveillance Network (EARS-Net, formerly EARSS), 2002 to 2009. Euro Surveill 16:19819.

Gladman S, Seemann, T, Victorian Bioinformatics Consortium. 2008. Velvet Optimiser: For automatically optimising the primary parameter options for the Velvet de novo sequence assembler.

Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: quality assessment tool for genome assemblies. Bioinformatics (Oxford, England) 29:1072-1075.

Hsu L-Y, Harris SR, Chlebowicz MA, Lindsay JA, Koh T-H, Krishnan P, Tan T-Y, Hon P-Y, Grubb WB, Bentley SD, Parkhill J, Peacock SJ, Holden MT. 2015. Evolutionary dynamics of methicillin-resistant Staphylococcus aureus within a healthcare system. Genome Biology 16:81.

Hull RA, Gill RE, Hsu P, Minshew BH, Falkow S. 1981. Construction and expression of recombinant plasmids encoding type 1 or D-mannose-resistant pili from a urinary tract infection Escherichia coli isolate. Infection and Immunity 33:933-938.

Hunt M, Silva ND, Otto TD, Parkhill J, Keane JA, Harris SR. 2015. Circlator: automated circularization of genome assemblies using long sequencing reads. Genome Biology 16:294.

Huson DH, Bryant D. 2005. Application of Phylogenetic Networks in Evolutionary Studies. Molecular Biology and Evolution 23:254-267.

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. Clinical Infectious Diseases 51:286-294.

Johnson JR, Porter S, Thuras P, Castanheira M. 2017. The pandemic H30 subclone of sequence type 131 (ST131) as the leading cause of multidrug-resistant Escherichia coli infections in the United States (2011–2012). Open Forum Infectious Diseases 4:ofx089.

Kallonen T, Brodrick HJ, Harris SR, Corander J, Brown NM, Martin V, Peacock SJ, Parkhill J. 2017. Systematic longitudinal survey of invasive Escherichia coli in England demonstrates a stable population structure only transiently disturbed by the emergence of ST131. Genome Research 27:1437-1449.

Letunic I, Bork P. 2016. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Research doi:10.1093/nar/gkw290.

Livermore DM, Mushtaq S, James D, Potz N, Walker RA, Charlett A, Warburton F, Johnson AP, Warner M, Henwood CJ. 2003. In vitro activity of piperacillin/tazobactam and other broad-spectrum antibiotics against bacteria from hospitalised patients in the British Isles. International Journal of Antimicrobial Agents 22:14-27.

Ludden C, Cormican M, Vellinga A, Johnson JR, Austin B, Morris D. 2015. Colonisation with ESBL-producing and carbapenemase-producing Enterobacteriaceae,

vancomycin-resistant enterococci, and meticillin-resistant Staphylococcus aureus in a long-term care facility over one year. BMC Infectious Diseases 15:168.

Ludden C, Raven K, 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 N, Holmes M, Parkhill J, Peacock S. 2019. One health genomic surveillance of Escherichia coli demonstrates distinct lineages and mobile genetic elements in isolates from humans versus livestock. mBio 10.

Matsumura Y, Pitout JDD, Peirano G, DeVinney R, Noguchi T, Yamamoto M, Gomi R, Matsuda T, Nakano S, Nagao M, Tanaka M, Ichiyama S. 2017. Rapid identification of different Escherichia coli sequence type 131 clades. Antimicrobial Agents and Chemotherapy 61:e00179-17.

Mushtaq S, Woodford N, Potz N, Livermore DM. 2003. Detection of CTX-M-15 extended-spectrum β-lactamase in the United Kingdom. Journal of Antimicrobial Chemotherapy 52:528-529.

Partridge SR, Tsafnat G, Coiera E, Iredell JR. 2009. Gene cassettes and cassette arrays in mobile resistance integrons. FEMS Microbiology Reviews 33:757-784.

Pelly H, Morris D, O'Connell E, Hanahoe B, Chambers C, Biernacka K, Gray S, Cormican M. 2006. Outbreak of extended spectrum beta-lactamase producing E. coli in a nursing home in Ireland, May 2006. Weekly releases (1997–2007) 11:3036.

Petty NK, Ben Zakour NL, Stanton-Cook M, Skippington E, Totsika M, Forde BM, Phan M-D, Gomes Moriel D, Peters KM, Davies M, Rogers BA, Dougan G, Rodriguez-Baño J, Pascual A, Pitout JDD, Upton M, Paterson DL, Walsh TR, Schembri MA, Beatson SA. 2014. Global dissemination of a multidrug resistant Escherichia coli clone. Proceedings of the National Academy of Sciences 111:5694-5699.

Poolman JT, Wacker M. 2016. Extraintestinal pathogenic Escherichia coli, a common human pathogen: Challenges for vaccine development and progress in the field. The Journal of Infectious Diseases 213:6-13.

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.

Pruitt KD, Tatusova T, Brown GR, Maglott DR. 2012. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. Nucleic Acids Research 40:D130-D135.

Public Health England. 2017. English Surveillance Programme for Antimicrobial Utilisation and Resistance (ESPAUR) Report 2017.

Rambaut A, Lam TT, Max Carvalho L, Pybus OG. 2016. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). Virus Evolution 2.

Roberts RR, Hota B, Ahmad I, Scott IIRD, Foster SD, Abbasi F, Schabowski S, Kampe LM, Ciavarella GG, Supino M, Naples J, Cordell R, Levy SB, Weinstein RA. 2009. Hospital and societal costs of antimicrobial-resistant infections in a Chicago teaching hospital: Implications for antibiotic stewardship. Clinical Infectious Diseases 49:1175-1184.

Rogers BA, Sidjabat HE, Paterson DL. 2011. Escherichia coli O25b-ST131: a pandemic, multiresistant, community-associated strain. Journal of Antimicrobial Chemotherapy 66:1-14.

Rottier WC, Ammerlaan HSM, Bonten MJM. 2012. Effects of confounders and intermediates on the association of bacteraemia caused by extended-spectrum β-lactamase-producing Enterobacteriaceae and patient outcome: a meta-analysis. Journal of Antimicrobial Chemotherapy 67:1311-1320.

Rumore J, Tschetter L, Kearney A, Kandar R, McCormick R, Walker M, Peterson C-L, Reimer A, Nadon C. 2018. Evaluation of whole-genome sequencing for outbreak detection of Verotoxigenic Escherichia coli O157:H7 from the Canadian perspective. BMC genomics 19:870-870.

Salipante SJ, SenGupta DJ, Cummings LA, Land TA, Hoogestraat DR, Cookson BT. 2015. Application of Whole-Genome Sequencing for Bacterial Strain Typing in Molecular Epidemiology. Journal of Clinical Microbiology 53:1072-1079.

Schwaber MJ, Carmeli Y. 2007. Mortality and delay in effective therapy associated with extended-spectrum β-lactamase production in Enterobacteriaceae bacteraemia: a systematic review and meta-analysis. Journal of Antimicrobial Chemotherapy 60:913-920.

Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics 30:2068-2069.

Sommer DD, Delcher AL, Salzberg SL, Pop M. 2007. Minimus: a fast, lightweight genome assembler. BMC Bioinformatics 8:64.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312-1313.

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, Group MMMI. 2016. Evolutionary history of the global emergence of the Escherichia coli epidemic clone ST131. mBio 7.

Suetens C. 2012. Healthcare-associated infections in European long-term care facilities: how big is the challenge? Euro Surveill 17:pii=20259.

Tchesnokova V, Riddell K, Scholes D, Johnson JR, Sokurenko EV. 2018. The uropathogenic Escherichia coli subclone sequence type 131-H30 is responsible for most antibiotic prescription errors at an urgent care clinic. Clinical Infectious Diseases 68:781-787.

Thaden JT, Fowler VG, Sexton DJ, Anderson DJ. 2016. Increasing Incidence of extended-spectrum β-lactamase-producing Escherichia coli in community hospitals throughout the Southeastern United States. Infection control and hospital epidemiology 37:49-54.

Tonkin-Hill, G., Lees, J.A., Bentley, S.D., Frost, S.D.W. & Corander, J. Fast hierarchical Bayesian analysis of population structure. Nucleic Acids Res. 47(11), 5539-5549; 10.1093/nar/gkz361 (2019).

Tsafnat G, Coiera E, Partridge SR, Schaeffer J, Iredell JR. 2009. Context-driven discovery of gene cassettes in mobile integrons using a computational grammar. BMC Bioinformatics 10:281.

Tumbarello M, Spanu T, Di Bidino R, Marchetti M, Ruggeri M, Trecarichi EM, De Pascale G, Proli EM, Cauda R, Cicchetti A, Fadda G. 2010. Costs of bloodstream infections caused by Escherichia coli and influence of extended-spectrum-β-lactamase production and inadequate initial antibiotic therapy. Antimicrobial Agents and Chemotherapy 54:4085-4091.

Vidal-Navarro L, Pfeiffer C, Bouziges N, Sotto A, Lavigne J-P. 2010. Faecal carriage of multidrug-resistant Gram-negative bacilli during a non-outbreak situation in a French university hospital. Journal of Antimicrobial Chemotherapy 65:2455-2458.

von Mentzer A, Connor TR, Wieler LH, Semmler T, Iguchi A, Thomson NR, Rasko DA, Joffre E, Corander J, Pickard D, Wiklund G, Svennerholm A-M, Sjoling A, Dougan G. 2014. Identification of enterotoxigenic Escherichia coli (ETEC) clades with long-term global distribution. Nat Genet 46:1321-1326.

Yu, G., Smith, D.K., Zhu, H., Guan, Y. & Lam, T.T.Y. 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; 10.1111/2041-210X.12628 (2017)

Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18:821-829.

Chapter 4: Complete assembly of *Escherichia coli* ST131 genomes using long DNA reads demonstrates antibiotic resistance gene variation within diverse plasmid and chromosomal contexts

Abstract

The incidence of infections caused by extraintestinal *Escherichia coli* (ExPEC) is rising globally, which is a major public health concern. ExPEC strains that are resistant to antimicrobials have been associated with excess mortality, prolonged hospital stays and higher healthcare costs. *E. coli* ST131 is a major ExPEC clonal group worldwide with variable plasmid composition and has an array of genes enabling antimicrobial resistance (AMR). ST131 isolates frequently encode the AMR genes *bla*$_{\text{CTX-M-14/15/27}}$, which are often rearranged, amplified and translocated by mobile genetic elements (MGEs). Short DNA reads do not fully resolve the architecture of repetitive elements on plasmids to allow MGE structures encoding *bla*$_{\text{CTX-M}}$ genes to be fully determined. Here, we performed long read sequencing to decipher the genome structures of six *E. coli* ST131 isolated from six patients. Most long read assemblies generated entire chromosomes and plasmids as single contigs, contrasting with more fragmented assemblies created with short reads alone. The long-read assemblies highlighted diverse accessory genomes with *bla*$_{\text{CTX-M-15}}$, *bla*$_{\text{CTX-M-14}}$ and *bla*$_{\text{CTX-M-27}}$ genes identified in three, one and one isolates, respectively. One sample had no *bla*$_{\text{CTX-M}}$ gene. Two samples had chromosomal *bla*$_{\text{CTX-M-14}}$ and *bla*$_{\text{CTX-M-15}}$ genes, and the latter was at three distinct locations, likely transposed by the adjacent MGEs: IS*Ecp1*, IS*903B* and Tn*2*. This study showed that AMR genes exist in multiple different chromosomal and plasmid contexts even between closely-related isolates within a clonal group such as *E. coli* ST131.

Publication: *mSphere* 2019 with Ludden C, Feltwell T, Judge K, Parkhill J, Downing T.

4.1 Introduction

Reported cases of bloodstream and urinary tract infections caused by extraintestinal pathogenic *Escherichia coli* (ExPEC) are increasing globally at an alarming rate (Poolman and Wacker 2016). As a key source of ExPEC isolates worldwide, *E. coli* sequence type 131 (ST131) is regarded as a serious threat to public health, given its high level of antimicrobial resistance (AMR), as well as the broad spectrum of infections it causes in community and hospital settings (Pitout et a. 2018; Goswami et al. 2018).

E. coli ST131 is virulent (Ender et al. 2009) and has an expansive range of virulence factors (Van der Bij et al. 2012; Calhau et al. 2013), especially those linked to uropathogenic *E. coli* (UPEC) (Goswami et al. 2018; Totsika et al. 2011; Ben Zakour et al. 2016). AMR and virulence genes allow ST131 to adapt to drug selection pressure and to survive in extraintestinal niches and are often encoded on mobile genetic elements (MGEs) (Forde et al. 2015), which means the exact set of virulence and AMR genes in a single ST131 isolate may vary (Ben Zakour et al. 2016; Johnson et al. 2010). ST131 encodes a range of extended-spectrum β-lactamases (ESBLs) that hydrolyse third-line drugs including cephalosporins, the most common of which encode cefotaximase *bla*$_{\text{CTX-M-15}}$. Within ST131, clade C2 has more AMR genes than other clades and is typically *bla*$_{\text{CTX-M-15}}$-positive, differentiating it from clade C1 that can be *bla*$_{\text{CTX-M-14}}$ or *bla*$_{\text{CTX-M-27}}$-positive (Goswami et al. 2018; Ben Zakour et al. 2016).

Most ST131 AMR genes are reported to be encoded on plasmids: circular self-replicating double-stranded DNA molecules that constitute part of the bacterial accessory genome (Juhas et al. 2009; Frost et al. 2005; Hinnebusch and Tilly 1993). Plasmids can reduce bacterial cell fitness, but a number of post-segregation killing and stable plasmid inheritance mechanisms allow the stable maintenance of IncF plasmids in ST131 (Woodford et al. 2009; Nicolas-Chanoine et al. 2014; Phan et al. 2015). The chromosomal integration of plasmid genes is most commonly facilitated by transposons, which can ensure acquisition and conservation of such elements if there is no subsequent local recombination (Harrison and Brockhurst 2012; MacLean andSan Millan 2015).

Identifying plasmid conjugation, recombination and transposition could have value in tracking AMR genes associated with disease outbreaks and antibiotic treatment failures. Plasmids may be classified using incompatibility (Inc), relaxase (MOB) and mating pair formation system typing (Shintani et al. 2015), but difficulties in plasmid genetic analysis and reconstruction arise with short read data due to rearrangements driven by recombination, dense arrays of repetitive elements including transposable elements (TEs), changes in gene copy numbers, and high sequence variation. Methods using short reads alone may fail to detect genomic segments exchanged between plasmids and the chromosome, limiting evaluation of the core and accessory genomes.

Whole genome sequencing has provided a high resolution of the genomic epidemiology of ST131 and plasmid-mediated AMR outbreaks (McNally et al. 2016). However, short reads alone are insufficient to resolve plasmids that often have numerous small MGEs of ~1 kb or less in size, e.g. TEs and insertion sequences (ISs) (Wick et al. 2017). Complex transposable units (TUs) consisting of multiple TEs or ISs can mobilise AMR genes by transposition, and this can sometimes be followed by recombination within the TU between one of the inverted repeats (IRs) flanking the TE and the IR of another local TE or an adjacent homologous sequence, resulting in different TU structures, locations and copy numbers. At present, the exact resolution of complex structural rearrangements of repetitive TUs containing AMR genes may be impossible with short reads (Arredondo-Alonso et al. 2018). Consequently, plasmid assembly is a challenge requiring accurate long reads and sufficient coverage to distinguish between independent plasmids with regions of sequence identity (Wick et al. 2017; Judge et al. 2016).

Long reads, such as those generated using Oxford Nanopore Technologies (ONT) or Pacific Biosciences platforms can provide a solution to this plasmid assembly problem (Leggett and Clark 2017; Roer et al. 2018; Goldstein et al. 2018). Here, we sequenced six ST131 using the ONT GridION X5 platform. Using the resulting high-coverage sequence data, we reconstructed and annotated the plasmids and chromosomal regions carrying *bla*$_{\text{CTX-M}}$ genes, as well as their genetic context and copy numbers.

4.2 Methods

Author contributions: As indicated by the in the published paper, I was involved in conceptualization, all methods, bioinformatic processing, genomic analysis, interpreting results, drafting the paper, editing the paper and visualization the results. Catherine Ludden helped with sample acquisition in Cambridge. Theresa Feltwell assisted with bacterial culturing and DNA isolation. Kim Judge arranged and managed the Nanopore GridION sequencing. Julian Parkhill contributed to project management and paper writing. Tim Downing helped with project design, bioinformatics and paper writing. I completed the majority of the work in this chapter and was involved in all aspects.

4.2.1 Sample collection

Six ESBL-producing *E. coli* ST131 clinical strains were isolated in June-October 2015 from patients at Addenbrooke's Hospital, Cambridge, as part of a study on antibiotic resistance (Table 4.1). Five samples were from faeces, and one was from blood. These were short-read sequenced in a multiplex run on an Illumina HiSeq 2500 platform and processed as previously outlined (Ludden et al. 2017).

4.2.2 High molecular weight DNA extraction

Frozen stocks of the six isolates were streaked onto LB agar plates and grown overnight at 37^{o}C. Single colonies were subcultured onto LB agar plates and incubated overnight at 37^{o}C. DNA was extracted using a Lucigen Masterpure Complete DNA and RNA Purification kit. For each sample, a swab was used to sweep half a plate of pure colonies and suspended in 1x phosphate buffer solution (PBS). Samples were processed according to the manufacturer's instructions, with elution in 70ul of Nuclease Free water. Pipetting was minimised to reduce shearing of the DNA prior to sequencing.

Strain	Source	Sampling date	Accession numbers		FigShare long read library locations
			Short reads	Long reads	
VRES1160	Faeces	26/08/2015	ERR1878359	ERR3284709	https://ndownloader.figshare.com/files/14039495
VREC0693	Faeces	03/06/2015	ERR2137889	ERR3284704	https://ndownloader.figshare.com/files/14039639
VRES0739	Faeces	05/06/2015	ERR1878196	ERR3284708	https://ndownloader.figshare.com/files/14039354
VREC1013	Faeces	19/08/2015	ERR2138591	ERR3284705	https://ndownloader.figshare.com/files/14039333
VREC1073	Blood	26/08/2015	ERR2138200	ERR3284706	https://ndownloader.figshare.com/files/14039345
VREC1428	Faeces	22/10/2015	ERR2138475	ERR3284707	https://ndownloader.figshare.com/files/14039351

Table 4.1. Sample collection source, sampling date and sequence read accession numbers.

4.2.3 Oxford Nanopore library preparation and sequencing

DNA was quantified using a Quant-iT™ HS (High Sensitivity) kit (Invitrogen). DNA purity was checked using a Nanodrop (ThermoFisher) and fragment size was confirmed by FEMTO Pulse (Nano Life Quest). The sequencing libraries were prepared using 1 µg DNA per sample and ligation sequencing kit 1D SQK-LSK109 with the barcoding extension kit EXP-NPB104 according to ONT protocols. The samples were combined using equimolar pooling and loaded onto a single 9.4.1 MIN-106 flow cell and sequenced on the GridION X5 platform under standard conditions.

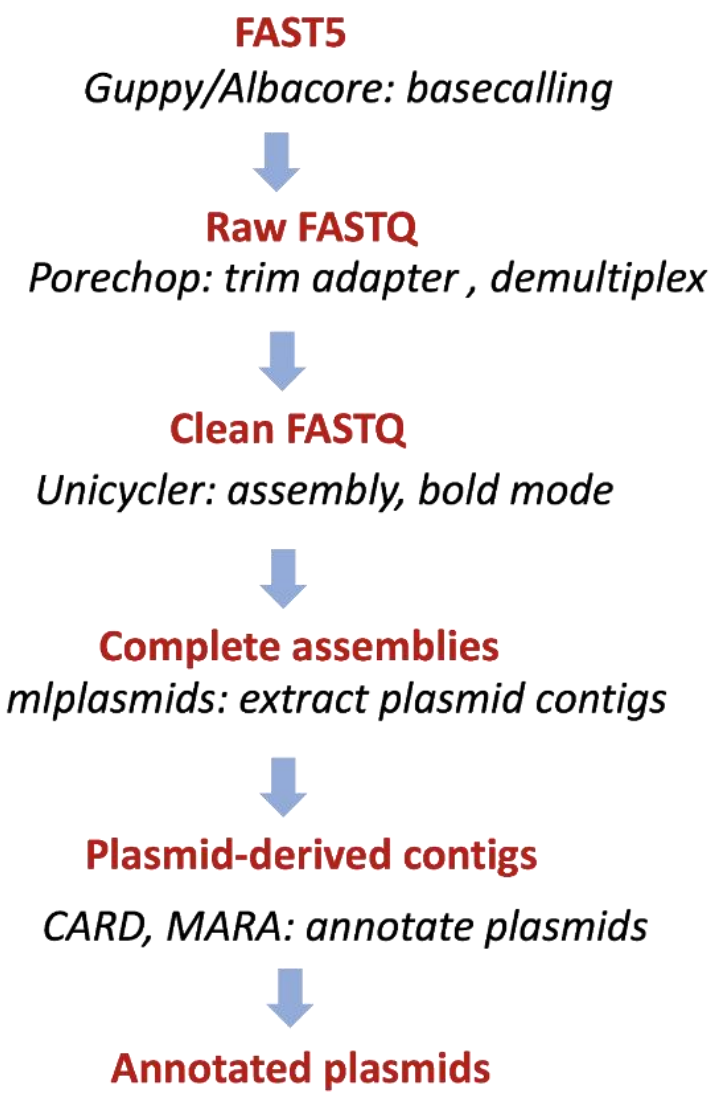
4.2.4 Illumina library preparation and sequencing

The short reads used in this study were created as follows: bacterial genomic DNA was extracted using the QIAxtractor (Qiagen, Valencia, CA, USA) according to the manufacturer's instructions. Library preparation was conducted according to the Illumina protocol and sequenced (96-plex) on an Illumina HiSeq 2500 platform (Illumina, San Diego, CA, USA) using 100 bp paired-end reads.

4.2.5 Oxford Nanopore base-calling and adapter trimming

The resulting fast5 read files (available at www.ncbi.nlm.nih.gov/sra/PRJEB30511, accession numbers ERR3284704-ERR3284709) were transferred to a separate Linux server 4.4.0 (Ubuntu 16.04.4) for analysis. Basecalling was performed during the GridION run using ONT's Guppy v0.5.1 and the resulting fast5 from the initial run was converted to fastq format with Albacore v2.0 (ONT). The statistical data of the sequencing run was processed with MinIONQC v1.3.5 (Lanfear et al. 2018) based on the default Q score cut-off of seven. Adapters and chimeric reads were removed from fastq files using Porechop v0.2.4 (Wick et al. 2017b) with demultiplex settings (Figure 4.1). Standard outputs were saved as log files and were then parsed. The quality of the final fastq files was assessed using FastQC v0.11.8 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and MultiQC v1.4 (Ewels et al. 2016).

Figure 4.1. Overview of genome assembly using Oxford Nanopore reads to recover plasmids with antibiotic resistance genes and mobile genetic elements (MGEs). Oxford Nanopore fast5 sequences were basecalled and converted to fastq format using Albacore v.2.0 and Guppy v.0.5.1. Forward, reverse and middle adapters were removed using Porechop v.0.2.4. The genomes were assembled using Unicycler v.4.6 (optionally including Illumina short reads for comparison). The probability that the resulting contigs were chromosomal or plasmid-associated was measured using mlplasmids. Contigs were annotated using the Comprehensive Antibiotic Resistance Database (CARD) and Multiple antibiotic Resistance Annotator (MARA) to resolve precise plasmid structure, *bla*$_{\text{CTX-M}}$ gene alleles, copy numbers and their adjacent regions.



4.2.6 Genome assembly and improvement

We assembled the genomes using the conservative, normal and bold modes of the long read-only assembly pipeline in Unicycler v4.6. Previous work has suggested that Unicycler outperforms alternatives (Wick et al 2017a) that struggle to resolve plasmids (George et al. 2017). This workflow included the assembly polisher, Racon, which ran iteratively to minimise error rates of called bases (Wick et al. 2017b). For comparison, short read-only and hybrid assemblies were also created using Unicycler v4.6. Briefly, during short read-only assembly, Unicycler v4.6 employed SPAdes v3.12 to assemble short reads then used Pilon v1.22 to polish the assembly. In hybrid assemblies, Unicycler v4.6 used Miniasm to piece long reads together first and applied SPAdes v3.12 to incorporate short reads and bridge gaps. Pilon was run 3-10 times for short read assemblies and 5-10 times for hybrid ones, until no further changes were required to achieve the most contiguous and completed genome assemblies. The average number of changes by Pilon was 74.3, 100.2 and 125.3 for short read assemblies, and 234.5, 257.7 and 377.0 across conservative, normal and bold modes (respectively).

4.2.7 Genome assembly assessment and error rate quantification

The quality of resulting assemblies was assessed using Quast 3.0 (Gurevich et al. 2013) according to the total assembly length, number of contigs, N50, GC content and degree of replicon circularization. Assembly graphs were visualized with Bandage (Wick et al. 2015). The resulting contigs in each assembly were classified as chromosomal or plasmid using machine learning algorithms implemented in mlplasmids (Arredondo-Alonso et al. 2018). Genome completeness was examined using the numbers of single-copy universal orthologous genes using Benchmarking Universal Single-Copy Orthologs (BUSCO) v3 with the gammaproteobacteria_odb9 database (Waterhouse et al. 2017).

4.2.8 Read depth estimation

The read depth of each replicon was estimated by aligning the short Illumina and long Oxford Nanopore reads to the completed genomes using Smalt v0.7.6 and BWA-MEM v0.7.17 (with the flag –x ont2d for ONT reads), respectively. SAMtools v1.7 was used to

process the SAM files to BAM format, remove duplicates, and identify the coverage at each base of each assembly. The median value for each replicon was noted and was normalized using the median chromosomal depth of the same assembly.

4.2.9 Genome annotation

The genomes were annotated using Prokka v1.13.3 (Seemann 2014). *Bla*$_{\text{CTX-M}}$ alleles and their contexts were detected using the Multiple Antibiotic Resistance Annotator (MARA) (Partridge and Tsafnat 2018) and by aligning the assemblies against the Comprehensive Antibiotic Resistance Database (CARD v3.0) to screen for matches with 100% ID only. Information on the detected AMR features and MGEs are retrieved from Galileo AMR (https://galileoamr.arcbio.com/mara/feature/list). Plasmid identification and typing was carried out using PlasmidFinder v2.0 (Carattoli et al. 2014). The plasmid-derived contigs from the assembled genomes were compared using BLAST v2.6.0 using a database of 10,892 complete plasmids (Brooks et al. 2019). Their sequence similarity and annotation were visualised using EasyFig v2.2.2 (Sullivan et al. 2011).

4.2.10 Phylogenetic analysis

To provide a phylogenetic context for these six isolates, the short Illumina reads of 63 from Ben Zakour et al. 2016 and 56 from Matsumura et al. 2017 published ST131 short read libraries were cleaned and trimmed using Fastp v0.12.3 (Chen et al. 2018a), as were the six isolates' short read libraries from this study. These 125 libraries were *de novo* assembled with Unicycler v4.6 using NCTC13441 as a reference and annotated using Prokka. The 126 genomes were processed using Roary v3.11.2 (Page et al. 2015) with a 95% BLAST v2.6.0 identity threshold to create a core genome alignment containing 4,457 SNPs using MAFFT v7.310 (Katoh and Standley 2013) spanning 3,250,343 bases and 3,350 genes of the NCTC13441 chromosome (a length similar to (McNally et al. 2016)). This core genome was used to construct a maximum likelihood phylogeny using RAxML v8.2.11 with the GTR model with gamma rate heterogeneity (Stamatakis 2014). Clade classification of the six isolates was based on published ST131 phylogenetic analysis (Ben Zakour et al. 2015) with associated classification and *bla*$_{\text{CTX-M}}$ allele data from (Ben Zakour et al. 2016) and (Matsumura et al. 2017).

4.3 Results

4.3.1 Oxford Nanopore long read quality control and filtering

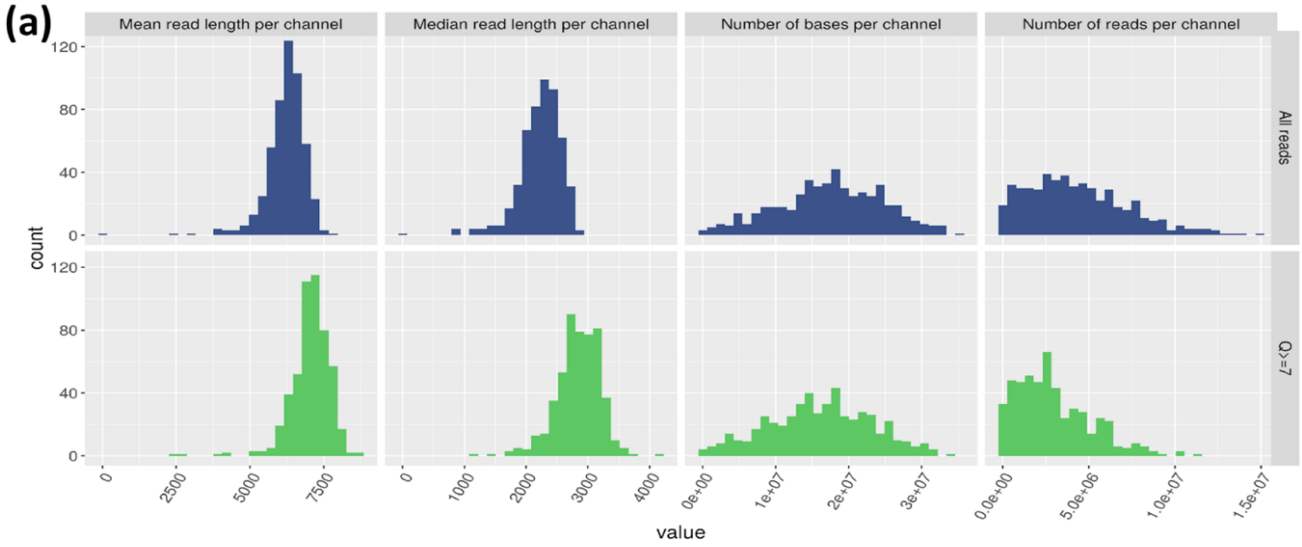
High molecular weight DNA from six *E. coli* ST131 isolates was sequenced using long Oxford Nanopore reads and short Illumina reads to assemble their genomes allowing for plasmid reconstruction and resolution of AMR genes, MGEs and associated rearrangements. The ONT GridION X5 sequencing generated 8.9 Gbases in total across 1,406,087 reads (mean length of 6.3 Kb, Table 4.2).

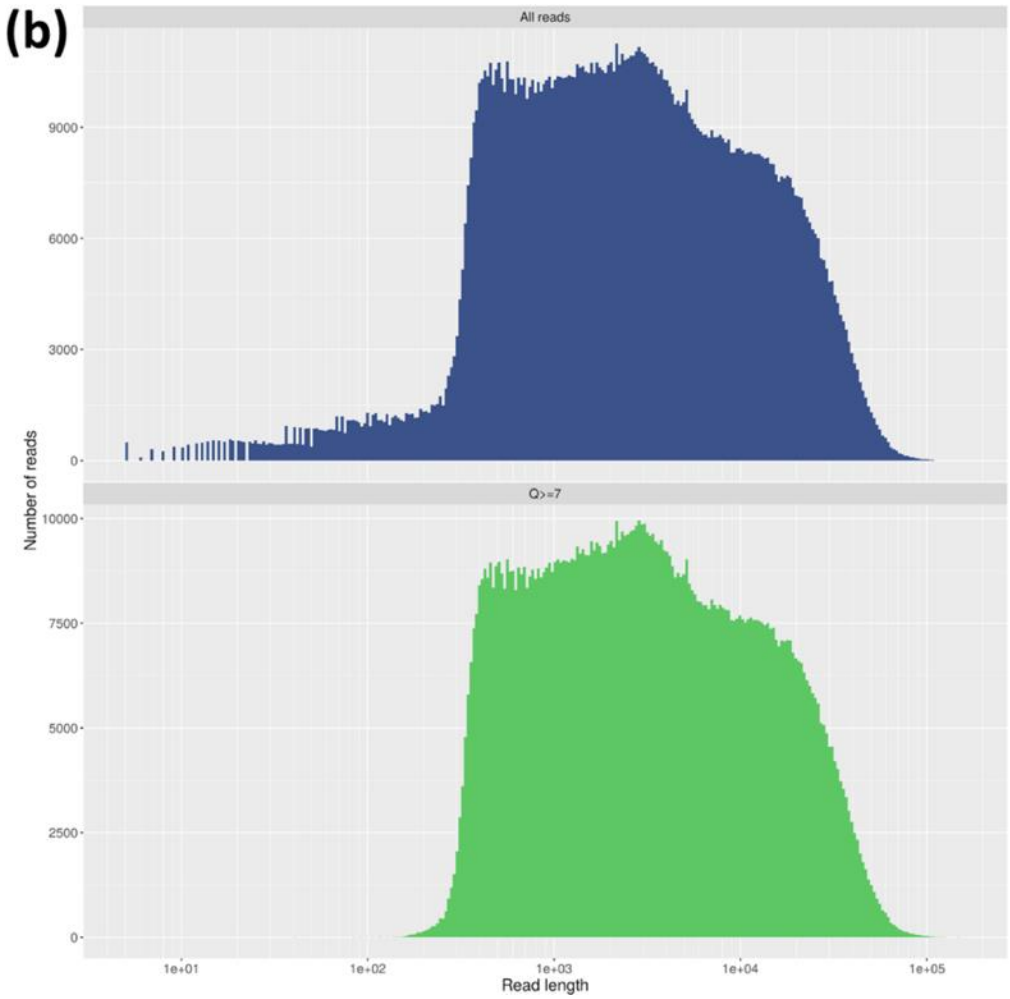
Parameter	All reads	Reads with Q≥7
Total bases	8,908,946	8,193,921
Total reads	1,406,087	1,142,067
Mean length (bp)	6,336	7,175
Median length (bp)	2,273	2,897
Mean Q score	9.1	10.2
Median Q score	10.0	10.5
Reads >100 Kb	85	81

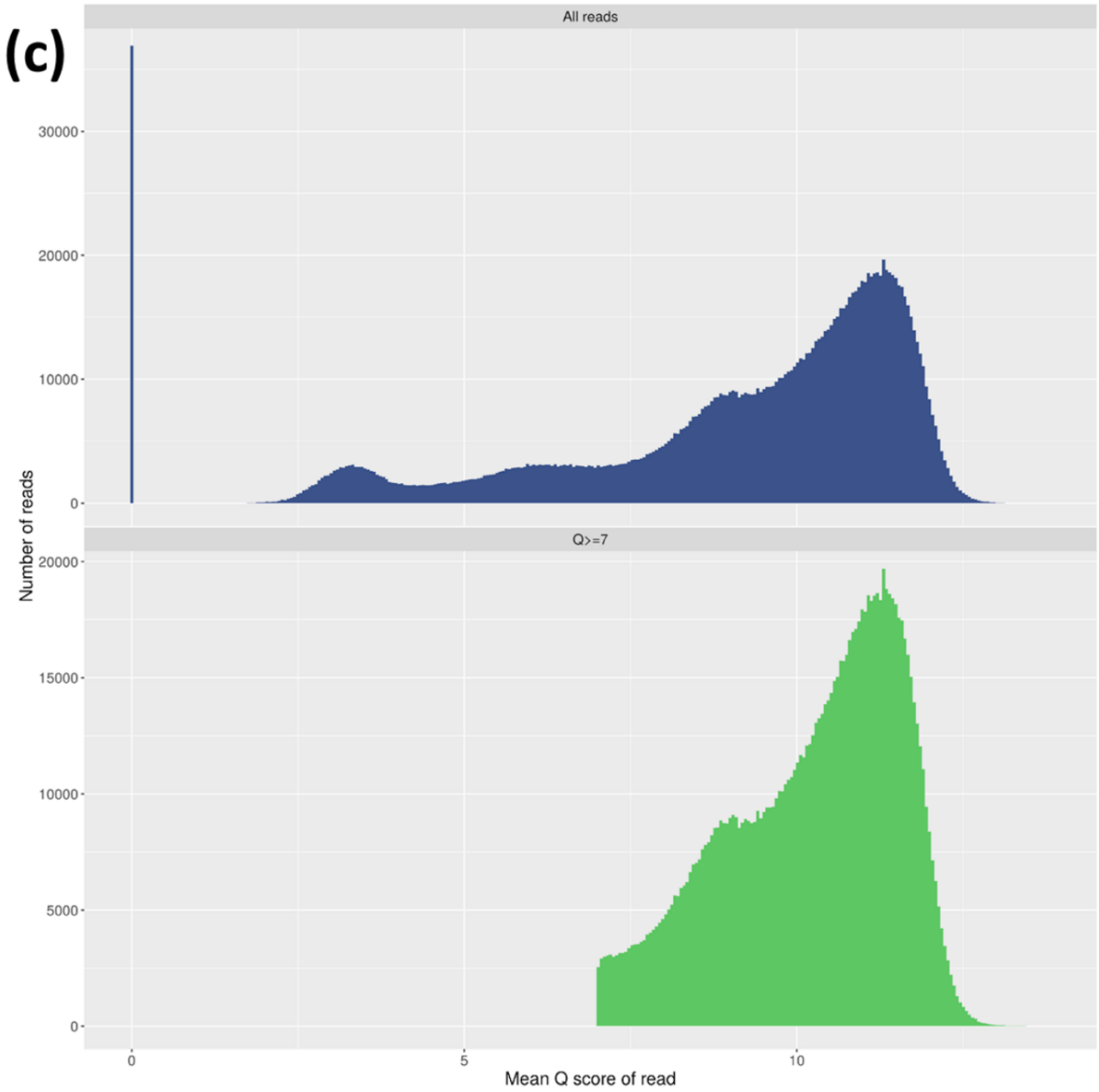
Table 4.2. Quality parameters indicated high-quality read libraries for the six ST131 samples from GridION X5 sequence data. A total of 264,020 of low-quality reads (with Q<7) totalling 715,024,800 bases were excluded.

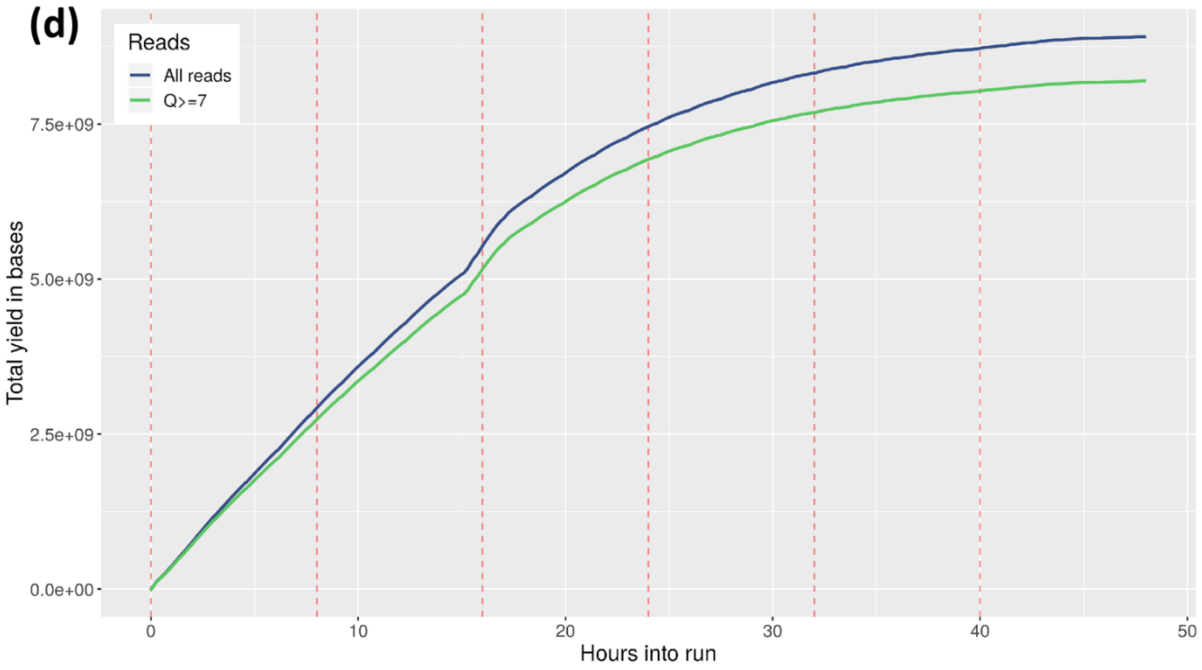
The number of reads generated per hour, total yield of bases over time, read length distribution, and read Q score distribution was examined (Figure 4.2a-h). Half of the reads were produced within 14 hours of sequencing, with the remainder produced over the subsequent 34 hours (Figure 4.2d). A median read length of 5.5 Kb for reads Q (quality) score $\geq$ 7 was achieved within one hour of sequencing (Figure 4.2e), and the median Q score declined slightly as the run proceeded (Supplementary Figure 4.2f). An average of 30-fold theoretical coverage from 954 Mbases with Q $\geq$ 7 was exceeded in this GridION run within three hours.

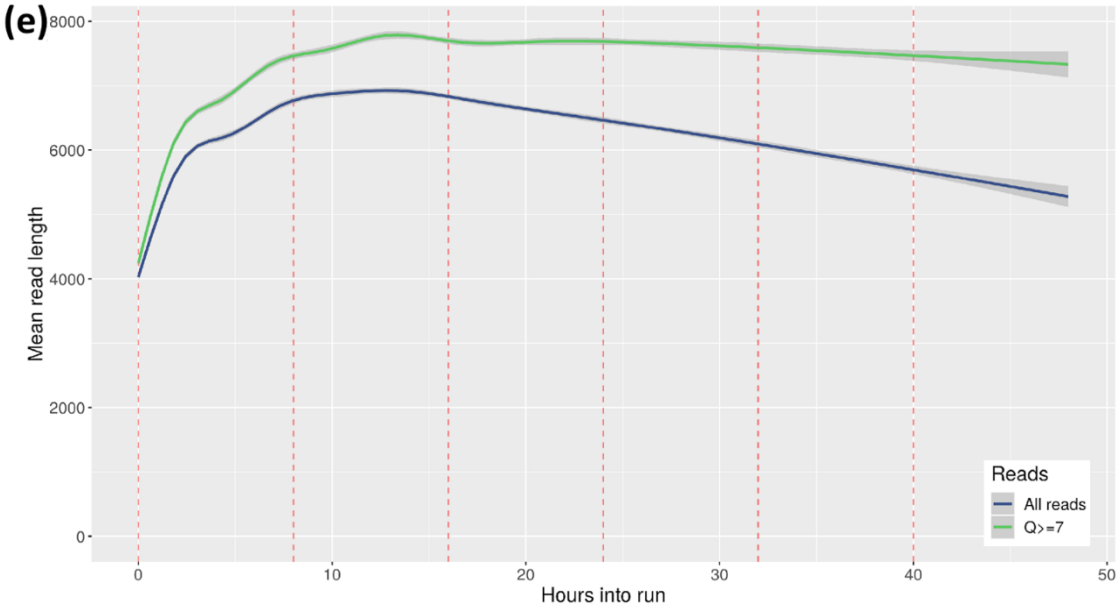
Figure 4.2a-g. Summary plots of the GridION X5 sequencing run for all (blue) and filtered (green) nanopore reads generated using MinIONQC. The graphs in (a) show the read count (y-axis) with the mean and median read length and the number of bases and reads per channel (x-axis), the overall read count (y-axis) vs length (x-axis) in (b) and read count (y-axis) vs the mean Q score (x-axis) in (c). Plots were also drawn to present the total amount of bases called (x-axis; d), the mean read length (x-axis; e) and the mean Q score (x-axis; f) per hour (in their y-axes); the total amount of bases (y-axis) contained in a minimum read length (x-axis) is shown in (g).

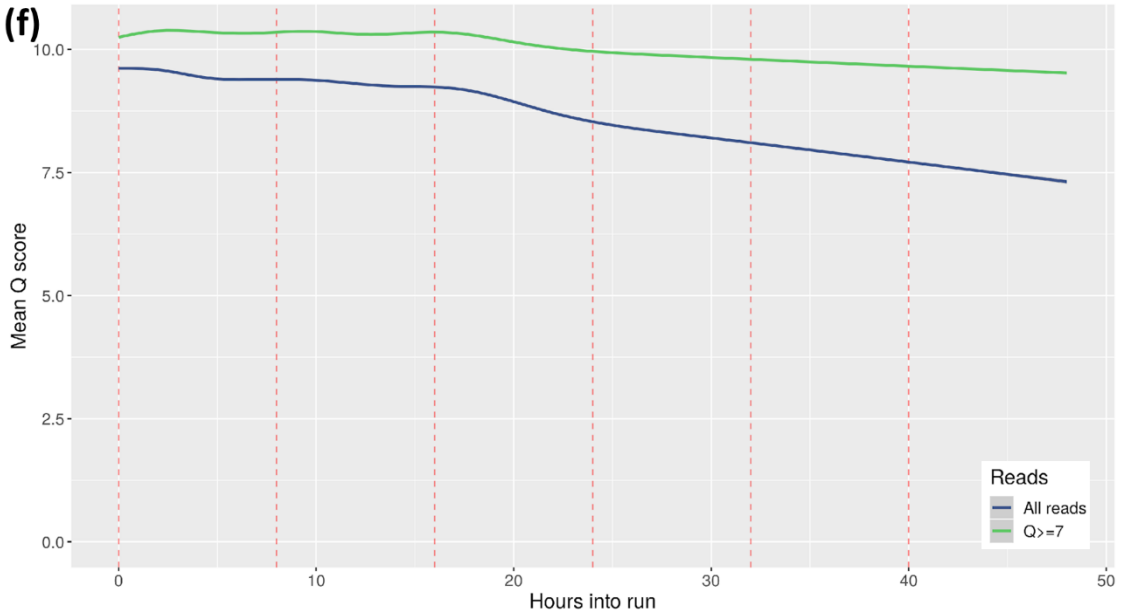


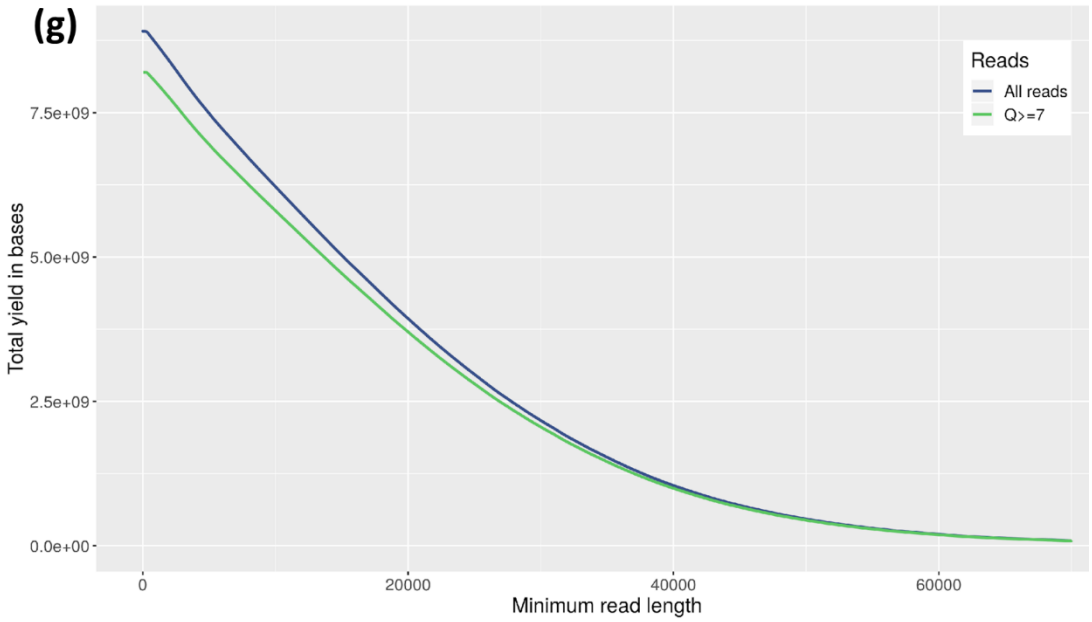

(b)
All reads
9000
6000
3000
0
Q>=7
10000
7500
5000
2500
0
Number of reads
1e+01
1e+02
1e+03
1e+04
1e+05
Read length


(c)
All reads
Q>=7
Number of reads
Mean Q score of read
0
5
10
30000
20000
15000
10000
5000


(d)
Reads
All reads
Q>=7
7.5e+09
5.0e+09
2.5e+09
0.0e+00
Total yield in bases
0
10
20
30
40
50
Hours into run


(e)
8000
6000
4000
2000
0
Mean read length
0
10
20
30
40
50
Hours into run
Reads
All reads
Q>=7


(f)
Mean Q score
10.0
7.5
5.0
2.5
0.0
0
10
20
30
40
50
Hours into run
Reads
All reads
Q>=7


(g)
Total yield in bases
7.5e+09
5.0e+09
2.5e+09
0.0e+00
0
20000
40000
60000
Minimum read length
Reads
All reads
Q>=7

Half of the bases with Q ≥ 7 were on reads of 18 Kb or longer (Figure 4.3). These metrics indicated sufficient GridION data in terms of quantity and quality. Initial screening removed reads with Q < 7, leaving 1,142,067 reads with 8.2 Gbp with a mean Q score of 10.2 and a mean length of 7.2 Kb for analysis. This included 81 reads longer than 100 Kb, including one of 155,312 bases. This corresponded to 257-fold theoretical coverage for six 5.3 Mb genomes.

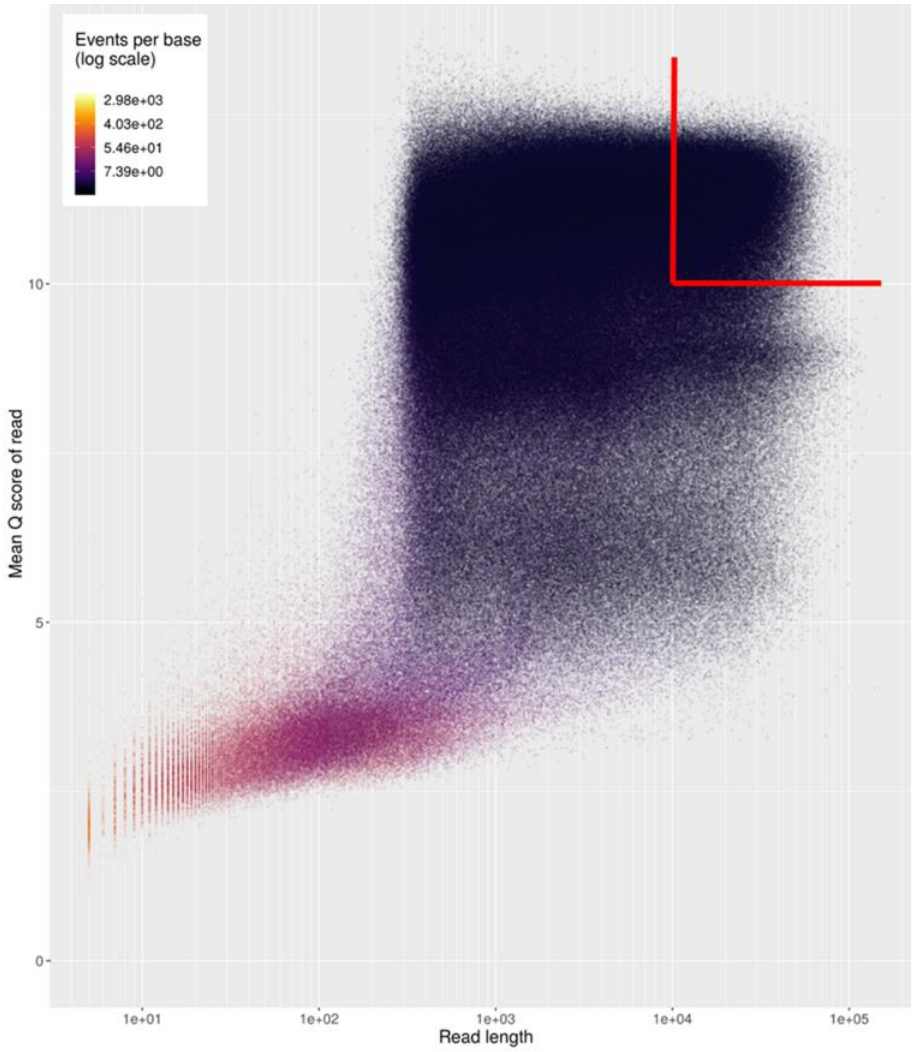


Figure 4.3. Summary of the GridION X5 sequencing run output showing the read length on a log10 scale (x-axis) versus the mean Q score of each read (y-axis) where points are coloured by events per base. The horizontal red line shows reads with lengths > 10 Kb and the vertical red line read with Q scores > 10. Together, this area showed the large number of long high-quality reads generated in this study. This plot emphasises that a high proportion of the bases were accurately called: these were subsequently used for downstream analysis.

The initial number of reads per library ranged from 127,118 to 510,253 and these were filtered using a series of steps to ensure that the reads used for each of the six assemblies had high quality. Bases were successfully called at an average of 97.9% of reads (Table 4.3). Identifying the consensus demultiplexed, duplicate-free and adapter-free reads from Porechop v0.2.4 eliminated a further 2.9% of the basecalled reads, yielding 120,123 to 487,482 reads per library (Table 4.3).

Strain	Initial reads (fast5)	Basecalled (fastq)	Adapter-free (fastq)	Average length (bp)
VRES1160	358,829	351,636	345,033	7,037
VREC0693	208,478	204,904	194,413	8,982
VRES0739	163,349	160,693	155,900	9,171
VREC1013	510,253	497,646	487,482	6,657
VREC1073	313,627	304,218	298,658	7,256
VREC1428	127,118	124,539	120,123	9,301

Table 4.3. Number of reads generated from GridION X5 sequencing data per library that passed filtering during basecalling with Albacore v2.0 and those that were adapter-free (using Porechop v0.2.4). The latter totalling 1,601,609 reads were used for downstream analyses. 80,045 reads were excluded during basecalling or adapter-trimming.

4.3.2 Long read genome assembly illuminates highly diverse accessory genomes

We compared short read-only, long read-only and hybrid assembly outputs from Unicycler v.4.6 using the long Oxford Nanopore reads and short Illumina reads to identify the most contiguous assemblies per sample across all three Unicycler modes (conservative, normal and bold). All six genome assemblies produced chromosomes of 4.81-5.38 Mb with differing numbers of plasmids with lengths spanning 4-156 Kb (Figure 4.4; Table 4.4). The numbers of contigs produced by long read assemblies of two samples (VREC0693, VRES0739) corresponded exactly to the chromosome and plasmids. The others had either one (VREC1073, VRES1160, VREC1013) or two (VREC1428) additional chromosomal contigs (Table 4.2).

Strain	Genome length (bp)	Number of contigs		N50	Chromosome size (Mb)	Number of plasmids	Plasmid sizes (Kb)
		Assembled	Minimum possible				
VRES1160	5,326,801	6	5	5,126,679	5.23	4	62, 16, 5, 4
VREC0693	5,260,741	3	3	5,039,909	5.04	2	132, 89
VRES0739	4,806,912	3	3	4,797,749	4.81	2	5, 4
VREC1013	5,223,433	3	2	3,699,451	5.14	1	90
VREC1073	5,539,158	3	2	5,286,804	5.38	1	156
VREC1428	5,236,419	7	5	4,924,536	5.13	4	92, 5, 5, 4

Table 4.4. Total size of assemblies, chromosomes and plasmids found in each strain based on their optimal whole genome assemblies using the GridION X5 long reads. Each assembly had seven or less contigs, and in three cases no fewer contigs were possible, consistent with full genome assembly (for VREC0693, VRES0739 and VREC1073). The optimal assembly with Unicycler used long reads alone (in bold mode), with exception of VREC1013, where a hybrid combining short Illumina reads with long Oxford Nanopore reads was best, with minor manual screening.

For five samples, the long read assemblies produced 2-7 contigs (with a median of three) with nearly identical results across modes, whereas the short read assemblies resulted in 76-230 contigs (a median of 124), and the hybrid assemblies also had more contigs (6-191 with a median of 44). For VREC0739 and VREC1428, the short read libraries resulted in over-bridging of contigs making it harder to classify contigs as chromosomal or plasmid-associated, perhaps because long reads already provided sufficient genome coverage and the assembler inserted the contigs produced by short reads at short homologous repetitive regions.

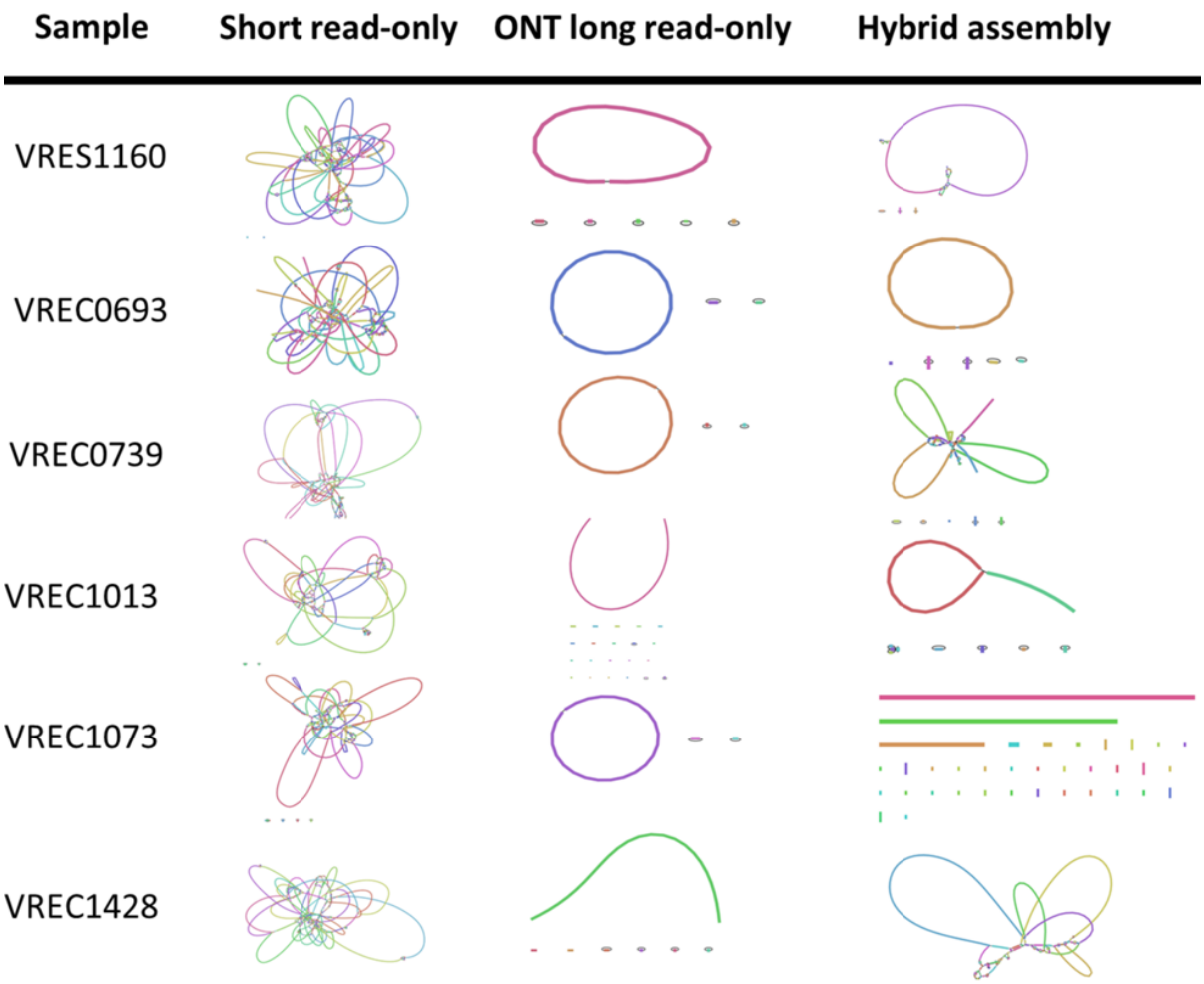


Figure 4.4. The assembly graphs of six *E. coli* ST131 genomes showed many connected edges for those created from short Illumina HiSeq reads only (left) but near-complete assemblies for those made with long Oxford Nanopore read-only (centre) and the hybrid assemblies of most of the strains (right). The assemblies were generated with Unicycler v.4.6 and were visualised using Bandage. Circularized contigs indicated complete assemblies.

For VREC1013, the hybrid assembly improved the long read assembly such that the final optimised version had three rather than 22 contigs and a smaller length (5.36 Mb, Table 4.5), after manual sequence alignment eliminated seven false-positive short contigs. Five contigs had depths of coverage <8% of the chromosomal median and may were the result of contig overbridging during assembly. Pairwise alignment of these five contigs with BLAST against the assembly showed that they had near-perfect matches with other contigs, showing that they were effectively duplicate contigs, and thus few reads mapped

to them. In contrast, the other four valid contigs acted positive controls and showed high homology to their own contigs only. As a result, duplicate contigs were removed from the VREC1013 hybrid assembly used for subsequent analyses.

Contigs were classified as chromosomal or plasmid-derived using mlplasmids given a probability threshold of 60% (Arredondo-Alonso et al. 2018), with further screening for plasmid-related gene content using MARA, CARD and PlasmidFinder (Table 4.5). The largest plasmid was a 156.3 Kb IncFIA one in VREC1073, its sole plasmid. VREC1428 and VRES1160 had 92.8 and 61.9 Kb IncFIA plasmids, respectively, along with three small Col plasmids each (Table 4.4). VREC0693 had a 132.0 Kb IncFIB plasmid and an 88.8 Kb IncB plasmid - IncB plasmids have the same Rep domains as IncFII plasmids (Partridge et al. 2018). VREC3013 had one 89.9 Kb IncFII plasmid. VRES0739 alone had no large plasmid, which was verified with the short read data.

Strain	Prediction	Prediction value (%)	Contig ID	Length (bp)	*bla*$_{CTX-M}$ allele	*bla*$_{CTX-M}$ count	Plasmid type	Median Depth	Normalized Depth
VRES1160	Chromosome	98	1	5,126,679			-	258	1.00
	Chromosome	70	2	113,086			-	213	1.00
	Plasmid	70	3	61,934	15	1	IncFIA	282	1.10
	Plasmid	85	4	15,803			ColRNAI	420	1.64
	Plasmid	81	5	5,203			ColRNAI	11	0.04
	Plasmid	83	6	4,096			Col8282	473	1.85
VREC0693	Chromosome	98	1	5,039,909	15	3	-	258	1.00
	Plasmid	61	2	132,042			IncFIB	213	0.83
	Plasmid	60	3	88,790			IncB	282	1.09
VRES0739	Chromosome	98	1	4,797,749			-	171	1.00
	Plasmid	96	2	5,162			Col156	436	2.55
	Plasmid	74	3	4,001			-	303	1.77
VREC1013	Chromosome	97	1	3,699,451			-	300	1.00
	Chromosome	97	2	1,434,037			-	335	1.00
	Plasmid	84	4	89,945	15	1	IncFII	1015	3.27
VREC1073	Chromosome	98	1	5,286,804			-	214	1.00
	Plasmid	68	2	156,298			IncFIA	172	0.80
	Chromosome	60	3	96,056	14	1	-	213	1
VREC1428	Chromosome	98	1	4,924,536			-	126	1.00
	Chromosome	97	2	103,034			-	57	1.00
	Chromosome	96	3	101,160			-	41	1.00
	Plasmid	64	4	92,750	27	1	IncFIA	85	0.67
	Plasmid	92	5	5,147			ColRNAI	168	1.33
	Plasmid	99	6	5,143			Col156	207	1.64
	Plasmid	73	7	4,649			ColRNAI	239	1.90

Table 4.5. Contigs were classified as chromosomal or plasmid-derived using the mlplasmids prediction value. Each contig were aligned against CARD to identify the presence/absence of *bla*$_{CTX-M}$ alleles and their copy numbers. Plasmid types were identified using PlasmidFinder.

Assembly	Mode	Metric	VRES1160	VREC0693	VRES0739	VREC1013	VREC1073	VREC1428
Short read-only	Conservative	Total length (bp)	5,142,342	5,146,205	5,181,497	5,208,807	4,967,093	5,375,468
		Number of contigs	168	159	200	148	117	230
		N50 (bp)	124,175	132,865	138,725	134,439	157,528	135,303
		#mismatches /	1.32	1.32	65.4	1.5	285.81	0.69
		#indels / 100 Kb	0.06	0.02	1.84	0.08	261.91	0.04
	Normal	Total length (bp)	5,158,728	5,171,710	5,227,751	5,240,888	4,989,316	5,416,180
		Number of contigs	110	106	123	94	76	148
		N50 (bp)	206,138	190,908	213,071	189,184	222,158	170,443
		#mismatches /	4.64	0.93	69.86	4.25	284.14	2.96
		#indels / 100 Kb	0.21	0.14	2.33	0.36	262.1	0.04
	Bold	Total length (bp)	5,159,662	5,163,846	5,207,686	5,226,735	4,977,746	5,411,973
		Number of contigs	124	120	146	108	86	140
		N50 (bp)	206,044	190,808	212,979	190,412	222,051	184,466
		#mismatches /	3.07	1.78	67.11	1.96	287.37	2.03
		#indels / 100 Kb	0.16	0.06	2.03	0.13	262.26	0.11
Long read-only	Conservative	Total length (bp)	5,326,801	5,260,741	4,806,912	6,307,464	5,539,158	5,236,419
		Number of contigs	6	3	3	22	3	7
		N50 (bp)	5,126,679	5,039,909	4,797,749	5,073,008	5,286,804	4,924,536
		#mismatches /	276.23	241.39	2,772.51	344.5	0	332.79
		#indels / 100 Kb	252.29	264.7	265	306.03	0	289.71
	Normal	Total length (bp)	5,326,801	5,260,741	4,806,912	6,307,464	5,539,158	5,236,419
		Number of contigs	6	3	3	22	3	7
		N50 (bp)	5,126,679	5,039,909	4,797,749	5,073,008	5,286,804	4,924,536
		#mismatches /	276.23	241.39	2772.51	344.5	0	332.79
		#indels / 100 Kb	252.29	264.7	265	306.03	0	289.71
	Bold	Total length (bp)	5,326,801	5,260,741	4,806,912	6,307,464	5,539,158	5,236,419
		Number of contigs	6	3	3	22	2	7
		N50 (bp)	5,126,679	5,039,909	4,797,749	5,073,008	5,286,804	4,924,536
		#mismatches /	276.23	241.39	2772.51	344.5	0	332.79
		#indels / 100 Kb	252.29	264.7	265	306.03	0	289.71
Hybrid	Conservative	Total length (bp)	5,272,824	5,275,251	5,215,332	5,323,049	5,055,625	5,492,517
		Number of contigs	52	6	191	34	51	107
		N50 (bp)	1,444,640	5,048,264	426,378	2,673,977	1,423,856	749,550
		#mismatches /	1.63	242.24	2,764.2	2.04	285.57	3.7
		#indels / 100 Kb	0.32	265.38	263.44	0.09	263.18	0.02
	Normal	Total length (bp)	5,276,305	5,275,251	5,291,108	5,327,833	5,098,966	5,516,886
		Number of contigs	42	6	110	33	44	74
		N50 (bp)	1,746,191	5,048,264	72,0730	2,675,388	1,762,353	1,243,293
		#mismatches /	1.56	242.24	44.59	2.28	284.11	1.65
		#indels / 100 Kb	0.28	265.38	4.07	0.13	266.82	0.02
	Bold	Total length (bp)	5,293,427	5,275,251	5,267,003	5,223,433	5,115,410	5,550,270
		Number of contigs	23	6	32	3	22	47
		N50 (bp)	3,801,465	5,048,264	1,222,073	3,699,451	4,958,323	1,266,683
		#mismatches /	271.47	242.24	2,770.38	321.64	283.97	296.99
		#indels / 100 Kb	252.55	265.38	264.11	268.47	268.27	268.29

Table 4.6. Comparison of short read-only, long read-only and hybrid genome assemblies generated using the conservative, normal and bold modes of Unicycler v.04.6. Assemblies

were assessed according to their total length, number of contigs produced, N50 (bp), numbers of mismatches per 100 Kb and numbers of indels per 100 Kb.

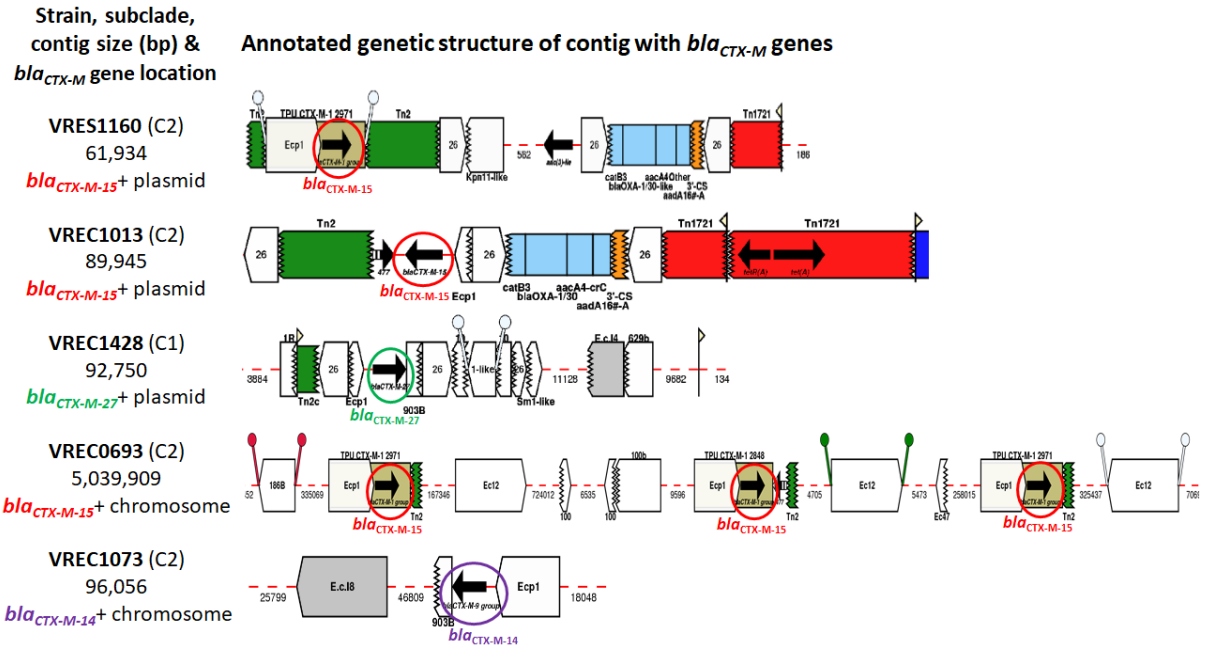


Figure 4.5. Two of the ST131's *bla*CTX-M genes were on chromosomal contigs (VREC0693 and VREC1073). VRES1160 and VREC1013 had IncFIA and IncFII plasmids, respectively, both of which had *bla*CTX-M-15 genes. VREC1428 had an IncFIA plasmid with *bla*CTX-M-27 gene. VRES0739 is not shown because it was *bla*CTX-M-negative and had no large plasmid. The contigs were classified as chromosomal or plasmid-derived by mlplasmids so that the *bla*CTX-M genes and their genetic flanking context could be examined. Annotation was derived from Galileo™ AMR based on the Multiple Antibiotic Resistance Annotator (MARA) and database. The *bla*CTX-M variants are labelled and encircled in red (*bla*CTX-M-15), purple (*bla*CTX-M-14) or green (*bla*CTX-M-27).

By mapping the long reads to the optimal assemblies, the read coverage of each chromosome and plasmid was estimated (Table 4.5). Each chromosome had between 126- and 310-fold median coverage, and the median coverage levels of large plasmids ranged from 85- to 282-fold, except for VREC1013's IncFII plasmid that had 1,015-fold coverage and a normalized depth of 3.3-fold. The normalised depth of plasmids compared to chromosomes suggested some cells in VREC1428 and VREC1073 may have lost their IncFIA plasmid, and the same for VREC0693 and its IncFIB plasmid. However, the IncFIA

plasmid in VRES1160 and the IncB plasmid in VREC0693 had higher than expected copy numbers (by 9% after normalisation), potentially indicating stable plasmid retention.

Across five assemblies in the Unicycler normal mode, the median indel error rates for short reads and hybrid assemblies were similar (0.21 and 0.28 per 100 Kb, respectively), but was much higher for long read assemblies (265.0 per 100 Kb, Table 4.6). Likewise, the median mismatch error rates for short reads and hybrid assemblies were comparable (4.25 and 2.28 per 100 Kb, respectively), but was much higher for long read assemblies (332.8 per 100 Kb, Table 4.6). These rates excluded VREC1073, for which some Quast metrics were zero values. Similarly, the recovery of conserved BUSCO genes was far higher for hybrid assemblies (>99.5%) than for long read ones (>82.3%).

4.3.3 The dynamic locations and genomic contexts of *bla*$_{\text{CTX-M}}$ genes in long read assemblies

The optimised assemblies provided an improved view of the genomic context of each *bla*$_{\text{CTX-M}}$ allele, whose effectiveness as a marker for ST131 clade classification and origin (Ben Zakour et al. 216) we explored here. The deeper resolution of genome architecture revealed surprising differences in *bla*$_{\text{CTX-M}}$ gene context (Figure 4.5; Table 4.5), including the discovery of chromosomal *bla*$_{\text{CTX-M}}$ genes in VREC0693 (three copies of *bla*$_{\text{CTX-M-15}}$) and VREC1073 (one copy of *bla*$_{\text{CTX-M-14}}$). All *bla*$_{\text{CTX-M}}$ genes were complete (876 bp) with adjacent IS*Ecp1* (1,658 bp with flanking IRs of 14-16 bp) and Tn*2* (5.8 Kb) elements: IS*Ecp1* and Tn*2* can transpose *bla*$_{\text{CTX-M}}$ and other ESBL genes (Lartigue et al. 2006; Barlow et al. 2008). The VRES0739 genome did not contain any region homologous to *bla*$_{\text{CTX-M}}$, most likely because it had lost an IncF plasmid, unlike the other isolates.

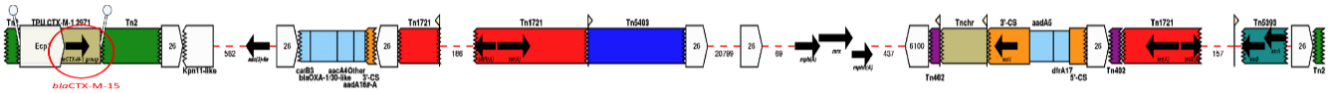
VRES1160, VREC0693 and VREC1013 all had *bla*$_{\text{CTX-M-15}}$ genes linked to isoforms of IS*Ecp1*, IS*26* and Tn*2*, implicating them in driving transposition of the TU (Figure 4.6). Each was similar to the ST131 clade C2 *ISEcp1-bla*$_{\text{CTX-M-15}}$-orf477**Δ** TU (Ben Zakour et al. 2016; Petty et al. 2014) but with distinct structural differences. VRES1160's single *bla*$_{\text{CTX-M-15}}$ gene was at 2,296 bp on its IncFIA plasmid and was flanked by IS*Ecp1* to its 5' and Tn*2* followed by IS*26* at its 3' end, with another Tn*2* 5' of IS*Ecp1*. VREC0693's three chromosomal *bla*$_{\text{CTX-M-15}}$ genes were not tandem repeats (chromosomal locations

2,781,074, 3,696,068 and 3,970,927), but each of these TUs were identical: all had IS*Ecp1* at the 5' ends and truncated Tn*2*s at the 3' ends. VREC1013's sole *bla*$_{\text{CTX-M-15}}$ gene was located at 13,226 bp on its IncFII plasmid and was flanked by a truncated IS*Ecp1* at its 5' end and Tn*2* at its 3' end, with IS*26* copies 5' and 3' of these segments.

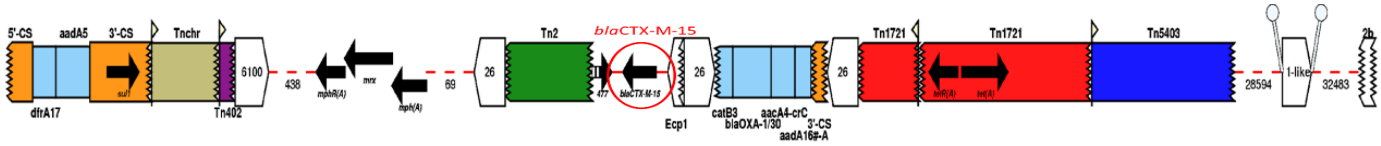
VREC1428's single *bla*$_{\text{CTX-M-27}}$ gene was on its IncFIA plasmid at position 6,018, and VREC1073's single chromosomal *bla*$_{\text{CTX-M-14}}$ gene started at contig position 19,746 (Figure 4.6). Both their *bla*$_{\text{CTX-M}}$ genes were flanked by a truncated IS*Ecp1* at the 5' ends and a shortened IS*903B* at the 3' ends suggesting that IS*Ecp1* and IS*903B* may have facilitated the transposition of the TU from the plasmid. Similar *bla*$_{\text{CTX-M}}$ gene transposition events have been observed in ST131 clade C1 (Ben Zakour et al. 2016).

Alignment of the plasmid-derived contigs of VRES1160 (IncFIA) to VREC1013 (IncFIB) showed that the *bla*$_{\text{CTX-M-15}}$-positive plasmids were much more similar (>83% identity) relative to VREC1428's *bla*$_{\text{CTX-M-27}}$-positive IncFIA plasmid, which was more distinct (Figure 4.7). In addition, VREC1428's plasmid had *traI* and *traD* genes indicating conjugation machinery (Table 4.7) as well as high homology to at least one published plasmid, unlike VRES1160's and VREC1013's plasmids. This suggested that the VRES1160 and VREC1013 plasmids had homology corresponding well with *bla*$_{\text{CTX-M}}$ gene and subclade classification, and that they were structurally different to published plasmids due to recombination.

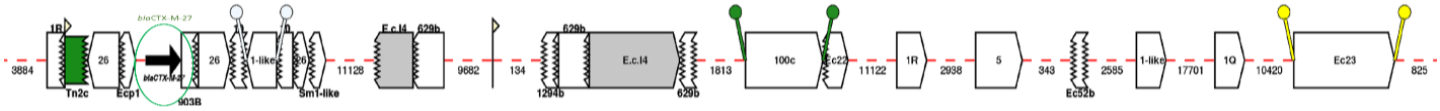
VRES1160 (subclade C2, 61,934 bp *bla*$_{CTX\text{-}M\text{-}15}$+ plasmid)



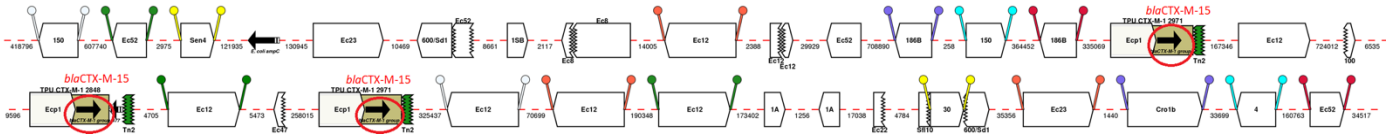
VREC1013 (subclade C2, 89,945 bp, *bla*$_{CTX\text{-}M\text{-}15}$+ plasmid).



VREC1428 (subclade C1, 92,750 bp *bla*$_{CTX\text{-}M\text{-}27}$+ plasmid)



VREC0693 (subclade C2, 5,039,909 bp *bla*$_{CTX\text{-}M\text{-}15}$+ chromosome with 3 distinct *bla*$_{CTX\text{-}M\text{-}15}$ genes in red - single chromosome is split below for visualisation)



VREC1073 (subclade C2, 96,056 bp *bla*$_{CTX\text{-}M\text{-}14}$+)

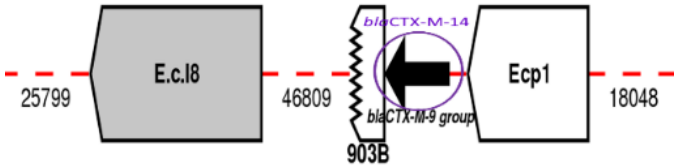
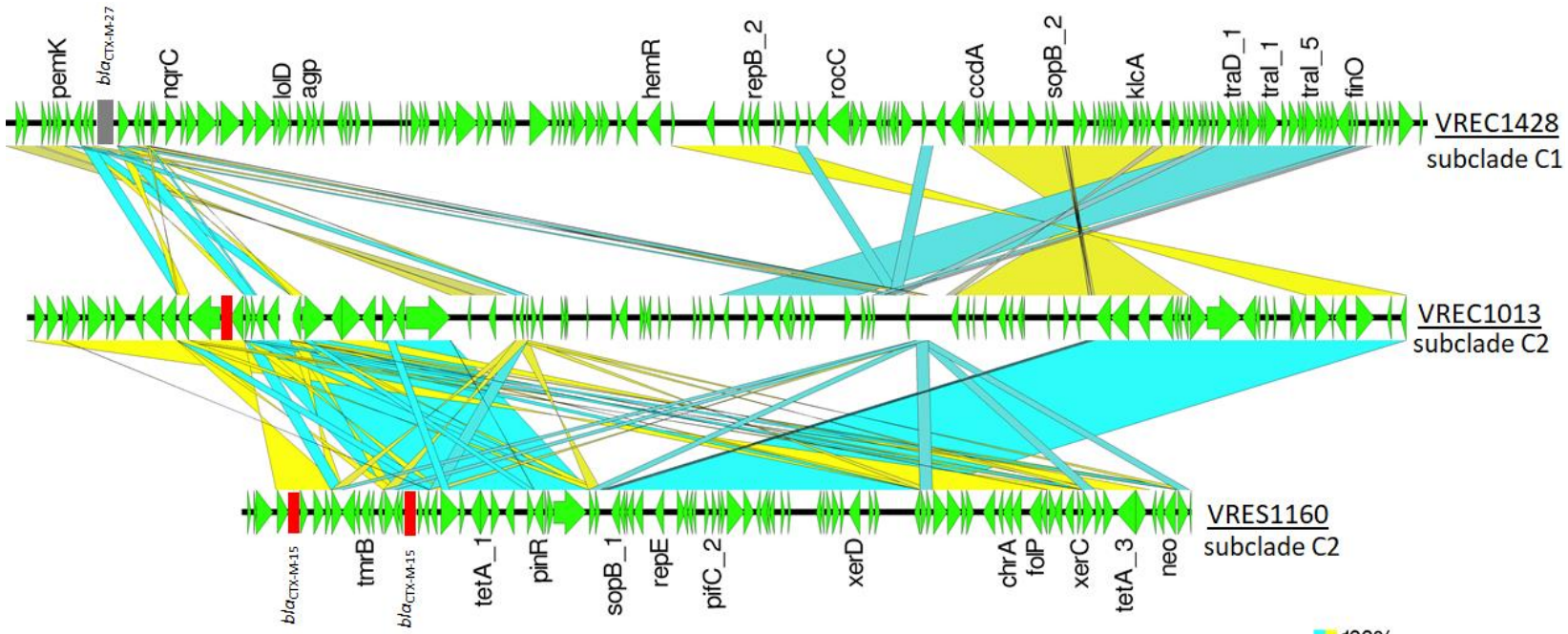


Figure 4.6. The contigs from the most optimal assembly mode of Unicycler v.4.6 of five out of six *E. coli* ST131 samples were identified as chromosomal or plasmid-derived using mlplasmids. These were annotated with *bla*$_{CTX\text{-}M}$ genes and their genetic flanking context using Galileo™ AMR based on the Multiple Antibiotic Resistance Annotator (MARA) and database (Seemann 2014); all *bla*$_{CTX\text{-}M}$ variants are labelled accordingly and encircled in red (*bla*$_{CTX\text{-}M\text{-}15}$), purple (*bla*$_{CTX\text{-}M\text{-}14}$) or green (*bla*$_{CTX\text{-}M\text{-}27}$). The definition of the other elements are listed at https://galileoamr.arcbio.com/mara/feature/list. The long VREC0693 chromosome is split into two parts so that the gene annotation is visible.



Strain	Size (bp)	Plasmid type	Gene count
VRES1160	61934	IncFIA	91
VREC1013	89945	IncFII	220
VREC1428	92750	IncFIA	143

Figure 4.7. Pairwise comparisons of the three *bla*$_{\text{CTX-M}}$-positive plasmid-associated contigs showed high sequence identity for the two from subclade C2 (VREC1013 and VRES1160) relative to one from C1 (VREC1428, top). The BLAST result was visualised with EasyFig v2.2.2 such that the middle blocks connecting regions of the contigs represent nucleotide homology: blue for homologous regions in the same direction, and yellow for inversions. Gaps or white spaces denote unique loci or regions present in a contig but not in the other. Gene models are in green with the direction of transcription shown by arrows. Genes of interest are labelled above each arrow. The *bla*$_{\text{CTX-M-27}}$ grey (top) is in mauve and the two *bla*$_{\text{CTX-M-15}}$ genes (middle, bottom) are in red. The table below shows the contig size, plasmid type and the number of genes per strain. The list and products of the annotated genes are in Table 4.8.

Table 4.7. List of genes (with count # indicated by "_") in the plasmid contigs of VREC1013, VRES1160 and VREC1428. The *bla*$_{CTX-M}$ (*bla*), *traI* and *traD* genes are in bold. Only isolate VREC1428 had *traI* and *traD* genes indicating conjugative capacity. VREC0693, VRES0739 and VREC1073 contigs did not have *tra* genes.

VREC1428	VREC1013	VRES1160
pemI	*xerD_1*	***bla_1***
pemK	*ccdA*	*tmrB*
bla	*vapC_1*	*cat*
nqrC	*vapC_2*	***bla_2***
lolD	*kdgT_1*	*aacA4*
agp	*kdgT_2*	*tetA_1*
hemR	*ridA*	*tetA_2*
repB_1	*yagE*	*pinR*
repB_2	*ugpA*	*sopB_1*
mmuM	*cpdA*	*sopB_2*
rocC	*tnpA*	*repE*
ccdA	*yknY*	*ccdB*
ccdB	*tpd*	*ccdA*
sopB_1	*xerD_2*	*pifC_1*
sopB_2	*dhfrI_1*	*pifC_2*
klcA	*ant1_1*	*pifC_3*
traD_1	*folP*	*repB_1*
traD_2	*srpC*	*repB_2*
traD_3	***bla***	*xerD*
traD_4	*xerD_3*	*chrA*
traD_5	*xerC*	*folP*
traD_6	*dhfrI_2*	*mdtJ*
traI_1	*ant1_2*	*xerC*
traI_2	*umuC*	*tetA_3*
traI_3	*lexA*	*tetR*
traI_4	*klcA*	*neo*
traI_5		*tnpR*
traI_6		
traI_7		
finO		

Table 4.8. Protein products encoded by the genes found in plasmids of VREC1013, VRES1160 and VREC1428 (Figure 4.7).

Gene	Protein product
agp	Glucose-1-phosphatase
ccDA	Antitoxin (Plasmid maintenance)
chrA	Response regulator
finO	Fertility inhibition protein
folP	Dihydropteroate synthase
hemR	Hemin TonB-dependent receptor
klcA	Antirestriction protein
lolD	Lipoprotein-releasing system ATP-binding
neo	Aminoglycoside 3'-phosphotransferase
nqrC	Na(+)-translocating NADH-quinone reductase
pemK	mRNA interferase
pifC	Transcriptional repressor protein
pinR	Serine recombinase protein
repB	Replication protein
repE	Replication initiation protein
rocC	Amino-acid permease
sopB	Inositol phosphate phosphatase
tetA	Tetracycline resistance protein
tmrB	Tunicamycin resistance protein
tnpR	Transposon gamma-delta resolvase
traD	Coupling protein
traI	Multifunctional conjugation protein
xerC	Tyrosine recombinase protein
xerD	Tyrosine recombinase protein

4.3.4 Long plasmid homology search and alignment

We examined the six long contigs (lengths > 20 Kb) classified as plasmid-derived by aligning them with a database of 10,892 complete plasmids (Carattoli et al. 2014) to identify the most similar plasmids using BLAST matches spanning more than one gene (match length > 1,000 bp) with a sequence ID threshold of 95%. This showed the most similar plasmids were isolates were spread across *Enterobacteriaceae* for five and one was in Gammaproteobacteria *Shewanella bicestrii* (VRES1160's plasmid), and that relatively high matching levels were detected for VREC0693's and VREC1428's plasmids, but not for VREC1013, VREC1073 nor VRES1160. The best match to *bla*$_{\text{CTX-M-15}}$-positive VRES1160's IncFIA 61,934 bp plasmid was to *S. bicestrii* strain JAB-1's 193,338 bp plasmid pSHE-CTX-M (NZ_CP022359) that had a length for matches >1 Kb of 30,225 bp. The best match to VREC0693's IncFIB 132,042 bp plasmid was to *Klebsiella pneumoniae* strain Kpn555's 142,858 bp plasmid pKPN-7c3 (NZ_CP015131) that had a length for matches >1 Kb of 98,455 bp. The best match to VREC0693's IncB 88,790 bp plasmid was to *Salmonella enterica* strain ST4/74 was for an 86,908 bp plasmid TY474p2 (NC_017675) that had a length for matches >1 Kb of 77,323 bp. The best match to *bla*$_{\text{CTX-M-15}}$-positive VREC1013's IncFII 89,945 bp plasmid was to *E. coli* strain M19's 11,321 bp plasmid D (NZ_CP010225) that had a length for matches >1 Kb of 5,925 bp. The best match to VREC1073's IncFIA 156,298 bp plasmid was to *Klebsiella pneumoniae* strain SKGH01 84,941 bp plasmid unnamed 3 (NZ_CP015503) that had a length for matches >1 Kb of 39,187 bp. The best match to *bla*$_{\text{CTX-M-27}}$-positive VREC1428's IncFIA plasmid was to *Shigella sonnei* strain 2015C-3566 was for a 55,820 bp plasmid unnamed1(NZ_CP022458) that had a length for matches >1 Kb of 53,995 bp.

4.3.5 Phylogenetic context of analysed isolates

Comparison of these six samples with 119 published ST131 (Ben Zakour et al. 2016; Matsumura et al. 2017) as short read assemblies scaffolded using reference genome NCTC13441 showed that all clustered in ST131 clade C (Figure 4.8). There was sufficient resolution across 4,457 core genome SNPs to confidently assign them to subclades C1 (n=1) or C2 (n=5) (Figure 4.9). VRES1160, VREC0693, VREC1013, VRES0739 and VREC1073 clustered with C2, whereas the *bla*$_{\text{CTX-M-27}}$-positive VREC1428 was in C1.

VRES1160, VREC0693 and VREC1013 all had IncF plasmids (IncFIA, IncFIB, IncFII) and *bla*$_{\text{CTX-M-15}}$ genes, consistent with C2 are typically *bla*$_{\text{CTX-M-15}}$-positive, which was observed for 77% of C2 isolates here (48 out of 62). However, VREC1073 was in C2 but had an a *bla*$_{\text{CTX-M-14}}$ gene, contradicting this pattern and was the sole *bla*$_{\text{CTX-M-14}}$-positive C2 isolate found here. The core genomes of VRES0739 and VREC0693 were identical, implying that VRES0739 has very recently lost its (*bla*$_{\text{CTX-M}}$-positive IncF) plasmid. The sole isolate clustering with C1 was VREC1428, which had an IncFIA plasmid with a *bla*$_{\text{CTX-M-27}}$ gene, and so may belong to the emerging subclade C1-M27 as evidenced by the presence of prophage-like regions like M27PP1/2 (Matsumura et al. 2017).

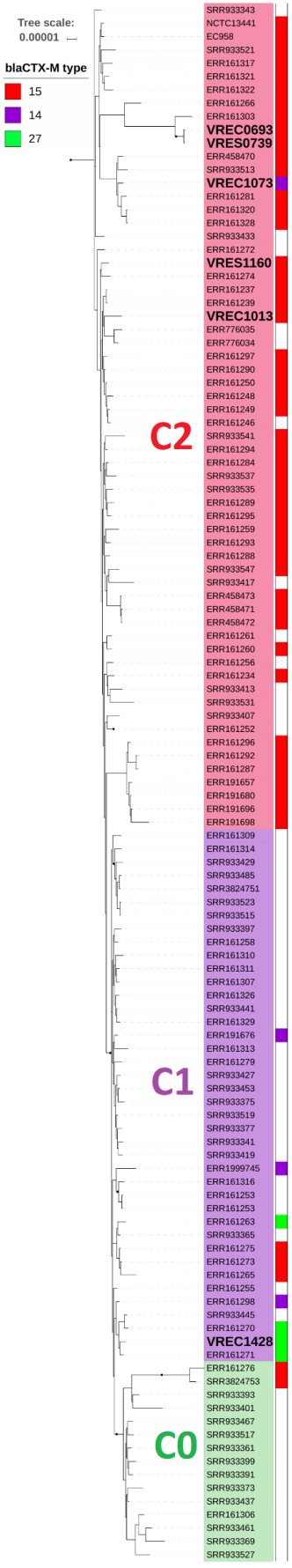


Figure 4.8. Phylogram of the six ST131 genomes showed that all except VREC1428 were in ST131 subclade C2 (red: VRES1160, VREC1073, VRES0739, VREC0693 and VREC1013). VREC1428 clustered in subclade C1 (purple). No new isolate was in C0 (green). The phylogram was built with RAxML v.8.2.11 and iTOL v4.3 using 3,603 non-recombinant SNPs from Gubbins v.2.3.4 where branch support was performed by 100 bootstrap replicates, and the scale bar indicates the number of substitutions per site. Clade classification was based on phylogenetic analysis by (Ben Zakour et al. 2016) by including the reference NCTC13441, n=63 isolates from (Ben Zakour et al. 2016) and n=56 from (Page et al. 2015) with associated classification and *bla*$_{\text{CTX-M}}$ allele data. The right-hand part shows *bla*$_{\text{CTX-M-15}}$ (red), *bla*$_{\text{CTX-M-14}}$ (purple) and *bla*$_{\text{CTX-M-27}}$ alleles (green). The six isolates' names are in large bold text. This mid-pointed rooted phylogeny included reference genome isolates EC958 and NCTC13441 (both in C2) and a clade B isolate as an outgroup (Figure 4.3). The C2 isolates were mainly *bla*$_{\text{CTX-M-15}}$-positive (48 out of 62, including VRES1160, VRES0739, VREC0693 and VREC1013), bar 13 that were *bla*$_{\text{CTX-M-}}$-negative and one that was *bla*$_{\text{CTX-M-14}}$-positive (VREC1073). The C0 isolates were mainly *bla*$_{\text{CTX-M-15}}$-negative (13 out of 15), as were the C1 (30 out of 40) isolates except for four that were *bla*$_{\text{CTX-M-27}}$-positive, three that were *bla*$_{\text{CTX-M-15}}$-positive and three that were *bla*$_{\text{CTX-M-14}}$-positive.

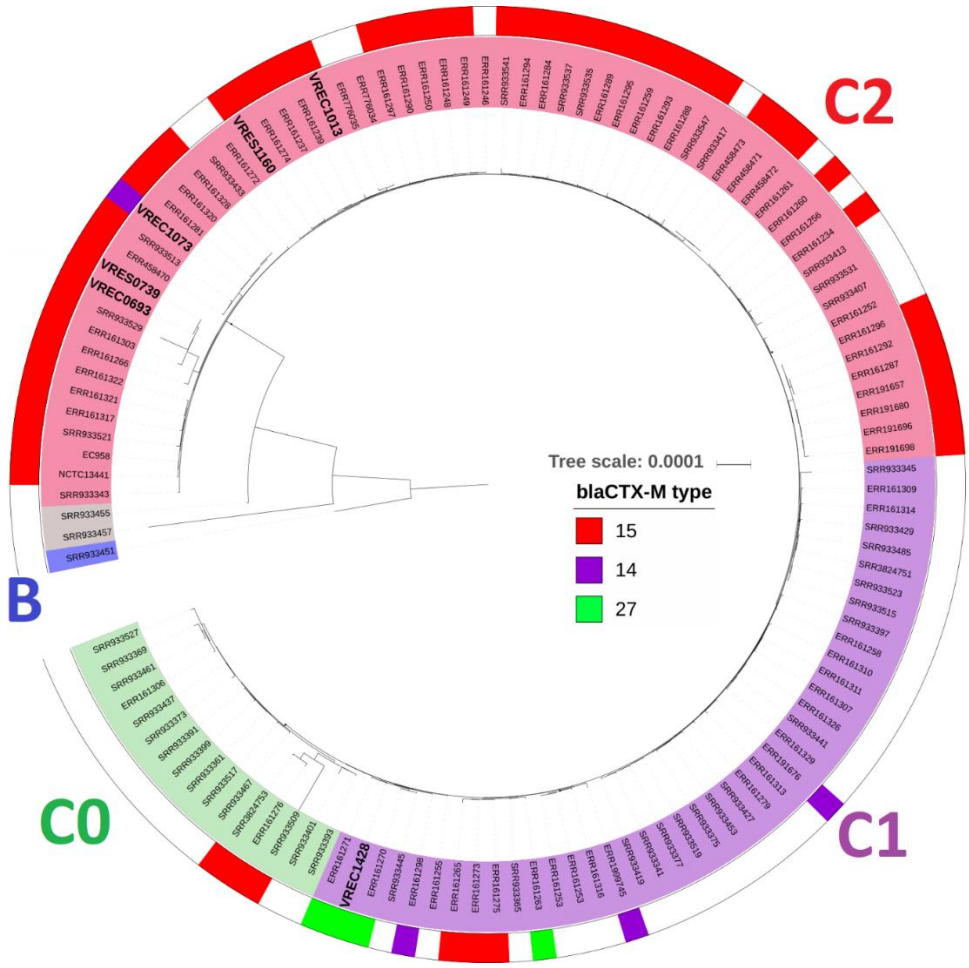


Figure 4.9. The phylogenetic context of the six ST131 genomes (names are in large bold font) showed that all except VREC1428 were in ST131 subclade C2 (red inner ring: VRES1160, VREC1073, VRES0739, VREC0693 and VREC1013). VREC1428 clustered in subclade C1 (purple inner ring). No new isolate clustered in C0 (green inner ring), B (blue inner ring) or an intermediate cluster (grey inner ring). Clade classification was based on phylogenetic analysis by (Ben Zakour et al. 2016) by including the reference NCTC13441, n=63 isolates from (Ben Zakour et al. 2016) and n=56 from (Matsumura et al. 2017) with associated classification and *bla*$_{\text{CTX-M}}$ allele data. VREC1073, and VREC0693 had chromosomal *bla*$_{\text{CTX-M}}$ genes. The outer ring shows *bla*$_{\text{CTX-M-15}}$ (red), *bla*$_{\text{CTX-M-14}}$ (purple) and *bla*$_{\text{CTX-M-27}}$ alleles (green). The phylogeny was built with RAxML v8.2.11 using 4,457 SNPs from a core genome alignment generated with Roary v3.11.2 and was visualised with iTOL v4.3. Branch support was performed by 100 bootstrap replicates, and the scale bar indicates the number of substitutions per site. This mid-pointed rooted phylogeny includes reference genome isolates EC958 and NCTC13441 (both in C2).

4.4 Discussion

My study resolved the plasmid architecture of several recent *E. coli* ST131 isolates, allowing investigation of AMR gene location, copy number and potential transposon-driven rearrangements. This advance was facilitated by the careful DNA handling during extraction to produce large volumes of high molecular weight DNA that was pure and free from contamination, which was avoided by performing separate extraction steps to obtain small plasmids (Lemon et al. 2017) overcoming a limitation for MinION sequencing (Wick et al. 2017a).

The long read genome assemblies illuminated significant variation in plasmids, MGEs and *bla*$_{\text{CTX-M}}$ gene composition that was not captured by short reads. ST131 is a globally pandemic *E. coli* clonal group (Nicolas-Chanoine et al. 2014) with diverse sources of transmission (Roer et al. 2018). Phylogenetic comparison with published genomes (Ben Zakour et al. 2016; Matsumura et al. 2017) showed that five out of six isolates were from subclade C2 with one from C1. The emergence of clade C has been associated with IncF plasmids, and clade C2 with IS*Ecp1* and Tn*2* elements flanking *bla*$_{\text{CTX-M-15}}$ genes (Stoesser et al. 2013; Branger et al. 2018). Our long read assemblies showed the excision of the entire TU from the IncFIB plasmid and chromosomal integration at three distinct locations for VREC0693, and similarly chromosomal translocation of the *bla*$_{\text{CTX-M-14}}$ gene from an IncFIA plasmid for VREC1073, mediated by IS*Ecp1* and IS*903B* based on previous work (Ben Zakour et al. 2016). These transposition events were likely driven by recombination at adjacent transposable elements. This highlights the value of long read sequencing to resolve the location of *bla*$_{\text{CTX-M}}$ genes and that chromosomal translocations are not rare in ST131.

A high resolution of the AMR gene structure, context and copy number is highly predictive of AMR phenotypes (Greig et al. 2018) and could lead to new insights into AMR mechanisms. However, the high indel and mismatch errors in long Oxford Nanopore reads (George et al. 2017; Lemon et al. 2017; Greig et al. 2018; Wang et al. 2014) limits power to identify AMR isoforms that could permit genome-based antimicrobial susceptibility testing (Partridge et al. 2018; Tamma et al. 2018; Tyson et al. 2015). Here, the five ONT assemblies together had an average of 447-fold higher indel and 48-fold

higher mismatch error rates than those for the corresponding Illumina reads, similar to previous work with MinION reads (Judge et al. 2016), and this impacted gene identification. Consequently, short reads and assembly polishing methods remain important for SNP identification and error detection until long read error rates can be reduced (Su et al. 2018).

My findings illustrated the diversity of AMR gene context even within recently emerged clones such as ExPEC ST131. The detection of multiple instances of chromosomally integrated ESBL genes using long reads here for *bla*$_{\text{CTX-M-15}}$ in *E. coli* has parallels elsewhere for *bla*$_{\text{OXA-181}}$ in *bla*$_{\text{CTX-M-15}}$-positive *K. pneumoniae* (Lutgring et al. 2018) and so highlights chromosomal ESBL gene IS*Ecp1*-mediated transposition as a potential adaptive mechanism in *Enterobacteriaceae*. Further studies are needed with larger sample sizes to identify the rates and mechanisms of these dynamic changes.

The sample size issue will be tackled directly in Chapter 5 by assembling the genome of a large number (4,071) of *E. coli* ST131. This is possible in part due to my work in this Chapter, the first to perform near-complete assembly of ST131 genome using long reads, and so is a novel contribution in this area. This is reflected in its publication in *mSphere*. Given these long read scaffolds and confirmation of the phylogenetic context of C1 and C2 along with their *bla*$_{\text{CTX-M}}$ gene isoforms, short read assembly can reveal the frequencies of known isoforms in ST131 globally. The latter may be limited in discovering novel ESBL context and associated, which was directly investigated in this chapter.

4.5 Data Summary

1. Illumina reads accession numbers: ERR2138475, ERR2138200, ERR2138591, ERR1878196, ERR2137889 and ERR1878359 in the European Nucleotide Archive (ENA) under BioProjects PRJEB21499 and PRJEB19918.

2. ONT reads accession numbers: ERR3284704, ERR328470, ERR3284706, ERR3284707, ERR3284708 and ERR3284709 - see www.ebi.ac.uk/ena/data/view/PRJEB30511 in the ENA or www.ncbi.nlm.nih.gov/sra/PRJEB30511 in the SRA under BioProject PRJEB30511; see also Figshare https://doi.org/10.6084/m9.figshare.7554293.v1

3. Unicycler assemblies: Figshare https://doi.org/10.6084/m9.figshare.7560458.v2

Ethical approval

The study protocol was approved by the National Research Ethics Service (ref:14/EE/1123), and the Cambridge University Hospitals NHS Foundation Trust Research and Development Department (ref: A093285).

Acknowledgements

We acknowledge Anne Parle-McDermott and Emma Finlay at Dublin City University (DCU, Ireland) for guidance on DNA extraction protocols, and also Emma Betteridge, Karen Oliver and the Long Read sequencing and data teams at the Wellcome Sanger Institute (U.K.) for their assistance with sequencing.

4.6 References

Arredondo-Alonso S, Rogers MRC, Braat JC, Verschuuren TD, Top J, Corander J, Willems RJL, Schürch AC. mlplasmids: a user-friendly tool to predict plasmid- and chromosome-derived sequences for single species. Microb Genom. 2018 4(11). doi: 10.1099/mgen.0.000224.

Barlow M, Reik RA, Jacobs SD, Medina M, Meyer MP, McGowan JE Jr, Tenover FC. High rate of mobilization for blaCTX-Ms. Emerg Infect Dis. 2008 14(3):423-8. doi: 10.3201/eid1403.070405.

Ben Zakour NL, Alsheikh-Hussain AS, Ashcroft MM, Khanh Nhu NT, Roberts LW, Stanton-Cook M, Schembri MA, Beatson SA. Sequential acquisition of virulence and fluoroquinolone resistance has shaped the evolution of *Escherichia coli* ST131. MBio. 2016 7(2):e00347-16. doi: 10.1128/mBio.00347-16.

Branger C, Ledda A, Billard-Pomares T, Doublet B, Fouteau S, Barbe V, Roche D, Cruveiller S, Médigue C, Castellanos M, Decré D, Drieux-Rouze L, Clermont O, Glodt J, Tenaillon O, Cloeckaert A, Arlet G, Denamur E. Extended-spectrum β-lactamase-encoding genes are spreading on a wide range of *Escherichia coli* plasmids existing prior to the use of third-generation cephalosporins. Microb Genom. 2018 4(9). doi: 10.1099/mgen.0.000203.

Brooks L, Kaze M, Sistrom M. A Curated, Comprehensive Database of Plasmid Sequences. Microbiol Resour Announc. 2019 8(1). pii: e01325-18. doi: 10.1128/MRA.01325-18

Calhau V, Ribeiro G, Mendonça N, Da Silva GJ. Prevalent combination of virulence and plasmidic-encoded resistance in ST131 *Escherichia coli* strains. Virulence. 2013 4(8):726-9. doi: 10.4161/viru.26552.

Carattoli A, Zankari E, García-Fernández A, Voldby Larsen M, Lund O, Villa L, Møller Aarestrup F, Hasman H. In silico detection and typing of plasmids using

PlasmidFinder and plasmid multilocus sequence typing. Antimicrob Agents Chemother. 2014 58(7):3895-903. doi: 10.1128/AAC.02412-14.

Chan KG, Chong TM, Yin WF, Upton M, Schembri MA, Beatson SA. Lineage-specific methyltransferases define the methylome of the globally disseminated *Escherichia coli* ST131 clone. MBio. 2015 6(6):e01602-15. doi: 10.1128/mBio.01602-15.

Chen S, Zhou Y, Chen Y, Gu J. Fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018 34(17):i884–i890.

Ender PT, Gajanana D, Johnston B, Clabots C, Tamarkin FJ, Johnson JR. Transmission of an extended-spectrum-β-lactamase-producing *Escherichia coli* (sequence type ST131) strain between a father and daughter resulting in septic shock and Emphysematous pyelonephritis. J Clin Microbiol. 2009 47(11):3780-2. doi: 10.1128/JCM.01361-09.

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.

Forde BM, Phan MD, Gawthorne JA, Ashcroft MM, Stanton-Cook M, Sarkar S, Peters KM, Frost LS, Leplae R, Summers AO, Toussaint A. Mobile genetic elements: the agents of open source evolution. Nat Rev Microbiol. 2005 3(9):722-32.

George S, Pankhurst L, Hubbard A, Votintseva A, Stoesser N et al. Resolving plasmid structures in *Enterobacteriaceae* using the MinION nanopore sequencer: assessment of MinION and MinION/Illumina hybrid data assembly approaches. Microb Genom 2017:1–8.

Goldstein S, Beka L, Graf J, Klassen J. Evaluation of strategies for the assembly of diverse bacterial genomes using MinION long-read sequencing. 2018 Biorxiv doi: https://doi.org/10.1101/362673

Goswami C, Fox S, Holden M, Connor M, Leanord A, Evans TJ. Genetic analysis of invasive *Escherichia coli* in Scotland reveals determinants of healthcare-associated versus community-acquired infections. Microb Genom. 2018 4(6). doi: 10.1099/mgen.0.000190.

Greig DR, Dallman TJ, Hopkins KL, Jenkins C. MinION nanopore sequencing identifies the position and structure of bacterial antibiotic resistance determinants in a multidrug-resistant strain of enteroaggregative *Escherichia coli*. Microb Genom. 2018 4(10). doi: 10.1099/mgen.0.000213.

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.
Harrison E, Brockhurst MA. Plasmid-mediated horizontal gene transfer is a coevolutionary process. Trends Microbiol. 2012 20(6):262-7. doi: 10.1016/j.tim.2012.04.003

Hinnebusch J, Tilly K. Linear plasmids and chromosomes in bacteria. Mol Microbiol. 1993 10(5):917-22.

Johnson JR, Johnston B, Clabots C, Kuskowski MA, Castanheira M. *Escherichia coli* sequence type ST131 as the major cause of serious multidrug-resistant *E. coli* infections in the United States. Clin Infect Dis. 2010 51(3):286-94. doi: 10.1086/653932.

Judge K, Hunt M, Reuter S, Tracey A, Quail MA, Parkhill J, Peacock SJ. Comparison of bacterial genome assembly software for MinION data and their applicability to medical microbiology. Microb Genom. 2016 2(9):e000085. doi: 10.1099/mgen.0.000085

Juhas M van der Meer JR, Gaillard M, Harding RM, Hood DW, Crook DW. Genomic islands: tools of bacterial horizontal gene transfer and evolution. FEMS Microbiol Rev. 2009 33(2):376-93. doi: 10.1111/j.1574-6976.2008.00136.x

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

Lanfear R, Schalamun M, Kainer D, Wang W, Schwessinger B. MinIONQC: fast and simple quality control for MinION sequencing data. Bioinformatics. 2018 doi: 10.1093/bioinformatics/bty654

Lartigue MF, Poirel L, Aubert D, Nordmann P. In vitro analysis of IS*Ecp1B*-mediated mobilization of naturally occurring β-lactamase gene blaCTX-M of *Kluyvera ascorbata*. Antimicrob Agents Chemother. 2006 50(4):1282-6.

Leggett RM, Clark MD. A world of opportunities with nanopore sequencing. J Exp Bot. 2017 68(20):5419-5429. doi: 10.1093/jxb/erx289.

Lemon JK, Khil PP, Frank KM, Dekker JP. Rapid Nanopore Sequencing of Plasmids and Resistance Gene Detection in Clinical Isolates. J Clin Microbiol. 2017 55(12):3530-3543. doi: 10.1128/JCM.01069-17.

Ludden C, Reuter S, Judge K, Gouliouris T, Blane B, Coll F, Naydenova P, Hunt M, Tracey A, Hopkins KL, Brown NM, Woodford N, Parkhill J, Peacock SJ. Sharing of carbapenemase-encoding plasmids between *Enterobacteriaceae* in UK sewage uncovered by MinION sequencing. Microb Genom. 2017 3(7):e000114. doi: 10.1099/mgen.0.000114.

Lutgring JD, Zhu W, de Man TJB, Avillan JJ, Anderson KF, Lonsway DR, Rowe LA, Batra D, Rasheed JK, Limbago BM. Phenotypic and Genotypic Characterization of *Enterobacteriaceae* Producing Oxacillinase-48-Like Carbapenemases, United States. Emerg Infect Dis. 2018 24(4):700-709. doi: 10.3201/eid2404.171377.

MacLean RC, San Millan A. Microbial Evolution: Towards Resolving the Plasmid Paradox. Curr Biol. 2015 25(17):R764-7. doi: 10.1016/j.cub.2015.07.006

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). pii: e00179-17. doi: 10.1128/AAC.00179-17.

McNally A, Oren Y, Kelly D, Pascoe B, Dunn S, Sreecharan T, Vehkala M, Välimäki 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. Combined Analysis of Variation in Core, Accessory and Regulatory Genome Regions Provides a Super-

Nicolas-Chanoine MH, Bertrand X, Madec JY. *Escherichia coli* ST131, an intriguing clonal group. Clin Microbiol Rev. 2014 27(3):543-74. doi:10.1128/CMR.00125-13.

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

Partridge SR, Kwong SM, Firth N, Jensen SO. Mobile genetic elements associated with antimicrobial resistance. Clinical Microbiology Reviews 31(4):e00088-17 2018 doi: 10.1128/CMR.00088-17

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.

Petty NK, Ben Zakour NL, 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, Phan MD, Forde BM, Peters KM, Sarkar S, Hancock S, Stanton-Cook M, Ben Zakour NL, Pitout JD, Upton M, Paterson DL, Walsh TR, Schembri MA, Beatson SA. Global dissemination of a multidrug resistant *Escherichia coli* clone. Proc Natl Acad Sci U S A. 2014 111(15):5694-9. doi: 10.1073/pnas.1322678111

Pitout JDD, DeVinney R. *Escherichia coli* ST131: a multidrug-resistant clone primed for global domination. *F1000Research* 2017 doi: 10.12688/f1000research.10609.1

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.
Resolution View into the Evolution of Bacterial Populations. PLoS Genet. 2016 12(9):e1006280. doi: 10.1371/journal.pgen.1006280.

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 fimH22 *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. JAntimicrob Chemother. 2018 doi: 10.1093/jac/dky484.

Schembri MA, Zakour NL, Phan MD, Forde BM, Stanton-Cook M, Beatson SA. 2015. Molecular characterization of the multidrug resistant Escherichia coli ST131 clone. Pathogens 4(3):422–30.

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

Shintani M, Sanchez ZK, Kimbara K. Genomics of microbial plasmids: classification and identification based on replication and transfer systems and host taxonomy. Front Microbiol 2015;6.

Sidjabat HE, Townell N, Nimmo GR, George NM, Robson J. 2015. Dominance of IMP-4-producing Enterobacter cloacae among carbapenemase-producing Enterobacteriaceae in Australia. Antimicrob Agents Chemother, 59:4059–4066.

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, Batty EM, Eyre DW, Morgan M, Wyllie DH, Del Ojo Elias C, Johnson JR, Walker AS, Peto TEA, Crook DW. Predicting antimicrobial susceptibilities for *Escherichia coli* and *Klebsiella pneumoniae* isolates using whole genomic sequence data. J Antimicrob Chemother 2013 68:2234-2244.

Su M, Satola SW, Read TD. Genome-based prediction of bacterial antibiotic resistance. J Clin Microbiol. 2018 doi: 10.1128/JCM.01405-18.

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

Tamma PD, Y Fan, Bergman Y, Pertea G, Kazmi A, Lewis S, Carroll KC, Schatz MC, Timp W, Simner P. Rapid optimization of antibiotic therapy for multidrug-resistant gram-negative infections using Nanopore whole genome sequencing. 2018 Available at SSRN: https://ssrn.com/abstract=3219539.

Totsika M, Beatson SA, Sarkar S, Phan MD, Petty NK, Bachmann N, Szubert M, Sidjabat HE, Paterson DL, Upton M, Schembri MA. Insights into a multidrug resistant *Escherichia coli* pathogen of the globally disseminated ST131 lineage: genome analysis and virulence mechanisms. PLoS One. 2011 6(10):e26578. doi: 10.1371/journal.pone.0026578.

Tyson GH, McDermott PF, Li C, Chen Y, Tadesse DA, Mukherjee S, Bodeis -Jones S, Kabera C, Gaines SA, Loneragan GH, Edrington TS, Torrence M, Harhay DM, Zhao S. WGS accurately predicts antimicrobial resistance in *Escherichia coli*. J Antimicrob Chemother 2015 70:2763-2769.

Upton M, Beatson SA, Schembri MA. Molecular characterization of a multidrug resistance IncF plasmid from the globally disseminated *Escherichia coli* ST131 clone. PLoS One. 2015 10(4):e0122369. doi: 10.1371/journal.pone.0122369.

Van der Bij AK, Peirano G, Pitondo-Silva A, Pitout JD. The presence of genes encoding for different virulence factors in clonally related *Escherichia coli* that produce CTX-Ms. Diagn Microbiol Infect Dis. 2012 72(4):297-302. doi: 10.1016/j.diagmicrobio.2011.12.011

Wang Y, Yang Q, Wang Z. The evolution of nanopore sequencing. Front Genet 2014 5:449.

Waterhouse RM, Seppey M, Simão FA, Manni M, Ioannidis P, Klioutchnikov G, Kriventseva EV, Zdobnov EM. BUSCO applications from quality assessments to gene prediction and phylogenomics. Mol Biol Evol. 2017 doi: 10.1093/molbev/msx319.

Wick RR, Judd LM, Gorrie CL, Holt KE. Completing bacterial genome assemblies with multiplex MinION sequencing. Microb Genom. 2017a 3(10):e000132. doi: 10.1099/mgen.0.000132.

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

Wick RR, Schultz MB, Zobel J, Holt KE. Bandage: interactive visualization of de novo genome assemblies. Bioinformatics. 2015 31(20):3350-2. doi: 10.1093/bioinformatics/btv383.

Woodford N, Carattoli A, Karisik E, Underwood A, Ellington MJ, Livermore DM. Complete nucleotide sequences of plasmids pEK204, pEK499, and pEK516, encoding CTX-M enzymes in three major *Escherichia coli* lineages from the United Kingdom, all belonging to the international O25:H4-ST131 clone. Antimicrob Agents Chemother. 2009 53(10):4472-82. doi: 10.1128/AAC.00688-09.

Chapter 5: The origin, evolution and population structure of 4,071 *E. coli* ST131 genomes

Abstract

Escherichia coli ST131 is a major cause of infection with extensive antimicrobial resistance (AMR) associated with the widespread use of beta-lactam antibiotics. This drug pressure has driven extended-spectrum β-lactamase (ESBL) gene acquisition and evolution in pathogens like ST131 and so a high resolution of the origin, evolution and spread of both ST131 and its ESBL genes is essential. These ESBL genes are embedded in mobile genetic elements (MGEs), which aid their transfer to new plasmid or chromosomal locations, during which these ESBL genes may be amplified, truncated or mutated. Plasmid recombination and conjugation can mobilise these ESBLs in MGEs further, thus entailing large-scale genomic epidemiology to investigate these processes more precisely. ST131 is as a paradigm for gram-negative bacteria that have a monomorphic core genome contrasting with a dynamic ESBL, MGE and plasmid composition. We extracted all available high-quality ST131 Illumina HiSeq read libraries to resolve the global population structure of ST131's three genetically distinct clades (A, B and C) and their subclades. We applied rigorous quality-control, genome *de novo* assembly and ESBL gene screening to the largest ST131 collection examined of 4,071 genomes. We reconstructed their evolutionary relationships across their core and accessory genomes by exploiting the published reference genomes, Nanopore and PacBio assemblies to use k-mer-based methods to contextualise pangenome diversity. We focus on the subclades of the most abundant clade (C) to provide a deep resolution of the epidemiology and genomic context of key ESBL genes. We show that core genome diversity within subclades is not correlated with that of the hypervariable accessory genome, including plasmids and key ESBL genes. Our findings underpin the potential to improve our understanding of the ESBL gene origin, evolution and spread using evolutionary pangenomics that may inform on accessory genome changes linked to emerging ST131 outbreaks.

Publication: in preparation for *mBio* 2019 with Downing T.

5.1 Introduction

Infections caused by multidrug-resistant (MDR) *Escherichia coli* sequence type (ST) 131 (the ST131 complex) are increasing worldwide (de Kraker et al. 2013, Poolman & Wacker 2016). ST131 are a type of extraintestinal pathogenic *E. coli* (ExPEC) that causes a significant amount of bloodstream and urinary tract infections globally and typically possess extended-spectrum β-lactamase (ESBL) (Banerjee & Johnson 2014), or more rarely carbapenemase (Peirano et al. 2011) genes. MDR ST131 is a major cause of ExPEC infections, because it has a 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 thus may be more pathogenic (Dautzenberg et al. 2016). ST131 is reported from around the globe, both in healthcare settings and in the community and is nearly always fluoroquinolone resistant (FQ-R) (Ben Zakour et al. 2016, Stoesser et al. 2016). The most predominant lineage within ST131 is known as clade C: this is FQ-R and has a *H30* variant of the type 1 fimbrial adhesin gene (*fimH*30) (Price et al. 2013, Petty et al. 2014). In contrast to the FQ-susceptible clades A and B, C can offset fitness costs of antimicrobial resistance (AMR), plasmid acquisition and maintenance through compensatory mutations at regulatory regions (McNally et al. 2016).

Understanding the mechanisms of AMR, host colonisation and pathogenicity in MDR ST131 requires a deep investigation of its population structure, selective process and the mechanisms by which its ESBL genes spread (Ben Zakour et al. 2016, Stoesser et al. 2016). Exploring the evolutionary origins, transmission and spread of outbreaks requires extensive sampling to contextualise the variation at key genes while simultaneously inferring the epidemiology and population structure (Croucher and Didelot, 2015). High-resolution large-scale bacterial epidemiology inferred from genomic data can address these questions (Lees *et al.* 2018).

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 that is only possible with whole genome sequencing, which also allows

profiling of all AMR genes (Sintchenko & Holmes 2015, Revez et al. 2017). A recent example of this applying cgMLST (core genome MLST) incorporated most (2,512) genes in the *E. coli* core genome: (Zhou et al. 2019). Computational efficiency has limited previous work, including one with 288 ST131 genomes in the context of 9,479 diverse *E. coli* such that only one specimen per rST was examined across 1,230,995 SNPs found in a 2.33 Mb core genome (Zhou et al. 2019). Given that rST1503 alone may account for approximately 81% of ST131 and that outbreaks may comprise a single rST (Ludden, Decano et al. 2019), investigating large isolate collections of individual STs can inform on past, present and emerging MDR ST131 outbreaks.

Deciphering the evolutionary relationships of a large ST131 collection based on its core genome provides a stable foundation to explore their accessory genomes. The gradual evolution of ST131 has been punctuated by plasmid conjugation, plasmid recombination and mobile genetic element (MGE) rearrangements of the cefotaximase (CTX-M) class of ESBLs (Canton et al. 2012, Decano et al. 2019) allowing resistance to 3rd-generation cephalosporins, particularly by *bla*$_{\text{CTX-M-14/15/27}}$ in ST131 (Mathers et al. 2015). This accessory genome dynamism is correlated strongly with the high prevalence of *bla*$_{\text{CTX-M-15}}$-positive ST131 subclade C2, the most common subclade (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 context results in negative frequency-dependent selection (NFDS) acting on the ST131 accessory genome, leading to a dynamic AMR gene repertoire (McNally et al. 2019) that has not yet to be explored in ST131's subclades. Consequently, the evolutionary pangenomics of ST131's clades and subclades can identify genetic changes corresponding to key transmission events whose phylogenetic pattern can be linked to outbreaks.

As identified in Chapter 5, investigating a large number of *E. coli* ST131 genomes has not yet been achieved at the scale planned here. This extended work on long read sequencing and assembly given the known ESBL gene isoforms, plasmids and subclade genetic similarities. Here, we aggregated all available ST131 sequence read libraries. I automated quality-control, genome *de novo* assembly, DNA read mapping and ESBL gene screening

in the largest ST131 sample collection examined thus far to reconstruct the core genome phylogenetic history as a basis to evaluate the distribution of clades and subclades across geographic region over time. We establish that the two most common C subclades (C1 and C2) have been and continue to co-circulate globally and that their ancestral *bla*$_{\text{CTX-M}}$ gene composition and context is flexible. We further show that the diversity of accessory genomes in isolates with virtually identical core genomes may be a function of the environment because MDR ST131's open pangenome has extensive AMR gene mobility.

5.2 Methods

5.2.1 Study selection and data extraction

4,870 *E. coli* ST131 genomes and linked metadata were collected using an automated text-mining algorithm that used a Python implementation of Selenium (Selenium-python.readthedocs.io) to extract data related to available *E. coli* ST131 samples from Enterobase (https://enterobase.warwick.ac.uk, Alikhan et al. 2018) on 10th September 2018 in a manner previously described (Kinderis et al. 2018). This ST131 data was used to query the European Nucleotide Archive (ENA) (www.ebi.ac.uk/ena) and NCBI Short Read Archive (SRA) databases. Only read libraries that were complete or not labelled as "traces" were selected for downloading as FASTQ files (Figure 5.1) from the ENA or SRA. Of the initial 4,870 read libraries, 4,264 were paired-end (PE) Illumina Hiseq and four were PacBio, in addition to PacBio-sequenced NCTC13441 genome that was used as a reference here. Although 495 libraries were available that were predominantly sequencing on Illumina MiSeq platforms, these were not examined to avoid platform-specific artefacts.

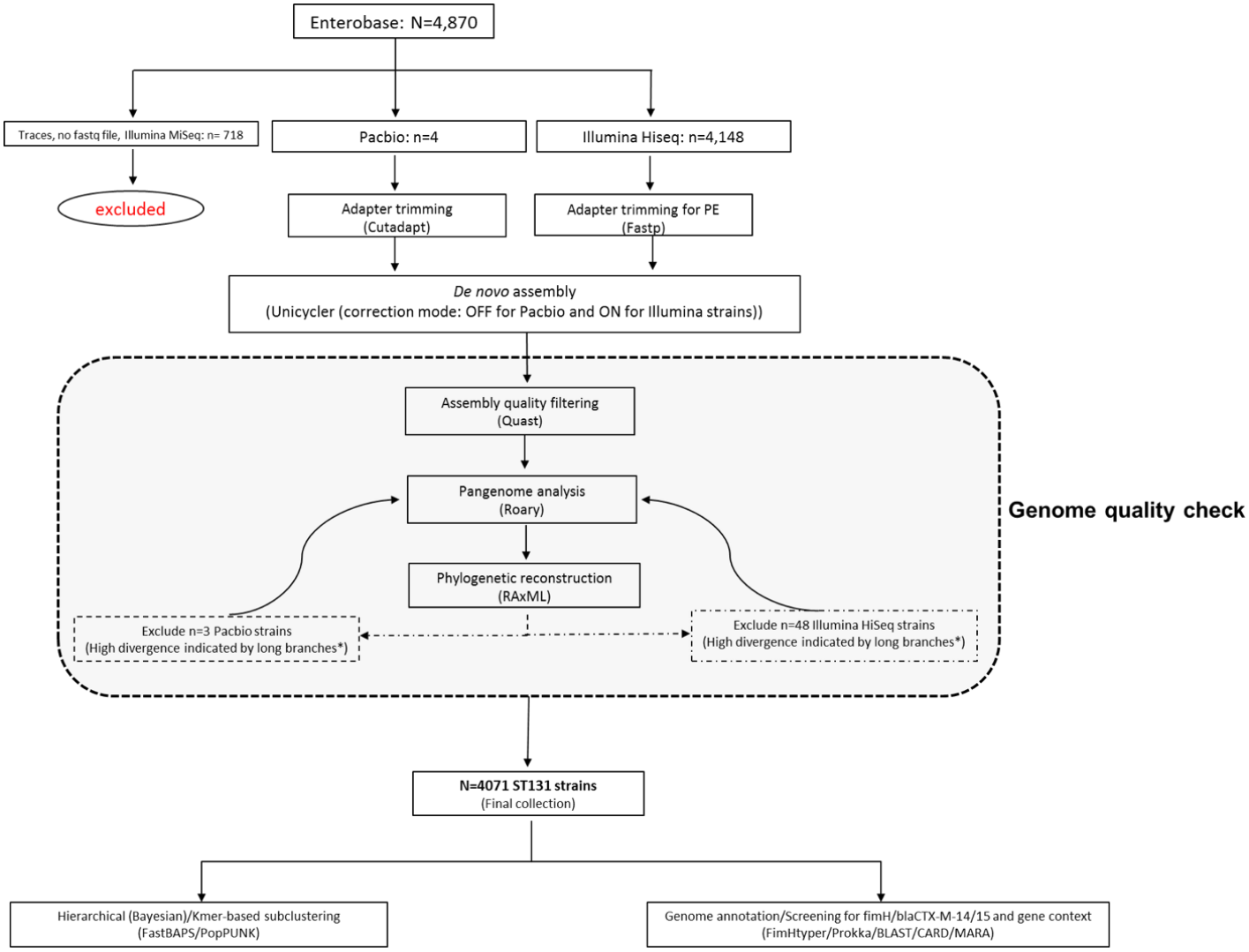


Figure 5.1. Methods summary of population structure of ST131 genomes. N=4,870 read libraries were downloaded from Enterobase on 10 September 2018: 718 were incomplete or have no FASTQ files and were hence excluded for downstream processing, four were long read libraries (Pacbio) and the rest were short paired-end reads (Illumina). The adapters of the four Pacbio and 4,147 Illumina reads were trimmed using Cutadapt and Fastp pipeline, respectively. The resulting adapter-free reads were assembled using Unicycler. The assembled genomes were annotated and screened for AMR genes (including *bla*$_{\text{CTX-M-14}}$ and *bla*$_{\text{CTX-M-15}}$) and their genetic context. Pangenome analysis was based on Roary and Prokka annotation files. Cleaned reads were mapped to the pangenome reference sequence from Roary to estimate gene copy numbers. Phylogenetic reconstruction was performed by running RAxML based on the core genome alignment of the samples; distinct sub-clusters from the produced phylogeny were then determined using Fastbaps (Fast Hierarchical Bayesian Analysis of Population Structure). Distances between the core and accessory genomes of the collection was estimated using PopPUNK (Population Partitioning Using Nucleotide K-mers).

5.2.2 Illumina HiSeq read data quality control, trimming and correction

Of these 4,264 PE Illumina HiSeq quality-checked read libraries, 4,147 passed stringent quality control screening. Quality filtering of Illumina reads was implemented using Fastp v.0.12.3 (Chen *et al.* 2018) to trim sequencing adapters, remove reads with poor 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. Individual bases in reads were corrected by BayesHammer in SPAdes v.3.11.1 using algorithms based on Hamming graphs and Bayesian sub-clustering (Nikolenko et al. 2013). Quality control metrics were examined at each step: on individual FASTQ files using FastQC v0.11.8 (www.bioinformatics.babraham.ac.uk/projects/fastqc/), and across the whole collection as a batch report using MultiQC v1.4 (Ewels et al. 2016). 117 Illumina Hiseq libraries were removed after this process.

5.2.3 Illumina HiSeq read library genome assembly

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

5.2.4 Reference PacBio genome quality control and assembly

The reference genome was NCTC13441, which was isolated in the UK in 2003 and belonged to ST131 subclade C2 (Brodrick et al. 2018). It was previously assembled into 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 v.1.18 (Martin 2011) followed by excluding duplicate reads with Unicycler v4.6. Base correction was implemented during Unicycler genome assembly with SPAdes v3.12, and its assembly was iteratively polished by Racon v1.3.1 (Vaser et al. 2017) until no further corrections were required. This additional 5,132,452 bp reference assembly had just five contigs, was in a different subclade (C1) compared to NCTC13441, had no IS*Ecp1*, and had 5,506 genes.

5.2.5 Genome assembly quality investigation

For the 4,147 Illumina Hiseq and single PacBio assembly, assembly quality was verified with Quast v.5.0 (Gurevich et al. 2013) based on N50, numbers of predicted genes and open-reading frames, and numbers of contigs with misassemblies. The quality of these 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).

5.2.6 Genome annotation identifies 4,071 assemblies for final examination

Initial annotation of the 4,147 Illumina Hiseq assemblies using Prokka v1.10 (Seeman 2014) suggested 77 assemblies had a distinct gene composition, indicating that they should not be included further 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 for the atlas and aligned against the reference genome NCTC13441 and PacBio assembly AR_0058 (Supplementary Table S5.2). This identified 4,829 genes on average per assembly with a minimum of 3,942 and maximum of 5,749 (Supplementary Figure S5.1). The variation in numbers of genes per assembly was largely explained by the total assembly length (r^2=0.959).

5.2.7 Pangenome analysis to identify the core and accessory genomes

We created a pangenome based on the above 4,072 annotation files using Roary v.3.11.2 (Page et al. 2015) with a 100% BLAST v2.6.0 identity threshold using the MAFFT v.7.310 (Katoh & Standley 2013) setting. The pangenome output generated a concatenated core

CDSs alignment spanning 1,244,619 bases and 3,712 genes scaffolded using NCTC13441, which was used for comparison of the core and accessory genomes, and core genome phylogenetics. Pangenomes for each clade and C subclade were also created for accessory (cloud) genome comparison.

5.2.8 Phylogenetic reconstruction to verify subclade assignments

The evolutionary relationships between the strains were inferred by generating a maximum likelihood phylogeny based on the core genome alignment of 4,071 genome assemblies with NCTC13441 as a reference across 1,244,619 sites containing 26,946 alignment patterns (from 30,029 SNPs) for 50 iterations of RAxML v8.2.11 (Stamatakis 2014) with a GTR model and gamma substitution rate heterogeneity. 3,585 (88%) of genome assemblies were genetically unique. The total execution time on an Ubuntu 16.04 computer server with 256 Gb RAM using 52 threads was 24.43 days. The resulting phylograms were drawn and annotated using iTol v4.3.2.

5.2.9 Population structure and subclade assignment

Clade classifications were initially based on published ST131 *fimH* phylogenetic analysis that associated clade A with *fimH41*, B with *fimH22*, B0 with *fimH27*, and C with *fimH30* (Price et al. 2013). To classify the C subclades for a large dataset that can be dissected as a sparse matrix, we used using genetic clustering based on a hierarchical Bayesian clustering algorithm implemented in Fastbaps v1.0 across the 30,029 core genome SNPs (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.

5.2.10 ESBL gene screening and contig visualisation

ESBL gene screening across the 4,071 assemblies' total of 505,761 contigs was implemented to detect contigs with *bla*$_{\text{CTX-M-14/15/27}}$ genes using BLASTn alignment of these three genes individually and the Comprehensive Antibiotic Resistance Database (CARD v3.0) requiring 100% identity for any match versus each assembled 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 gene structure. Frequencies of ST131 clades, subclades and their *bla*$_{\text{CTX-M-14/15/27}}$ genes over geographic region 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.

5.2.11 Accessory genome composition across clades and subclades

The relative pairwise genetic distances of the core (π) and accessory (a) genomes were compared across the pangenome of all 4,071 assemblies, for each clade, each C subclade and all *bla*$_{\text{CTX-M}}$-positive clade C samples using Poppunk (Population Partitioning Using Nucleotide Kmers, Lees et al. 2019). Poppunk has high power to distinguish closely related genomes (Lees et al. 2019) and used variable length DNA 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 pair of samples. This annotation- and alignment-free approach complemented the Fastbaps, RAxML and Roary results.

In addition, we determined the expected shell gene number ($E[a_p]$) from the Roary output for a given pooled set of samples p originally from groups $i=1..k$ based on the shell gene number of group i (a_i) weighted by the 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 percentage excess was determined as: $(a_p - E[a_p])/E[a_p]$. 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:

$$\boldsymbol{E[d_p] = \frac{\sum_{i=1}^{k} n_i d_i}{\sum_{i=1}^{k} n_i} - \frac{\sum_{i=1}^{k} n_i (c_i - c_p)}{\sum_{i=1}^{k} n_i} - \frac{\sum_{i=1}^{k} n_i (s_i - s_p)}{\sum_{i=1}^{k} n_i} - \frac{\sum_{i=1}^{k} n_i (a_i - a_p)}{\sum_{i=1}^{k} n_i}}$$

Again, the percentage excess was determined as: $(d_p - E[d_p])/E[d_p]$.

We quantified pangenome openness (*alpha*) from $\Delta n = kN^{-alpha}$ where Δ*n* was the number of newly added genes across *N* genome assemblies with *n* genes in total as estimated by Roary with R packages poweRlaw v0.70.2, igraph v1.2.4.1 and VGAM v1.1.1. This power-law regression approximates Heaps' law well, such that an open pangenome has *alpha* < 1, and a closed one *alpha* > 1 (Tettelin et al. 2008, Park et al. 2019). Previously, diverse *E. coli* had *alpha = 0.625* where the latter was largely stable with a slight decline as *N* increased (Park et al. 2019), and similarly *alpha* was approximately 0.877 for ST131 clade C, 0.898 for B, 0.958 for A, and 0.951 for all combined, suggesting *alpha* was higher when genetically distinct clades were combined (McNally et al. 2019).

5.3 Results

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

We collated SRA and ENA accession IDs and linked metadata on 4,870 global ST131 strains from Enterobase using a text mining algorithm on 10th September 2018. Of these, 4,267 were genome-sequenced using Illumina HiSeq or PacBio platforms from 188 BioProjects. Following thorough filtering steps, a final collection of 4,071 high quality *de novo* genome assemblies whose DNA was isolated in 1967-2018 from diverse sources across 170 BioProjects (Supplementary Table S5.1) were created for investigation and further analyses (Supplementary Table S5.2). 721 assemblies were not examined further because they were sequenced using a different platforms (Illumina MiSeq), or had poor library base quality metrics.

These 4,070 Illumina HiSeq read libraries assembled using the Unicycler v.0.4.6 pipelines (bold mode) generated draft genomes with a mean N50 of 195,830±57,037 bp (mean ± standard deviation), a mean assembly length of 5,136,890±121,402 bp, an average of 124.3±74.8 contigs, and an average of 4,829±142 genes (Supplementary Table S5.2). The final assembly was generated from PacBio reads (AR_0058) to make five contigs with a N50 of 4,923,470 bp and an assembly length of 5,132,452 bp.

5.3.2 A ST131 core genome of 3,712 genes and an accessory genome of 22,525 genes

We assembled the 4,071 assemblies' pangenome using NCTC13441 as the reference with Roary (Page *et al.* 2015) resulting in 26,479 genes, most of which were rare. The hard core genome was composed of 3,712 genes present in all samples (100%), though an additional 242 comprised the soft core genome (present in ≥95% of samples) (Supplementary FigureS5.3). 22,525 CDSs formed the accessory genome, along with 242 in the soft core (>95% of samples), 1,018 shell genes found in 15-95% of samples, and 21,507 (81% of the total) cloud genes in <15% of samples.

5.3.3 Population structure classification shows three dominant ST131 C subclades

Of the final 4,071 assemblies, clades A (n=414, 10.1% of the total), B (n=420, 10.3%) and B0 (n=13, 0.3%) were relatively rare in comparison to the 3,224 assigned to clade C (79%). This was based on *fimH* typing that showed 91% of clade A had *fimH41*, 66% of clade B had *fimH22*, 99% of clade C had *fimH30,* and unexpectedly all 13 isolates in subclade B0 were in *fimH30*, rather than *fimH27* (Table 5.1). Nine isolates were *fimH54*, of which eight were in clade B (Matsumura et al. 2017).

ST131 subclades within the main clades were determined by clustering based on 30,029 core SNPs with Fastbaps v1.0 (Figure 5.2). This came from the core genome of 3,712 genes spanning 1,244,619 bp of the reference NCTC13441 chromosome and did not include the 242 soft core genes whose absence may have been due to assembly errors or other artefacts (Supplementary Figure S5.1). This divided the 4,071 isolates into nine genetically distinct subclades (clusters 1-9) and two groups of unassigned isolates (clusters 10 and 11) (Supplementary Figure S5.2). Clade A was mainly assigned to cluster 2 (n=407, 98.3%) and seven were unassigned (cluster 11). 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%) that were unassigned (n=34 in cluster 10, n=17 in cluster 11). Subclade B0 strains entirely belong to cluster 8, suggesting that this group could be considered one of a number of lineages in clade B.

Although clade C was dominant, it had only three main subclades determined by Fastbaps. C0 (n=52) was mainly assigned to cluster 11, consistent with its heterogeneous classification (Ben Zakour et al. 2016). C1 was composed of 1,121 isolates, 1,113 of which were in cluster 6 (referred to as C1_6) with eight unassigned in cluster 10 (Figure 5.3). C2 had 2,051 isolates, of which 1,651 were assigned to cluster 9 (C2_9) and 386 to cluster 4 (C2_4). One C2 genome was assigned to cluster 6 and 13 to cluster 10, perhaps due to SNP calling accuracy.

Clade/ subclade	Fastbaps Cluster IDs	Isolate count	*fimH* allele			
			41	22	30	Others
A	2	414	376			38

B	1, 3, 5, 7, 8	420	8	277		135
B0	8	13			13	
C0	-	52			51	1
C1	6	1,121			1,111	10
C2	4, 9	2,051			2,032	19
Total		4,071	384	277	3,207	203

T**able 5.1**. Number of ST131 in clades A, B, B0 and subclades 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 others and were assigned to multiple Fastbaps groups. Clade C mainly had *fimH30* or *fimH*-like alleles and were assigned to Fastbaps cluster 6 for C1 (aka C1_6), or clusters 4 and 9 for C2 (aka C2_4 and C2_9).

5.3.4 ST131 subclades' relative frequencies stable over time

Previous work has shown that 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). This pattern was present in this study in the clades (A, B, C) and three main C subclades, C1_6, C2_4 and C2_9 for 2002-2017 where sufficient annual sampling was available (Supplementary Figure S5.4): clade A was first sampled here in Japan in 2004, and C2_4 in 2008 in the USA, their relative rates soon stabilised after emergence, consistent with NFDS. 53% of the 4,071 had no source information and only 12% of the remainder was from non-human sources, suggesting that although a relatively higher rate of clade B isolates were from animal sources (OR=9.0) here, this may be a consequence of biased sampling.

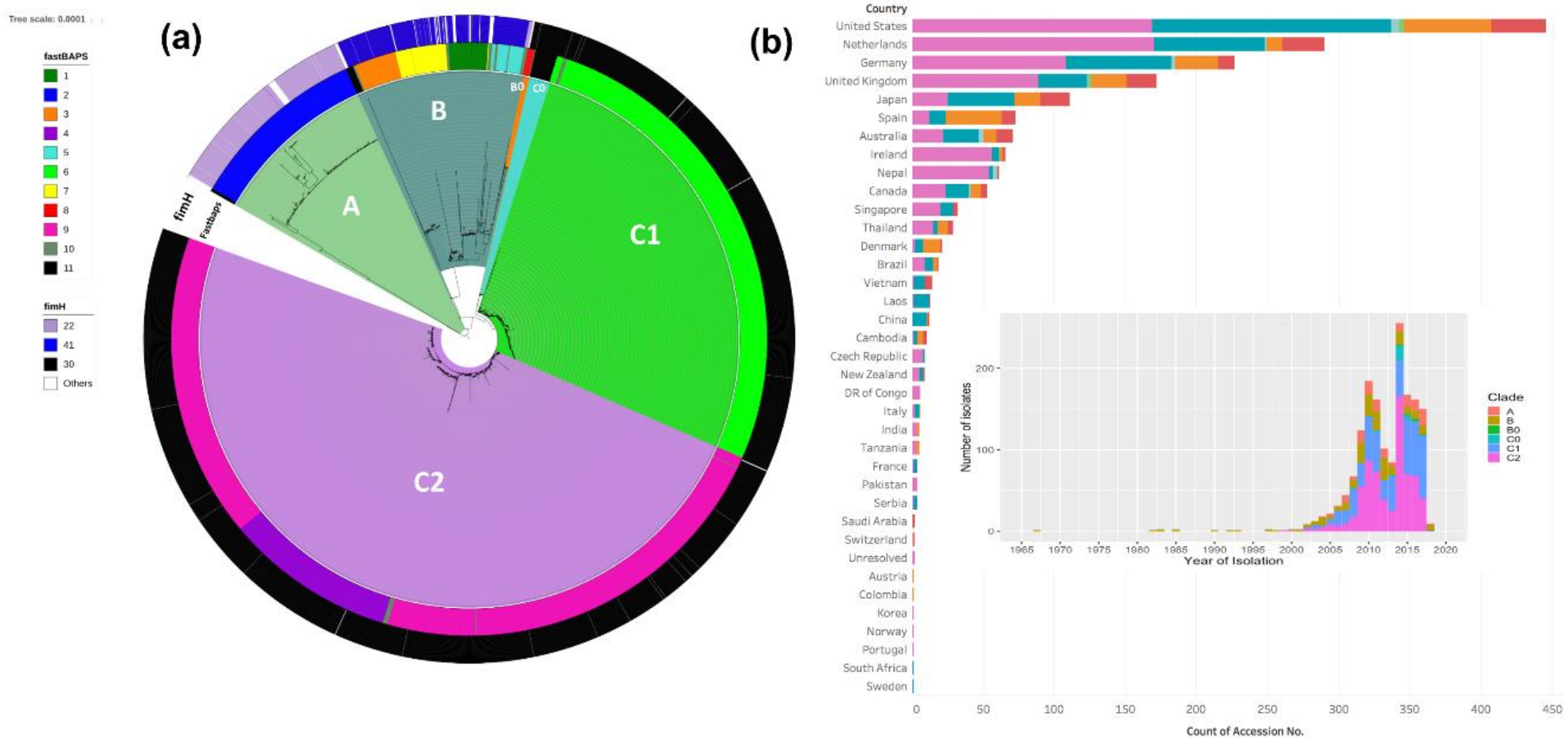


Figure 5.2. Maximum likelihood phylogeny of n=4071 global ST131 (a) and the distribution of *E. coli* ST131 samples across continents and sources over time (b). The phylogram showed clades A (n=414), B (n=420), the intermediate subclade B0 (n=13) and C (n=3,224; n=52 of these were from C0, n=1,121 were from C1 and 2,051 belong to C2). The mid-point rooted phylogram was constructed with RAxML from the 30,029 chromosome-wide SNPs arising by mutation and visualized with iTol. The colored strips surrounding the tree represent the subgroups formed from hierarchical Bayesian clustering generated using Fastbaps and the major type of each strain. The histograms in (b) show that 2,051 out of the 4,071 ST131 genomes isolated from 1999 to 2018 belong to C2.

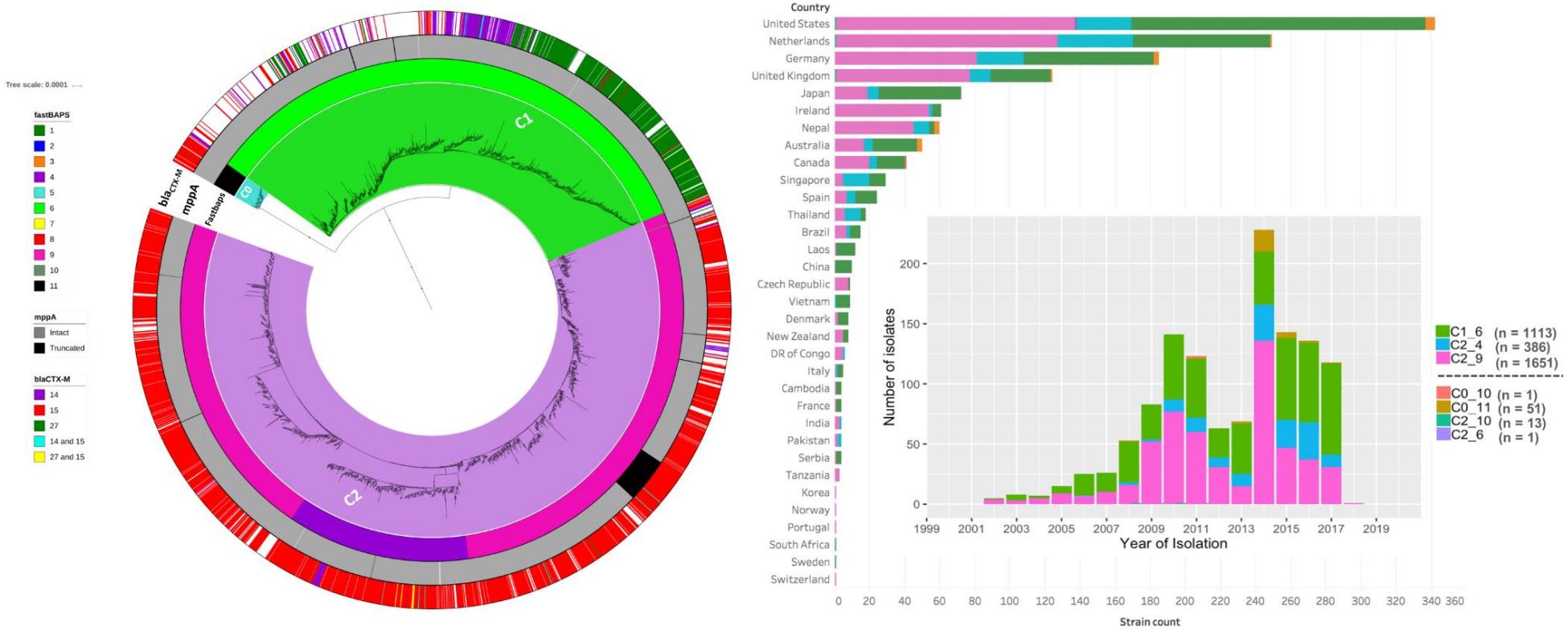


Figure 5.3. Phylogenetic reconstruction (a) and the geographic distribution of n=3,224 Clade C strains over time (b). Clade C0 isolates in (a) were from Fastbaps cluster 11 (n=52) while a singleton belonged to cluster 10. Of the 1,121 C1 samples, 1,113 form the Fastbaps cluster 6 and eight were more closely related to cluster 10. The C2 subclades corresponded to Fastbaps clusters 9 (C2_9, n=1,651 samples) and 4 (C2_4, n=386). 2,416 from Clade C had *bla*$_{\text{CTX-M-14}}$, *bla*$_{\text{CTX-M-15}}$, or *bla*$_{\text{CTX-M-27}}$ genes: 177 *bla*$_{\text{CTX-M-14}}$, 1,790 *bla*$_{\text{CTX-M-15}}$ and 424 *bla*$_{\text{CTX-M-27}}$.

5.3.5 Epidemic ST131 subclades C1 and C2 co-circulating globally

Subclades C1 and C2 were prevalent globally with no evidence of population structure, suggesting they were (and are) a co-circulating epidemic, likely associated with frequent host switching (McNally et al. 2016). This comes from the 819 isolates from Europe, 499 from North America, 294 from Asia, 80 from Oceania, 20 from South America, 12 from Africa - the remaining 2,347 (58%) of the 4,071 had no recorded geographic information. Annual sampling peaked at 255 in 2014: these had a broad geographic origin including countries in Asia, Africa, Europe, North America and Oceania (Figure 5.3; Supplementary Table S5.2). There were minor differences in the relative rates of C1_6, which was more common in North America (OR=1.57, 95% CI 1.25-1.96, p=0.0004) and less common in Europe (OR=0.67, 95% CI 0.53-0.81, p=0.0004), and also for C2_4, which was more frequent in Asia (OR=1.75, 95% CI 1.18-2.56, p=0.019) and less common in North America (OR=0.61, 95% CI 0.40-0.91, p=0.042).

The most likely origin of clade C can be based on the common ancestor with clade B, whose sample distribution in 1967-1997 was solely in the USA across five Fastbaps clusters here (1, 3, 5, 7, 10) until one isolate in Spain in 1998. This is because previous work has timed origins of clade C in 1985, the *fimH*30 allele to 1986, the FQ-R C1/C2 ancestor with mutations at the DNA gyrase and DNA topoisomerase genes to 1991 (Ludden, Decano et al. 2019) (or potentially earlier in 1986, Kallonen et al. 2017), consistent with a North American origin of C. However, the earliest isolate from clade C in our data was isolated in Norway in 1999 from a cancer patient (ERR1912633 from C2_9) that was FQ-R with no $bla_{\text{CTX-M}}$ gene but did have a $bla_{\text{TEM-1B}}$ one (Knudsen et al. 2017). C1_6 was first detected in 2002 in Japan in a $bla_{\text{CTX-M-14}}$-positive sample, and C0 later in 2008 in Nepal.

Subclade	Number	*bla*$_{CTX-M}$ allele numbers per isolate					*mppA*	
		14	**14+15**	**15**	**15+27**	**27**	**Intact**	**Truncated**
A	414	65	1	66	2	51	414	
	%	15.7	0.2	15.9	0.5	6.5	100	
B	420	7	1	9		1	409	11
	%	1.7	0.2	2.1		0.2	97.4	2.6
C0	52			46		1	52	
	%			88.5		2.0	100	
C1_6	1113	149	6	59	3	418	1,108	3
	%	13.4	0.5	5.3	0.3	37.6	99.6	0.3
C2_4	386	12	3	339	7	1	382	1
	%	3.1	0.8	87.8	1.8	0.3	99.0	0.3
C2_9	1651	16	6	1,338		4	1,561	90
	%	1.0	0.4	81.0		0.2	94.5	5.5

Table 5.2. Genetic characterization of ST131 subclades' *bla*$_{CTX-M-14/15/27}$ genes. Also shown is the chromosomal gene *mppA* as intact or truncated, where truncation of this gene may indicate chromosomal insertion of a TU containing a *bla*$_{CTX-M-15}$ gene. Seven samples (0.2% of all) had undetermined *mppA* contexts due to small contig sizes. Subclade B0 (n=13) is not shown because it had no *bla*$_{CTX-M}$ genes and all its *mppA* genes were intact (Supplementary Figure S5.5).

5.3.6 Variable prevalence of *bla*$_{CTX-M-14/15/27}$ genes across time, geography and ST131 subclades

Alignment of the 4,071 assemblies *bla*$_{CTX-M-14/15/27}$ genes with BLAST and CARD showed that these genes were more common in clades A (45%) and C (75%) than B (4%) (Figure 5.1), and that a limited number of isolates were both *bla*$_{CTX-M-14/15}$-positive (0.4%) or *bla*$_{CTX-M-15/27}$-positive (0.3%), but none were *bla*$_{CTX-M-14/27}$-positive (Table 5.2). Clade A had higher rates of *bla*$_{CTX-M-14/15}$ alleles than *bla*$_{CTX-M-27}$ ones, perhaps because the former were detected in isolates from A in 2005-6 versus 2011 for the latter, and *bla*$_{CTX-M-27}$ in clade A was slightly more common in Asia.

Two thousand four hundred eight (2,408) clade C samples had *bla*$_{CTX-M}$ genes: 1,782 *bla*$_{CTX-M-15}$, 424 *bla*$_{CTX-M-27}$, 177 had *bla*$_{CTX-M-14}$, 15 *bla*$_{CTX-M-14/15}$, and 10 *bla*$_{CTX-M-15/27}$ (Figure 5.4). The rate of *bla*$_{CTX-M}$-positive isolates was highest in C2_4 (93.8%) followed by C0 (90%), C2_9 (82.6%) and then C1_6 (57%) (Supplementary Figure S5.7). The earliest *bla*$_{CTX-M}$-positive clade C strain was from a human isolate in Canada (ERR161284 from C2_9, Supplementary Table S5.2, Petty et al. 2014). For this C2_9 group, 81% (1,338 of 1,651) were *bla*$_{CTX-M-15}$-positive with limited geographic or temporal structure (Supplementary Figure S5.6c). Two C2_9 samples that also had *bla*$_{CTX-M-14}$ genes were from Japan, as was one that had *bla*$_{CTX-M-14}$ only and another that had *bla*$_{CTX-M-27}$ only (Matsumura et al. 2016). Similarly, the majority of C2_4 assemblies (339 or 386, 88%) were *bla*$_{CTX-M-15}$-positive, with no geographic or temporal structure indicating their global spread (Supplementary Figure S5.6b). This reiterates that the C2 ancestor was *bla*$_{CTX-M-15}$-positive and that subsequent gains of other *bla*$_{CTX-M}$ genes were likely local conjugative events.

C1_6 had a different *bla*$_{CTX-M}$ gene rates to C2, most probably because its earliest acquisition of *bla*$_{CTX-M-15}$ was 2008 here, whereas *bla*$_{CTX-M-14}$-positive isolates were found from 2002 and *bla*$_{CTX-M-27}$-positive ones from 2004 (Supplementary Figure S5.6a). This later date for *bla*$_{CTX-M-27}$ may be because it differs from *bla*$_{CTX-M-14}$ by a single D240G mutation that arose sometime before 2000, and this allows higher ceftazidime resistance and is shared with *bla*$_{CTX-M-15}$ (Bonnet et al. 2003). *Bla*$_{CTX-M-15}$ was marginally more common in Europe (OR=3.3, 95% CI 1.38-8.70, p n/s), and *bla*$_{CTX-M-14}$ was more common in Asia (OR=4.4, 95% CI 2.21-8.85, p=0.00007), whereas *bla*$_{CTX-M-27}$ was common globally. Consequently, *bla*$_{CTX-M-27}$ (38%) was much more common in C1_6 than *bla*$_{CTX-M-14}$ (14%) or *bla*$_{CTX-M-15}$ (6%) (Table 5.2). In addition, C1_6 was only found in Japan in 2002-2004, before detection in China and Canada (both 2005), Europe by 2007 and Africa by 2008. This historical context is associated with the 11,894 bp M27PP1 region in the C1-M27 lineage (Matsumura et al. 2016), numbering 421 here.

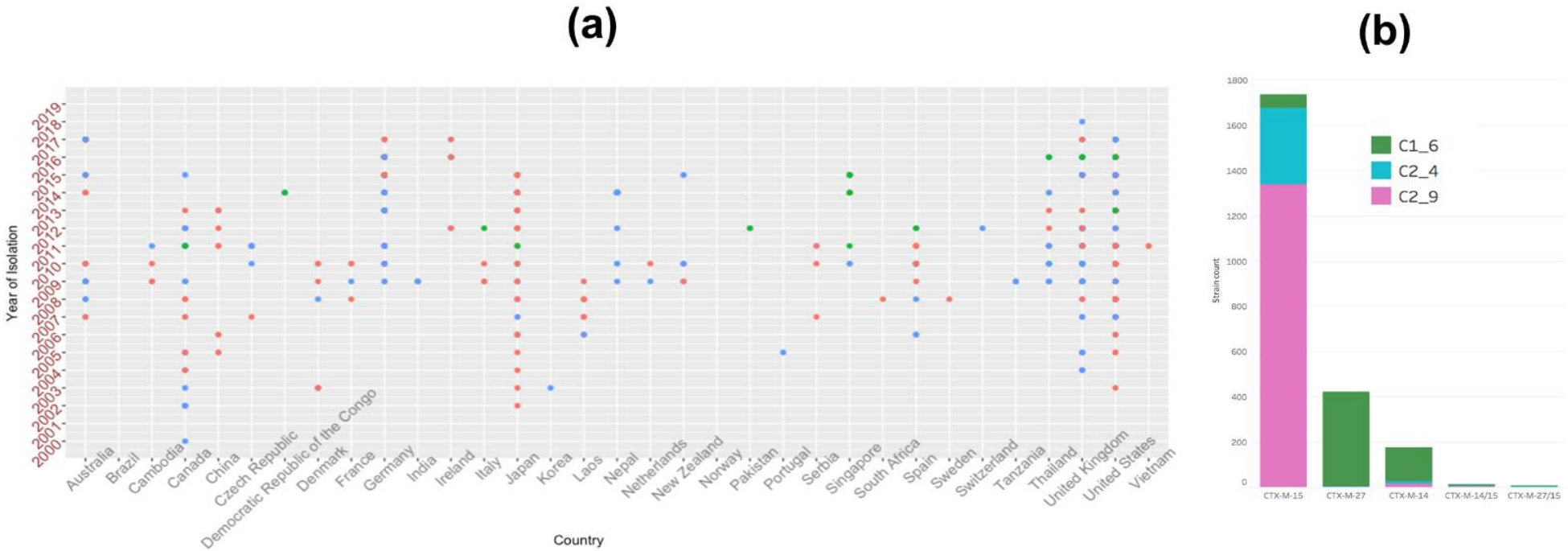


Figure 5.4. Distribution of ST131 samples that belong to subclades C1_6, C2_4 and C2_9 over geography (country) and time (year; a) and the type of blaCTX-M allele that they contain (b). Samples from C1_6 are represented by green dots/squares, those from C2_4 are indicated by blue dots/squres; red dots/pink squares are strains from C2_9. Data were plotted and drawn using R v.1.1.463 for (a) and Tableau v.10.1 for (b).

5.3.7 Genomic locations and structures of the *bla*$_{\text{CTX-M-14/15/27}}$ genes' contigs across ST131 subclades

Screening all 505,761 contigs from the 4,071 assemblies for *bla*$_{\text{CTX-M-14/15/27}}$-positive ones identified different local structures and multiple genomic contexts. These contigs with *bla*$_{\text{CTX-M}}$ genes were annotated with Prokka and MARA so that isoforms in C1_6, C2_4 and C2_9 could be examined. C2 isolates generally had *bla*$_{\text{CTX-M-15}}$ genes in a transposition unit (TU) flanked by a 1,658 bp 5' IS*Ecp1* and 3' a orf477 as a 2,971 bp IS*Ecp1*-*bla*$_{\text{CTX-M-15}}$-orf477Δ TU, usually with a 5.8 Kb Tn*2* at the 3' end (Supplementary Figure S5.8). Some isolates had incomplete TU cassettes due to the small contig lengths recovered from short reads, but this TU structure was verified previously using long reads in Chapter 4 (Decano et al. 2019).

Normally this TU is encoded on an IncF plasmid (F2:A1/ B-), but selected C2_9 isolates had a chromosomal insertion of a *bla*$_{\text{CTX-M-15}}$ gene here (Supplementary Figure S5.9). Previously, an outbreak detected in C2 isolates from Europe and North America had this TU chromosomally inserted at the *mppA* gene (Ludden, Decano et al. 2019). Here, we found 13 additional C2_9 isolates with the same insertion that had both time and geographic data, indicating that this genetic lineage was present in Thailand (Stoesser et al. 2016), Singapore and the Democratic Republic of Congo in 2014 (Irenge et al. 2019), consistent with a global circulation pattern (Supplementary Figure S5.6c, Supplementary Figure S5.5) from an origin in about 1998 (Ludden, Decano et al. 2019). One C2_4 isolate from Pakistan in 2012 had this TU inserted at *mppA* (SRR1610051, Sheppard et al. 2018), affirming that insertions at *mppA* will recur due to local sequence homology to IS*Ecp1*'s 14-bp 3' inverted repeat (IRR) (Lartigue et al. 2006, Canton et al. 2012, Poirel et al. 2013).

5.3.8 Inter-clade but not intra-clade accessory genome divergence

Previous work suggests that the accessory genome was specific to ST complexes because it was associated with ecological niche specialisation driving NFDS such that pangenome variation was higher in C than B than A, even though B was more diverse than A or C (McNally et al. 2016). We compared the ST131 clades and subclades' accessory genomes to determine diversity levels in terms of the core and intermediate (15-95%) frequency

shell genes, pangenome openness and accessory gene overlap. Clade C's more recent origin corresponded to a larger core genome (3,916 genes) that was higher in subclades C0 (4,031), C2_4 (4,109) and C2_9 (4,019) but not C1_6 (3,843), consistent with an older origin for C1 than C2 (Stoesser et al. 2016). Most of the 22,525 accessory genes across the 4,071 assemblies were cloud genes that had a low frequency of <15% across the collection (21,507 or 95.5%) (Supplementary Figure S5.3). Across the three clades and four C subclades, cloud gene rates were proportional to group sample size, symptomatic of an open pangenome increasing sub-linearly without convergence (Supplementary Figure S5.10), for which sample number explained most (r^2=0.846, p=0.00012) of variation in cloud gene number, but not core (r^2=0.162), soft core (r^2=0.258) nor shell (r^2=0.001) gene numbers (Table 3).

Our 4,071 genome assemblies had more open pangenomes compared to related work (McNally et al. 2019) with *alpha* = 0.823 (Supplementary Figure S5.10). As expected, clade A had a less open pangenome (0.807) than B (0.762, Table 5.3), and like previous studies B in this collection had fewer core (3,771 genes) but more shell and cloud genes (Supplementary Figure S5.11). Clade C (0.806) was less open than clade B here, whereas previously 648 clade C ST131 had a more open pangenome than 140 clade B isolates and 70 from clade A (McNally et al. 2019). One factor was that here *alpha* was relatively stable once >250 genomes were sampled (Supplementary Figure S5.10), like previous findings (Park et al. 2019). A second was that the average *alpha* for all 4,071 as increasing numbers were added was similar to the estimate above (0.8123±0.024), though the variance of *alpha* was inversely proportional to sample size (Table 5.4). Within C here, the most prevalent subclade C2_9 had a less open pangenome (0.822) than C2_4 (0.696) or C1_6 (0.755), suggesting that a third factor may be the dependency of pangenome openness with a more recent origin.

NFDS predicts that accessory genes are maintained intermediate levels, presumably due to functional relevance to the ecological niche (McNally et al. 2019). This was supported by the high correlation of pairwise core and accessory genome distances across the 4,071 assemblies between clades A, B and C measured with Poppunk (Lees et al. 2019), consistent with work on a diverse *E. coli* dataset including 218 ST131 (Kallonen et al. 2018). A higher *alpha* corresponding with more population structure was supported by a lower *alpha* for the more recent lineage C2_4 compared to the other clusters. Consequently, we examined pangenome openness using Roary results for clade and subclade combinations and found that C combined with B (0.813) showed less population structure than when combined with A (0.835), consistent with previous work (Table 5.3). Given the pangenome openness for the subclades individually, the *alpha* values for C2_4 with C1_6 (0.760) or C2_9 (0.808), and for C1_6 and C2_9 combined (0.819) showed proportionally similar pangenome differences (Table 5.4).

To explore the accessory gene overlap further, the observed numbers of cloud and shell genes were compared to the expected values weighted by the numbers of pooled samples adjusted for gene category changes (Methods). Given the relatively conserved core genomes adjusted for cloud gene rates, pooled groups with extensive population

structure resulting in accessory genome divergence should have more shell genes, contrasting to h a lack of population structure where there would be no excess of shell genes. Pairwise combinations of A, B and C had a 17-61% cloud gene excess proportional to pooled sample sizes, and shell gene levels indicating higher divergence between A with B (shell gene excess 6%) or C (1%) than for B and C whose deficit of 6% indicated some shell gene overlap.

Within and across the C subclades, the pairwise core and accessory genome distance correlation did not hold because their accessory genomes varied extensively even if the core genomes were nearly identical (Figure 5.5). The older lineages C1_6 and C2_9 had a large cloud gene excess (41%) but a small shell gene one (3%), indicating extensive accessory gene sharing that was absent with C2_4 was compared with both C1_6 (22%) and C2_9 (23%) (Table 5.4). This more unique shell gene set in C2_4 was also found when this was compared with A (41% excess) or B (5%) in contrast to rates for C1_6 (16% with A, -8% with B) and C2_9 (16% with A, -11% with B). These results highlighted that the diverse accessory genes present in any given isolate was independent of the core genome composition, but that the panel of potential accessory genes within each clade was estimatable. They also suggested that newer lineages like C2_4 that are initially rare may initially possess more open pangenome due to NFDS and will gain more mixed accessory genomes like the other ST131 groups if they share environmental niches due to NFDS.

	Clades	All	A	B	C	A & B	A & C	B & C	C0	C1_6	C2_4	C2_9
Gene set	**#isolates**	4,071	414	433	3,200	847	3,616	3,635	52	1,113	384	1,651
Total		26,479	12,163	16,323	21,304	18,639	22,912	25,084	6,427	16,490	10,322	15,485
Core	**c**	3,712	3,798	3,771	3,916	3,708	3,738	3,776	4,031	3,843	4,109	4,019
Soft core	**s**	242	292	281	334	232	260	325	447	424	380	354
Shell	**a**	1,018	764	1,437	731	1,317	980	881	766	571	642	566
Cloud	**d**	21,507	7,309	10,834	16,323	14,573	17,934	20,102	1,183	11,652	5,191	10,546
	weighted[d]	14,726.3			10,135.9	9,111.0	15,281.9	15,660.2				
Shell	**a**	1,018	764	1,437	731	1,317	980	881	766	571	642	566
	weighted[a]	805.1			580.1	1,108.0	734.4	814.7				
Core	**Deficit**	175.4			52.8	76.2	164.4	122.7				
Soft core	**Deficit**	81.5			49.0	54.4	69.2	2.7				
Shell	**Deficit**	-206.9			-150.9	-209.0	-245.1	-65.8				
Shell	**E[a]**	1,062.1			681.9	1,238.6	967.9	940.0				
Shell	**Difference**	-4.1%			7.2%	6.3%	1.2%	-6.3%				
Cloud	**E[d]**	14,776.3			10,086.8	9,032.7	15,270.4	15,719.7				
Cloud	**Difference**	45.6%			61.8%	61.3%	17.4%	27.9%				
Openness	**alpha**	0.8231	0.8066	0.7619	0.8059	0.8132	0.8349	0.8128	0.9392	0.7549	0.6957	0.8222
Openness	**mean_alpha**	0.8123	0.7798	0.7985	0.7658					0.7490	0.7541	0.7989
Openness	**SD_alpha**	0.0244	0.0451	0.0463	0.0310					0.0171	0.0767	0.0378

Table 5.3. The pangenome composition of ST131 clades and subclades showed stable core, soft core and shell genomes with open pangenomes (*alpha*). The ST131 clades and subclades dynamic cloud gene rates near-linearly correlated with sample size following a power law model. For pooled groups with independent genomes, the fractions of core, soft core and shell genes should decrease proportionally and that for cloud genes should increase. If a pool group had a near-identical accessory genome, the fractions of core, soft core and shell genes would be constant with a cloud gene set that was dependent on the pooled sample size. The excess percentage was expressed as a function of the expected numbers. The difference between the observed and expected shell and cloud gene set counts showed the percentage excess genes where the expected value was the weighted average of the pooled groups. The B0 subclade (n=13) was included with clade B above. Eight C1 (C1_10) and 16 C2 (C2_6 and C2_10) isolates were not examined in this analysis because they were not assigned to clear clusters during Fastbaps analysis.

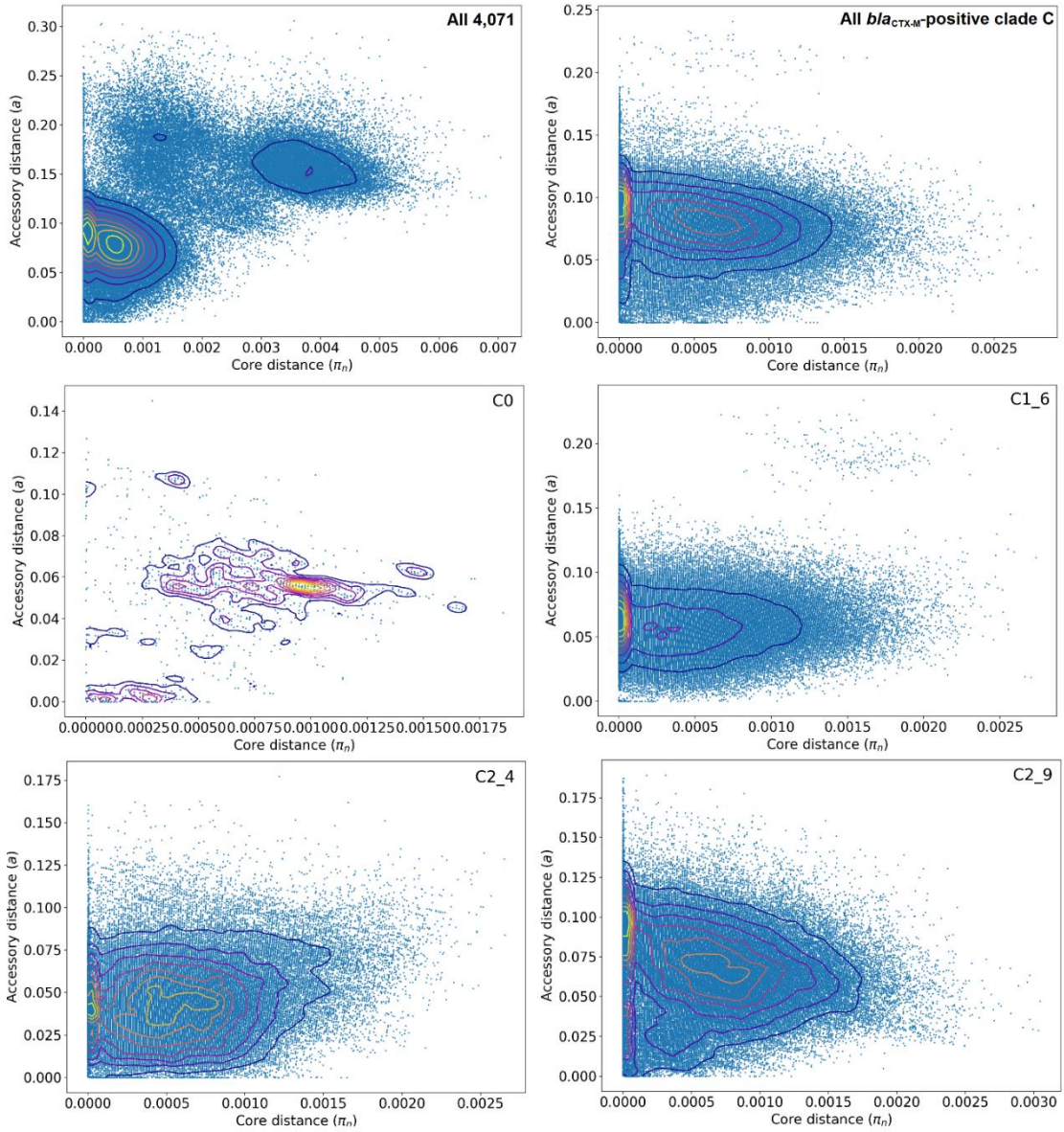


Figure 5.5. Majority of Clade C isolates have more diversity across their accessory genome compared to their core genome. Each plot shows the distribution of core (π, x-axis) and accessory genome pairwise distances (a, y-axis) as blue dots with contour indicating dot density. Top left: All 4,071 assemblies 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: All 2,416 *bla*$_{CTX-M}$-positive clade C strains were grouped with the other clade C isolates. Middle left: The 52 subclade C0 isolates had a peak density at π=0.001, a=0.055. Middle right: The 1,113 C1_6 isolates had a peak density mainly at π≤0.001, a=0.06. Bottom left: The 386 C2_4 isolates had peak densities at π=0.0006, a=0.045 and π=0.0001, a=0.04. Bottom right: The 1,651 C2_9 isolates had peak densities at π=0.0007, a=0.065 and π≤0.0001, a=0.09.

	Subclades	C2_4 & C2_9	C1_6 & C2_4	C1_6 & C2_9	A, B & C2_9	A, B & C2_4	A, B & C1_6
Gene set	**#isolates**	2,035	1,497	2,764	2,498	1,233	1,960
Total		16,637	18,315	20,346	22,200	19,933	23,368
Core	**c**	4,023	3,875	3,914	3,754	3,740	3,681
Soft core	**s**	358	406	348	196	215	242
Shell	**a**	733	772	658	1,098	1,405	1,113
Cloud	**d**	11,523	13,262	15,426	17,152	14,573	18,332
	weighted[d]	9,535.5	9,994.7	10,991.4	11,911.4	11,627.5	12,914.3
Shell	**a**	733	772	658	1,098	1,405	1,113
	weighted[a]	580.3	589.2	568.0	820.6	1,104.6	893.4
Core	**Deficit**	13.0	36.2	34.1	159.5	92.9	103.7
Soft core	**Deficit**	0.9	6.7	34.2	116.6	63.1	99.0
Shell	**Deficit**	-152.7	-182.8	-90.0	-277.4	-298.1	-219.6
Shell	**E[a]**	594.2	632.2	636.3	1,096.8	1,260.6	1,096.1
Shell	**Difference**	23.4%	22.1%	3.4%	0.1%	11.5%	1.5%
Cloud	**E[d]**	9,396.8	9,854.8	10,969.7	11,910.3	11,485.4	12,897.4
Cloud	**Difference**	22.6%	34.6%	40.6%	44.0%	26.9%	42.1%
Openness		0.8080	0.7598	0.8188	0.8476	0.7918	0.8188

	Subclades	A & C2_9	A & C2_4	A & C1_6	B & C2_9	B & C2_4	B & C1_6
Gene set	**#isolates**	2,065	800	1,527	2,084	819	1,546
Total		17,755	14,578	19,011	20,683	17,972	21,843
Core	**c**	3,778	3,789	3,680	3,823	3,812	3,711
Soft core	**s**	267	266	294	274	231	336
Shell	**a**	1,015	1,308	1,007	855	1,332	894
Cloud	**d**	12,695	14,573	14,030	15,731	12,597	16,902
	weighted[d]	9,897.0	6,274.1	10,474.5	10,605.8	8,161.7	11,422.9
Shell	**a**	1,015	1,308	1,007	855	1,332	894
	weighted[a]	605.7	703.5	623.3	747.0	1,060.7	813.5
Core	**Deficit**	196.7	158.3	150.8	144.5	117.6	111.8
Soft core	**Deficit**	74.6	68.2	94.2	64.8	96.3	47.9
Shell	**Deficit**	-409.3	-601.2	-383.7	-108.0	-268.0	-80.5
Shell	**E[a]**	877.0	930.0	868.3	956.3	1,274.6	973.3
Shell	**Difference**	15.7%	40.7%	16.0%	-10.6%	4.5%	-8.2%
Cloud	**E[d]**	9,759.0	5,899.3	10,335.9	10,707.1	8,107.6	11,502.2
Cloud	**Difference**	30.1%	147.0%	35.7%	46.9%	55.4%	46.9%
Openness		0.8351	0.7573	0.7880	0.8246	0.7842	0.8195

Table 5.4. The pangenome composition of ST131 clades and subclades comparisons showed stable core, soft core and shell genomes with open pangenomes (*alpha*). The excess percentage was expressed as a function of the expected numbers. The difference between the observed and expected shell and cloud gene set counts showed the percentage excess genes where the expected value was the weighted average of the pooled groups. The B0 subclade (n=13) was included with clade B above. Eight C1 (C1_10) and 16 C2 (C2_6 and C2_10) isolates were not examined in this analysis because they were not assigned to clear clusters by Fastbaps.

5.4 Discussion

In Chapter 4 (published in *mSphere*), we showed the diverse AMR contexts of *bla*$_{\text{CTX-M}}$ alleles at plasmids and chromosomes in ST131. We sequenced six clade C ST131 strains using Oxford Nanopore technology. While my results provided a high resolution of the *bla*$_{\text{CTX-M}}$ genomic locations and local structures, the sample size was small. Analysing the largest ST131 sample collection so far thus strengthens this previous chapter.

By collating all available ST131 genomes to produce 4,071 high quality draft assemblies, we reconstructed their phylogenetic relationships using a core genome of 3,712 genes to show that ST131 is dominated by subclades C1 and C2. Although they have different origins and thus had different ancestral ESBL gene compositions, these subclades have been co-circulating globally since at least 2002 with relatively stable globally frequencies, with minor differences in rates presumably due to differing evolutionary patterns after emerging in North America (Stoesser et al. 2016). This worldwide circulation coupled with NFDS (McNally et al. 2019) suggests new genetic types with different host adhesion abilities (as seen by *fimH30,* Paul et al. 2013) or AMR variants (like FQ-R or *bla*$_{\text{CTX-M-15}}$) will become an additional co-circulating lineage, and we found such a pattern in our study for the minority C2 subgroup called C2_4.

Our analysis confirmed that clade A had variable *bla*$_{\text{CTX-M}}$ gene isoforms where present (39%) but clade B seldom was *bla*$_{\text{CTX-M}}$-positive (4%) in spite of its high diversity, suggestive of potential differences in niche specialisation but no difference in source types were evident across clades in this study. Subclades C0 and C2 were typically (81-88%) *bla*$_{\text{CTX-M-15}}$-positive in contrast to subclade C1 had either *bla*$_{\text{CTX-M-14}}$ (14%) or *bla*$_{\text{CTX-M-27}}$ (38%) genes where present. Although the C1 ancestor may have been *bla*$_{\text{CTX-M-14}}$-positive, the higher ceftazidime resistance of *bla*$_{\text{CTX-M-27}}$ (Bonnet et al. 2003) may explain its higher incidence in C1 that we can expect to continue into the future.

We previously highlighted our observations of chromosomally inserted *bla*$_{\text{CTX-M-15}}$ mostly unique to a clade from an Irish outbreak (Chapter 3). We noted a similar pattern in our bigger sample population: the same Irish samples formed a subgroup with this unique chromosomal insertion at *mppA*, with some exceptions dispersed across different clades.

The clonal expansion of the C2_9 isolates with the 2,971 bp IS*Ecp1*-*bla*$_{\text{CTX-M-15}}$-orf477-*Tn2* TU *mppA* chromosomal insertion has spread worldwide. Prior to 1991, ESBL-positive isolates were *bla*$_{\text{SHV}}$-positive (Piddock et al. 1997), and subsequently this TU was obtained from *Kluyvera* (Humeniuk et al. 2002, Barlow et al. 2008), was detected in *Enterobacteriaceae* in 1999 in India (Karim et al. 2001), and the first plasmid sequenced containing it was isolated in 1999-2000 in Canada from extended-spectrum cephalosporin-resistant ST131 (Boyd et al. 2004). This reiterates that tracking plasmid, MGEs and ESBL genes must be a key component of disease monitoring to consider potential future ST131 outbreaks' spectrum of AMR.

Horizontal DNA transfer has allowed *E. coli* to adapt to new ecological niches (Chen et al. 2006) and contributes to its dynamic accessory genome (Welch et al. 2002). Our low ratio of core (3,712) to accessory (22,525) genes was consistent with previous work (Chaudhuri et al. 2010) and showed that cloud gene number is a function of isolate number with a median of 2.1 genes per additional isolate in this large collection of 4,071. Similarly, a more diverse set of 1,509 *E. coli* including 266 ST131 had a core genome of 1,744 genes (>99% of isolates) and 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). Likewise, a previous study of 283 predominantly ST131 ExPEC samples had a total of 16,236 genes in an open pangenome, with a core of 3,079 genes (Salipante et al. 2015), 21% less than the core gene complement here. These pangenome compositions are comparable to other *Proteobacterial* collections (Livingstone et al. 2018).

Although the ST131 accessory genome had 22,525 genes, the NFDS hypothesis posits that the 1,018-shell gene set of intermediate (15-95%) frequency genes may drive adaptation to new hosts and environments (McNally et al. 2019). The open pangenome (Medini et al. 2005) and level of shell gene sharing across clades supported inter-clade structure resulting from ecological specialisation, with clade A more different to B and C. Within C, we found C2_9 had a smaller set of shell genes, perhaps symptomatic of stronger host-environment NFDS. The more recently emerged C2_4 lineage had a shell gene set more distinct from C1_6 and C2_9, indicating that core and accessory gene relationships may differ within subclades. We illustrated this by showing that the inter-clade correlation of accessory genome distance with that of the core across isolate pairs was not found within

subclades. Moreover, isolates from C1 and C2 within minimal core genome differences often had divergent accessory genomes (Achtman 2012), implying that approaches using core genes may miss the constellation of changes at plasmids, AMR genes and MGEs (Lanza et al. 2014) and thus that outbreaks can be better understood using pangenomic epidemiology (Stoesser et al. 2016).

This study's atlas of ST131 genomes will aid in tracing of historical and current transmission in humans, for which resolving inferred evolutionary transmissions to the level of genealogies may be limited by sampling depth (Hanage 2019). Clonal outbreaks may reflect transmission patterns shared by a common ancestral sample or gene and can be enhanced by adding epidemiological information to prospective and routine genome-sequencing data (Raven et al. 2019) to allow inference of past, present and emerging fitness in different lineages (Azarian et al. 2018).

Overall, Chapter 5's results confirm ST131's rapid global dissemination driven by ESBL genes. Future work could explore ideas in a recent study showing that case notification data can produce accurate estimates of a pathogen's reproductive number over time, comparable to actually analysing the genomes throughout the duration of an outbreak (Duchêne et al. 2019). Analyses of the metadata alone enabled the authors to simulate birth-death process and infer epidemiological dynamics (Duchêne et al. 2019). Given numerous large-scale pathogen sequencing projects (Harrison et al. 2019) like GenomeTrackr, which has over 317,000 sequenced isolates as of 31st March 2019 and is adding >9,500 more per month, more rapid analytic methods are essential to tackle infection outbreaks more effectively by harnessing global epidemiological information. These now entail global coordination of data processing and bioinformatic interpretation to maximise the output from large datasets to identify, trace and control disease outbreaks (Pijnacker et al. 2019).

5.5 Data summary

All raw sequence data (reads and/or assembled genomes) for the *E. coli* genomes analysed in this publication are publicly available under the project numbers given in Supplementary Table S5.2, with more detail on their Bioproject IDs and associated study DOIs in Supplementary Table S5.3.

An interactive version of the phylogeny generated by Poppunk using k-mer genetic distances for all 4,071 ST131 assemblies is available in MicroReact at https://microreact.org/project/oD6K_fL2d. The tree file (Newick format) is also avaiable for download via this link.

5.6 Supplementary Tables and Figures

Supplementary tables for Chapter 5 are publicly available in Figshare: https://figshare.com/s/d7f57048f104aa45ae79.

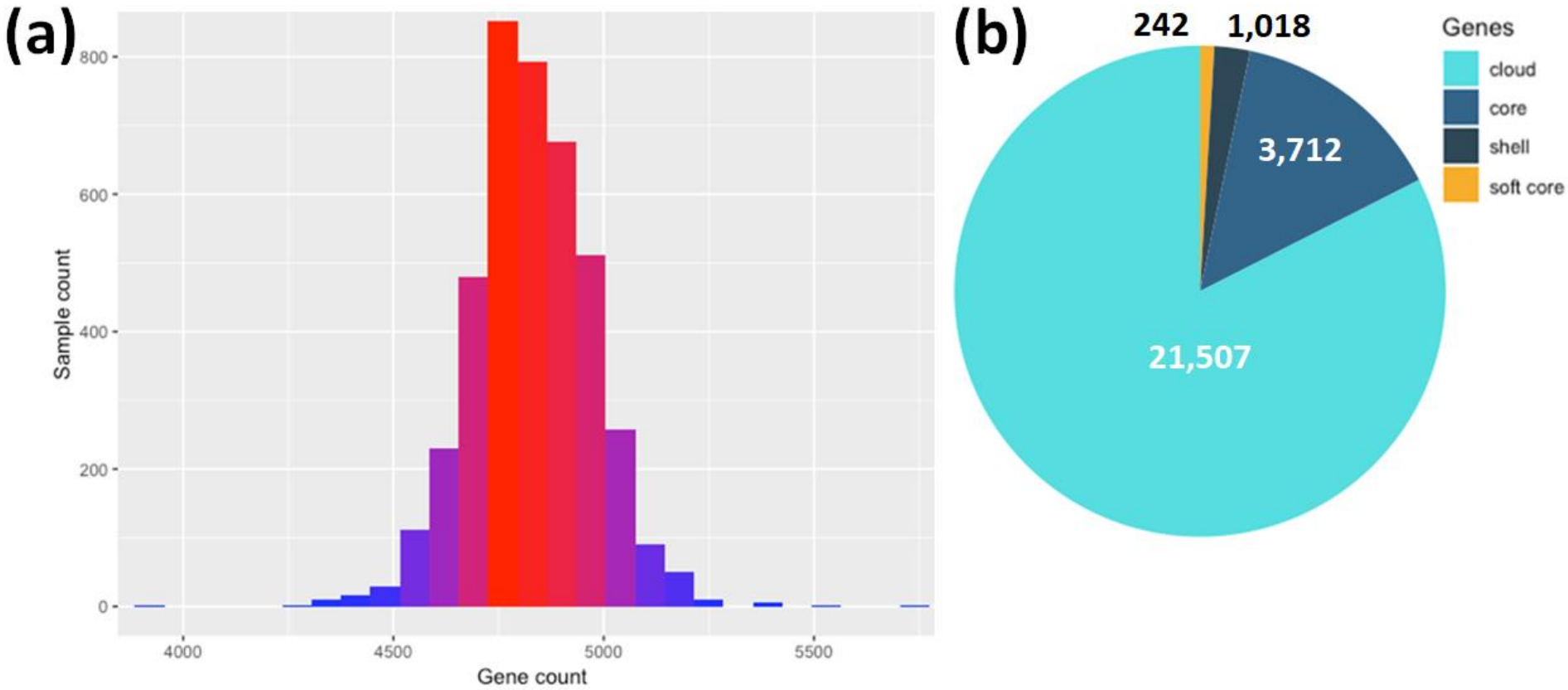
Supplementary Table S5.1. Metadata of 4,071 high quality ST131 genomes used in this study downloaded from Enterobase.

Supplementary Table S5.2. Quality statistics of the final ST131 read libraries used in this study. Shown are the proportion of duplicate reads, the average GC content, mean sequence length (bp) and total number of sequences (millions) per library.

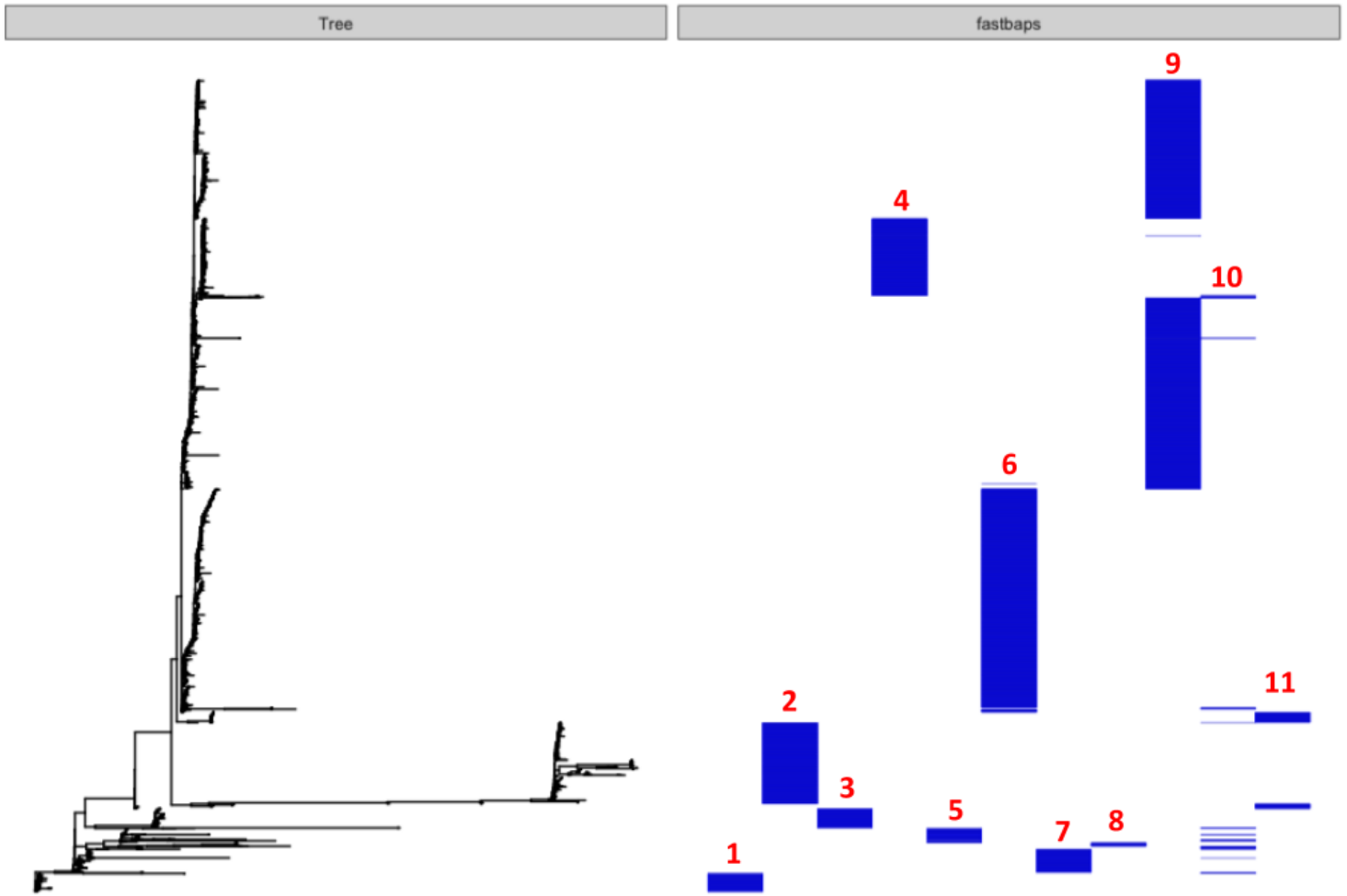
Supplementary Table S5.3. Bioproject IDs and associated study DOIs of published studies where the genomes in this study were sourced from.

Acknowledgements

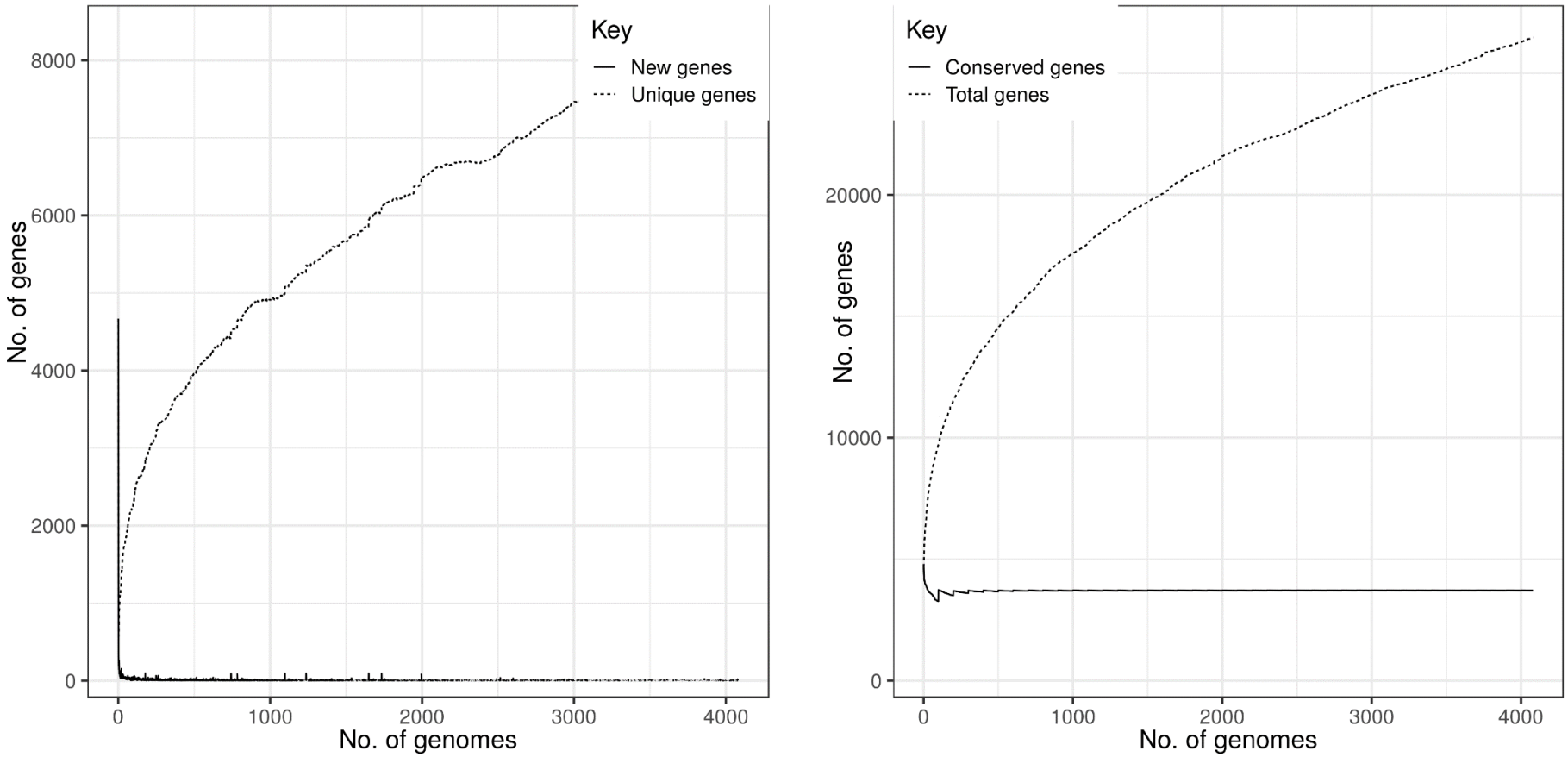
I thank Marius Kinderis (Dublin City University, Ireland) for assistance in implementing the text mining algorithm. This project was funded by a DCU O'Hare Ph.D. fellowship and a DCU Enhancing Performance grant.



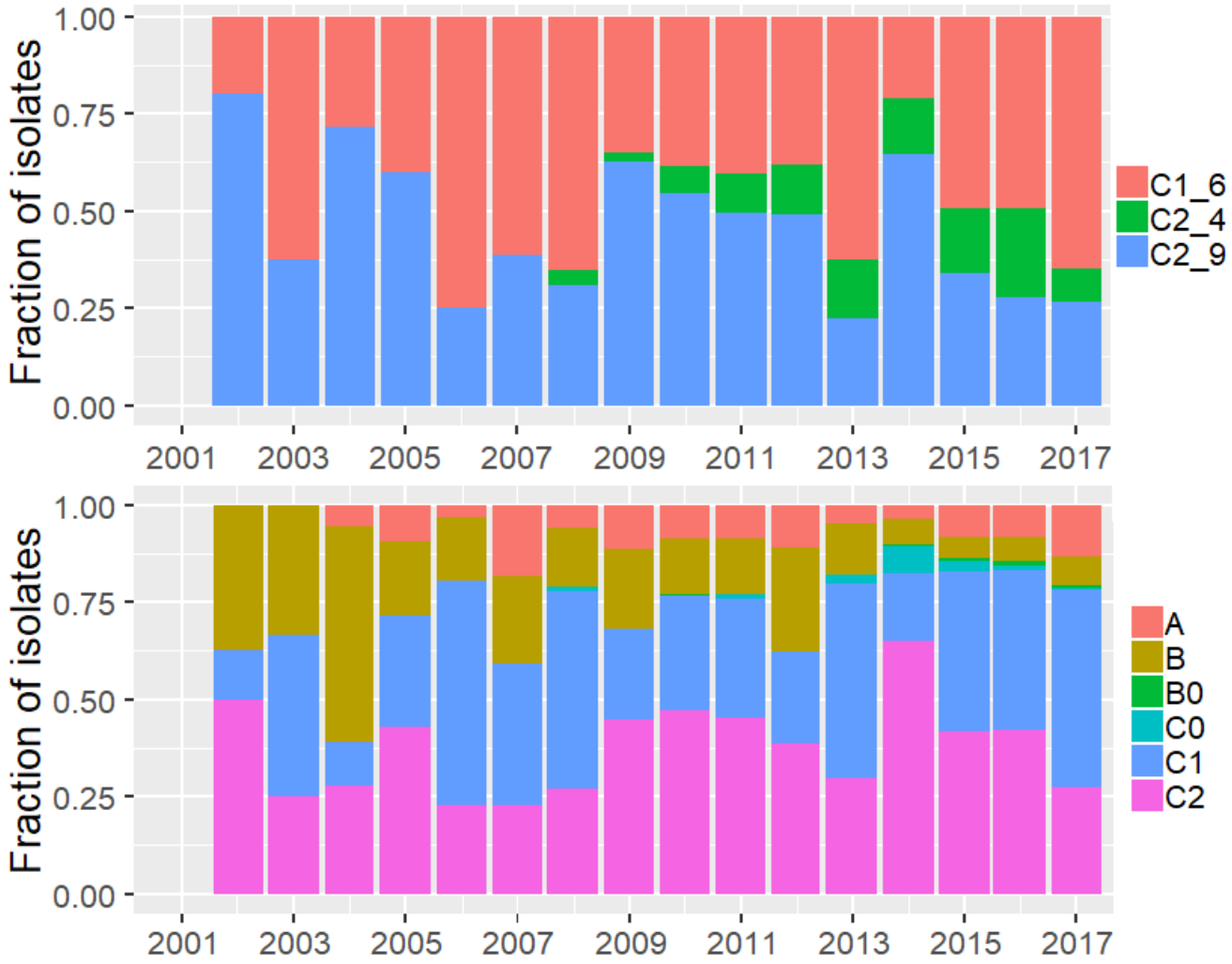
Supplementary Figure S5.1. Number of genes in the 4,071 ST131 genomes (along with NCTC13441) annotated using Prokka showed that (a) this identified 4,829 genes on average per assembly with a minimum of 3,942 and maximum of 5,749. (b) Of the total 26,479 gene clusters detected using Roary, 3,712 comprised 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).



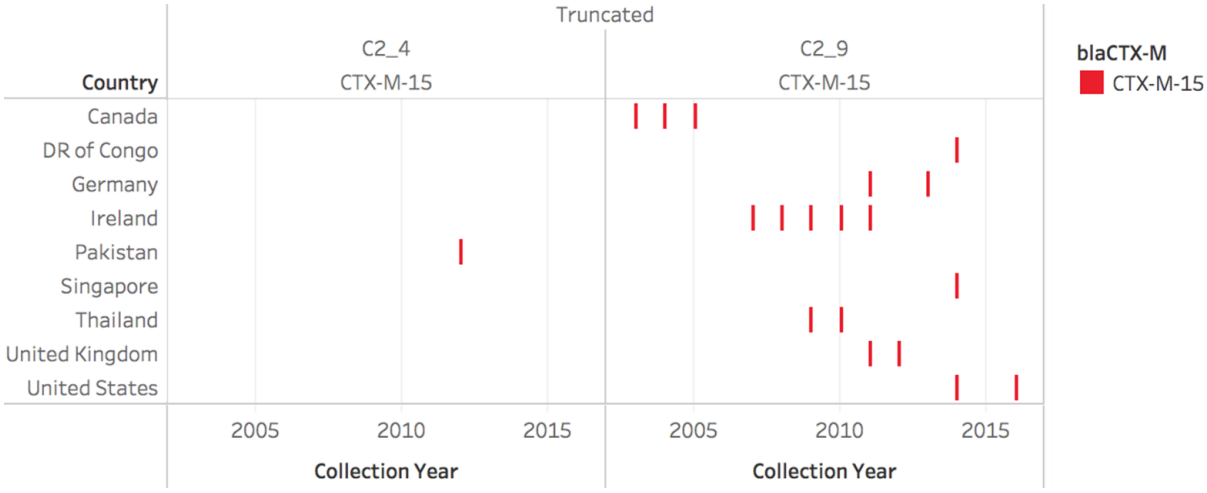
Supplementary Figure S5.2. Hierarchical sub-clustering of 4,071 strains using Fastbaps based on 30,029 SNPs. Grouping is indicated by numerical numbers in bold red font on top of the blue bars. There were nine major clusters found while two (clusters 10 and 11) were dispersed among the major clades



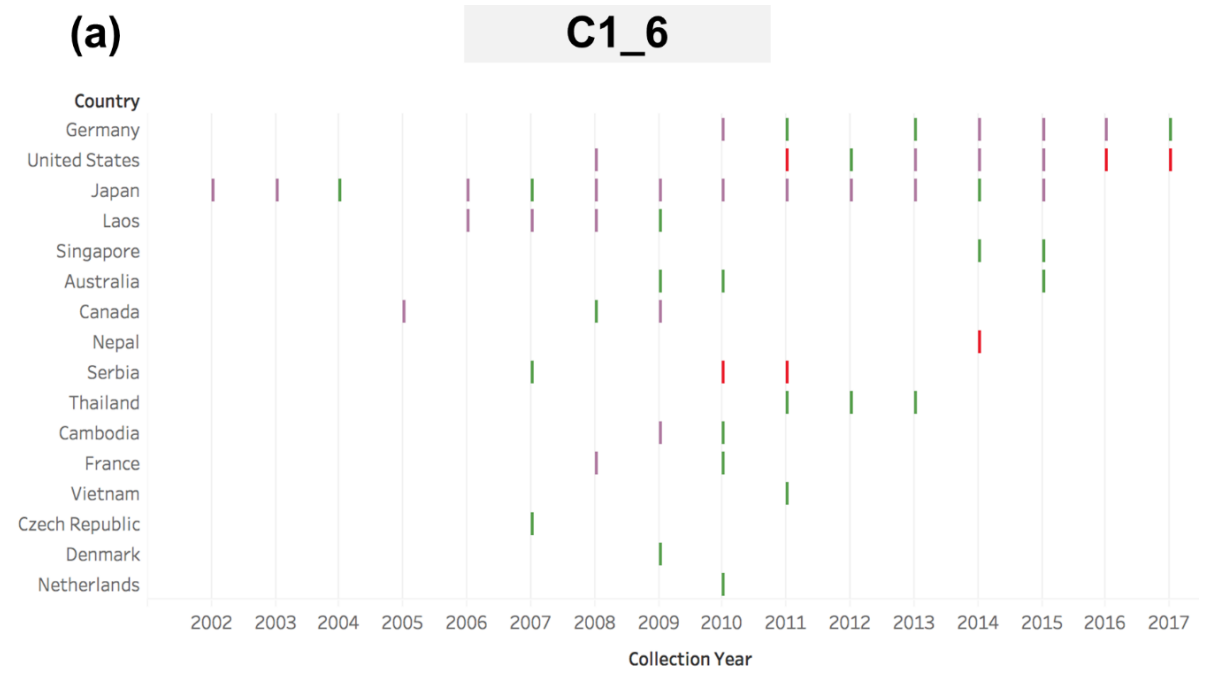
Supplementary Figure S5.3. Pangenome construction showed (left) that few new genes discovered (black bars) as the number of genomes included increased (x-axis), but that the number of unique genes associated with the cloud gene set increased consistently (dashed line). (Right) The core genome composition across all 4,071 assemblies was stable once >200 genomes were included (solid line), whereas the total number of cloud genes increased at a much higher rate (dashed line).



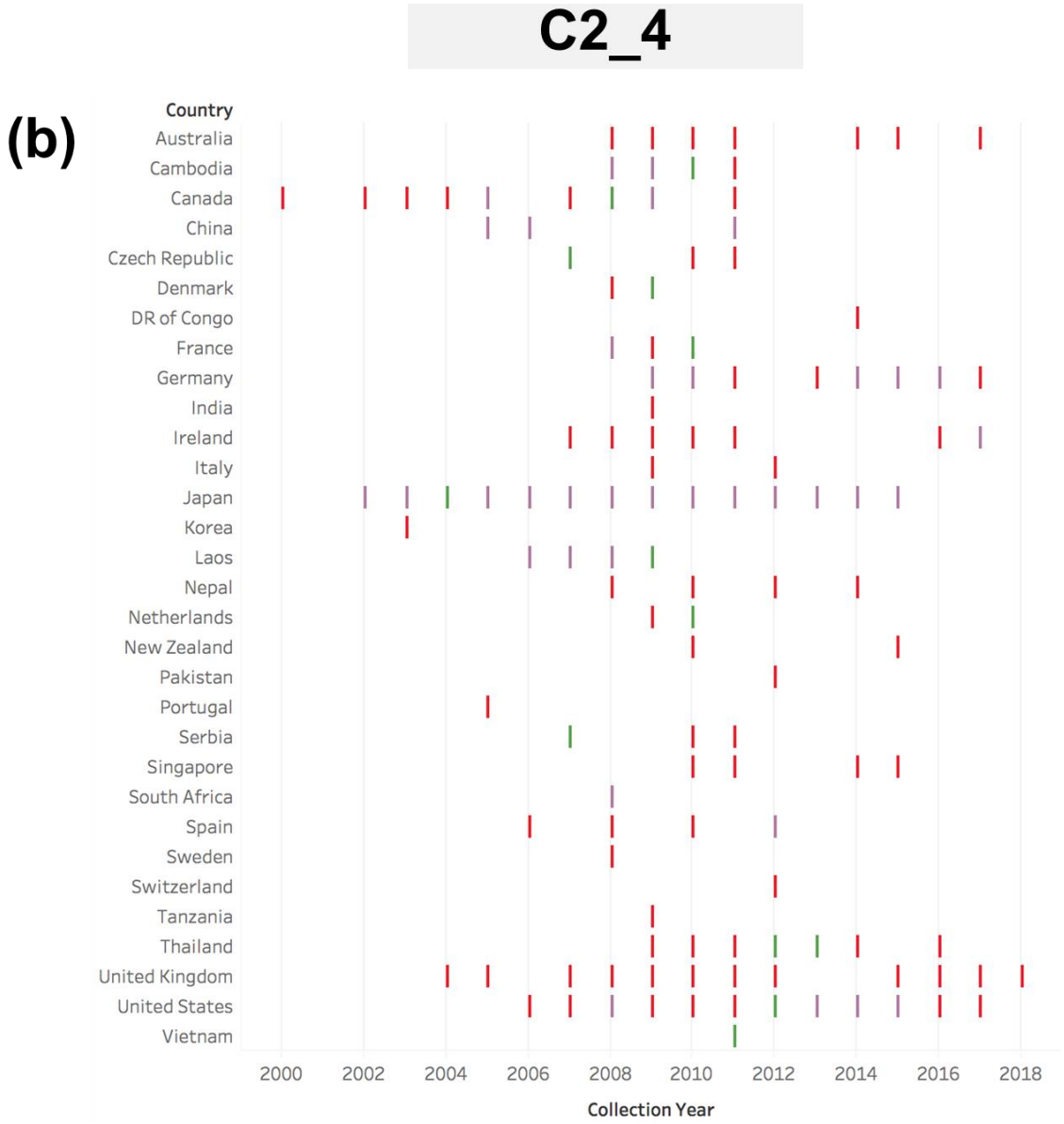
Supplementary Figure S5.4. Frequencies of ST131 subclades C1_6, C2_4 and C2_9 (top) and clades (bottom) over time showed relatively equal rates per year.

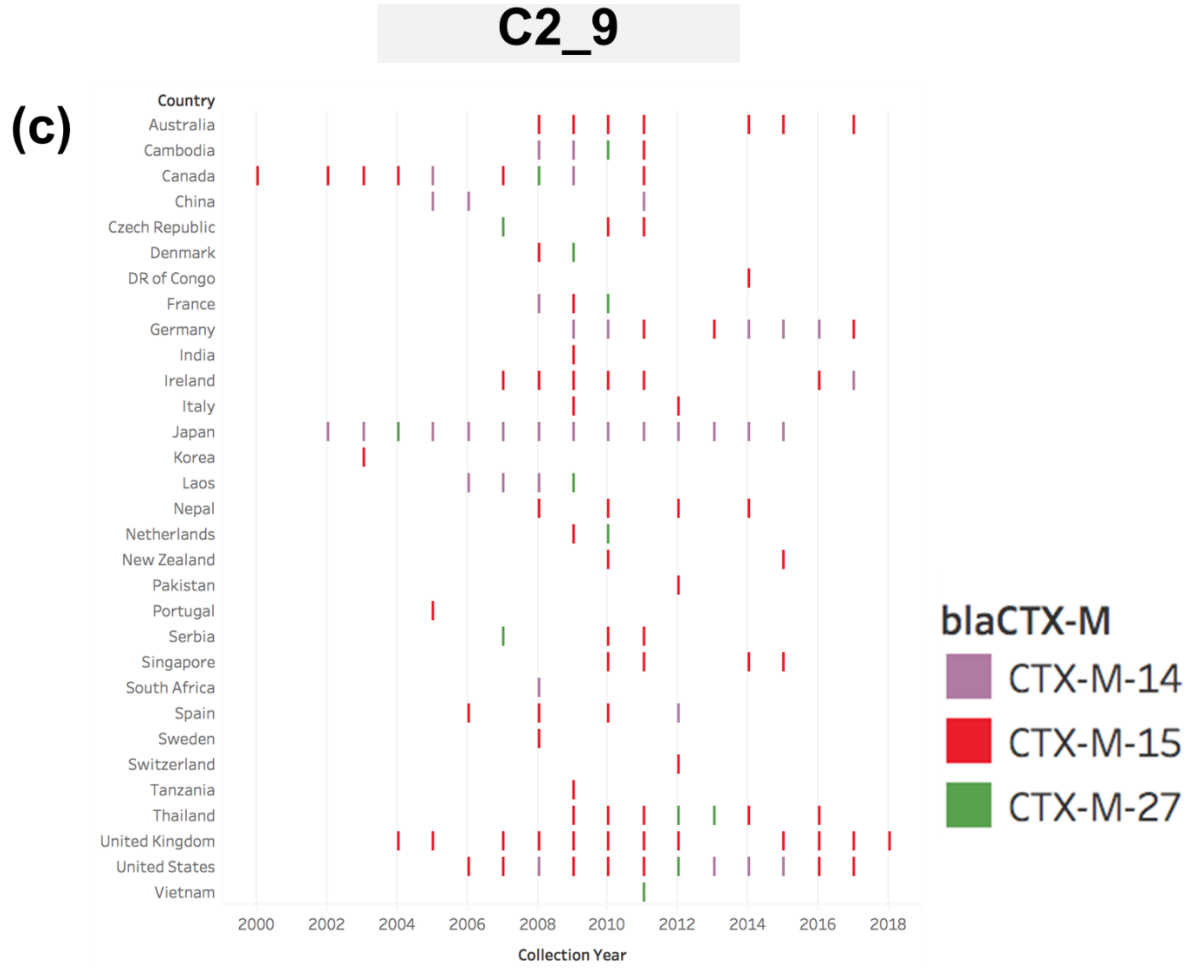


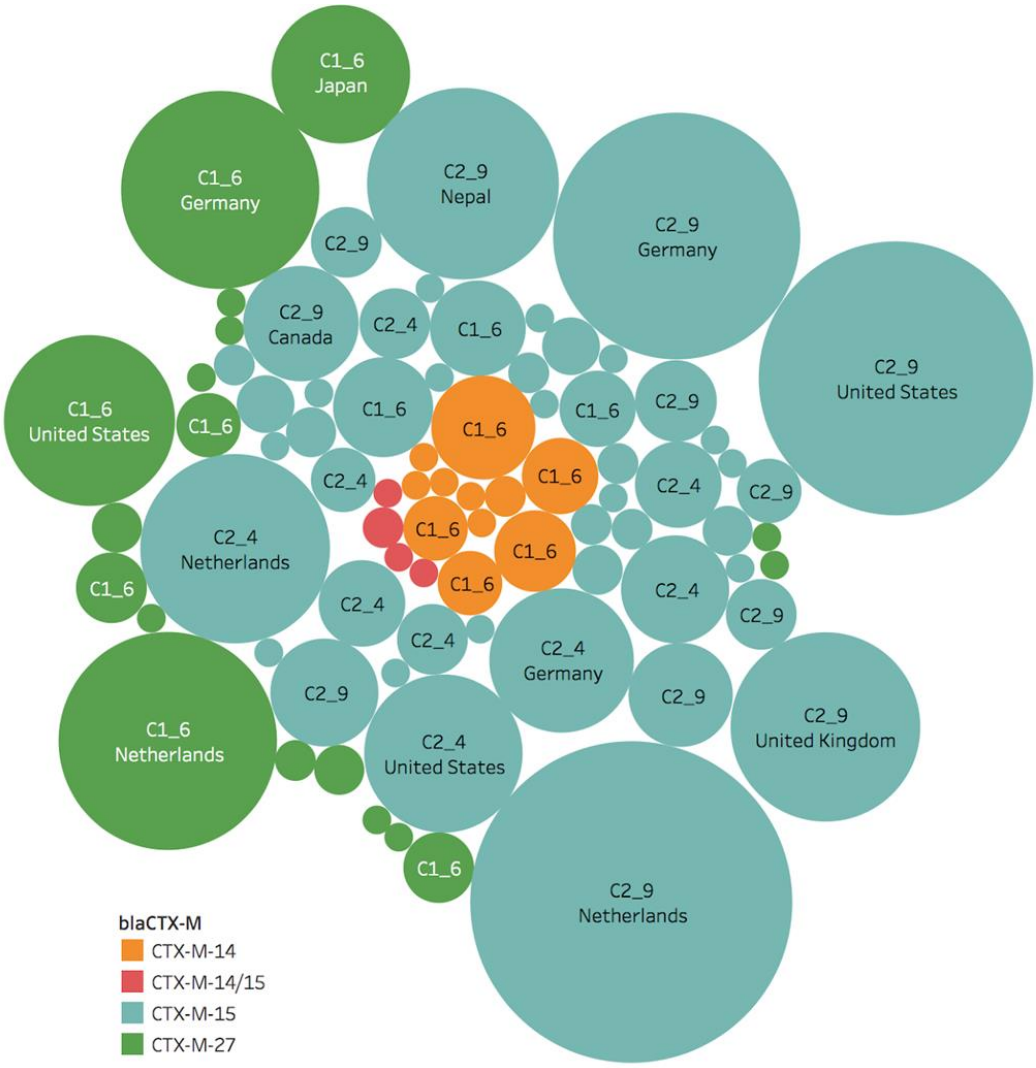
Supplementary Figure S5.5. Distribution of the ESBL *bla*$_{CTX\text{-}M\text{-}15}$ allele among C2_4 and C2_9 strains and their geographic (country) and temporal (isolation year) origins**.** Data were plotted and drawn using Tableau v.10.1.



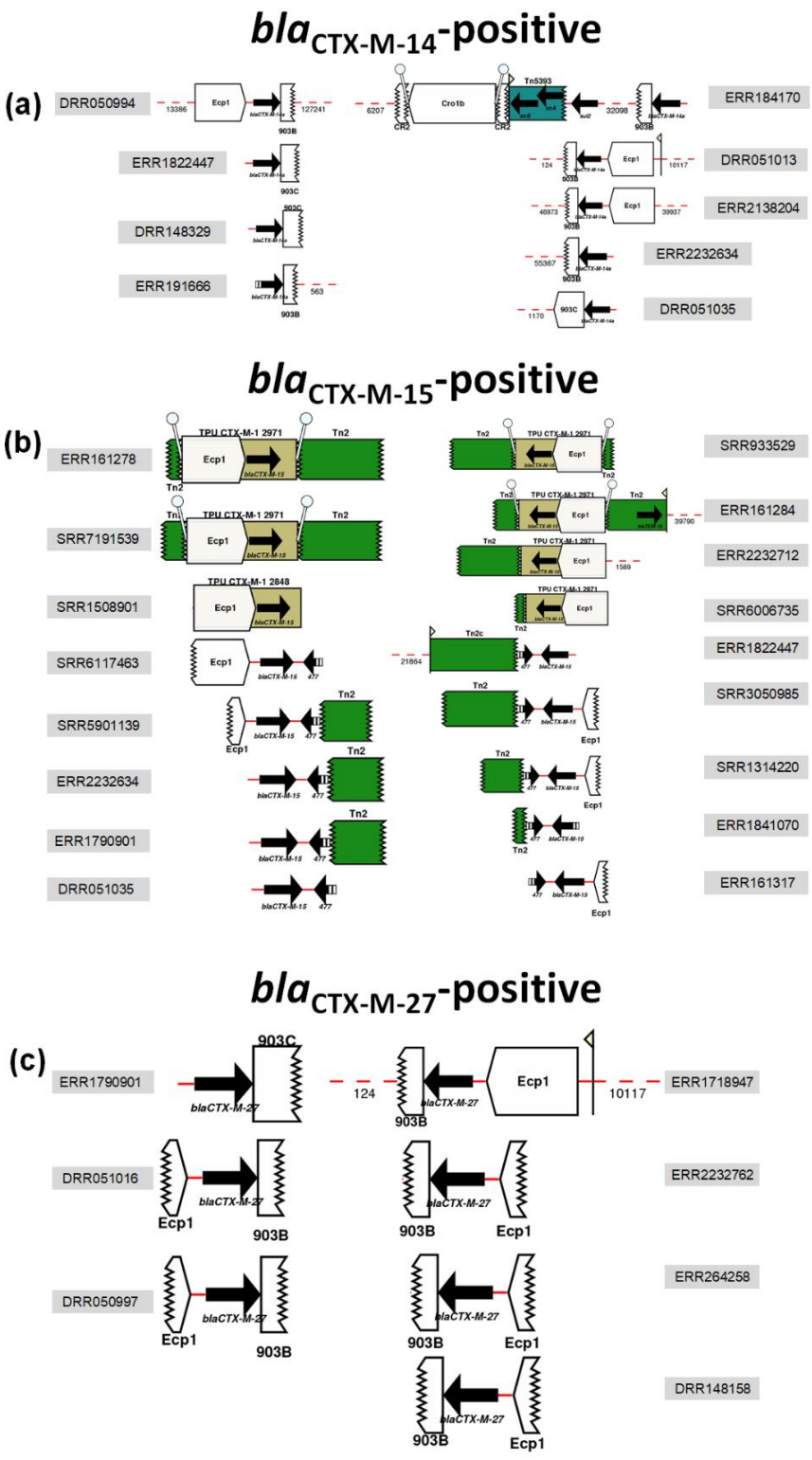
Supplementary Figure S5.6. Distribution of *bla*$_{CTX\text{-}M\text{-}14,}$ *bla*$_{CTX\text{-}M\text{-}15}$ and *bla*$_{CTX\text{-}M\text{-}27}$ among ST131 samples from C1_6 (a), C2_4 (b) and C2_9 (c) and their geographic location (country) over time (isolation year). *Bla*$_{CTX\text{-}M\text{-}14}$was 1st acquired by strain in Japan in 2004 co-occurring with an isolate that was noted with blaCTX-M-27 in the same year. Initial records of *bla*$_{CTX\text{-}M\text{-}15}$ were in C2 strains (both C2_4 and C2_9) in Canada in 2000. Data were plotted and drawn using Tableau v.10.1.


C2_4
(b)
Country
Australia
Cambodia
Canada
China
Czech Republic
Denmark
DR of Congo
France
Germany
India
Ireland
Italy
Japan
Korea
Laos
Nepal
Netherlands
New Zealand
Pakistan
Portugal
Serbia
Singapore
South Africa
Spain
Sweden
Switzerland
Tanzania
Thailand
United Kingdom
United States
Vietnam
2000
2002
2004
2006
2008
2010
2012
2014
2016
2018
Collection Year

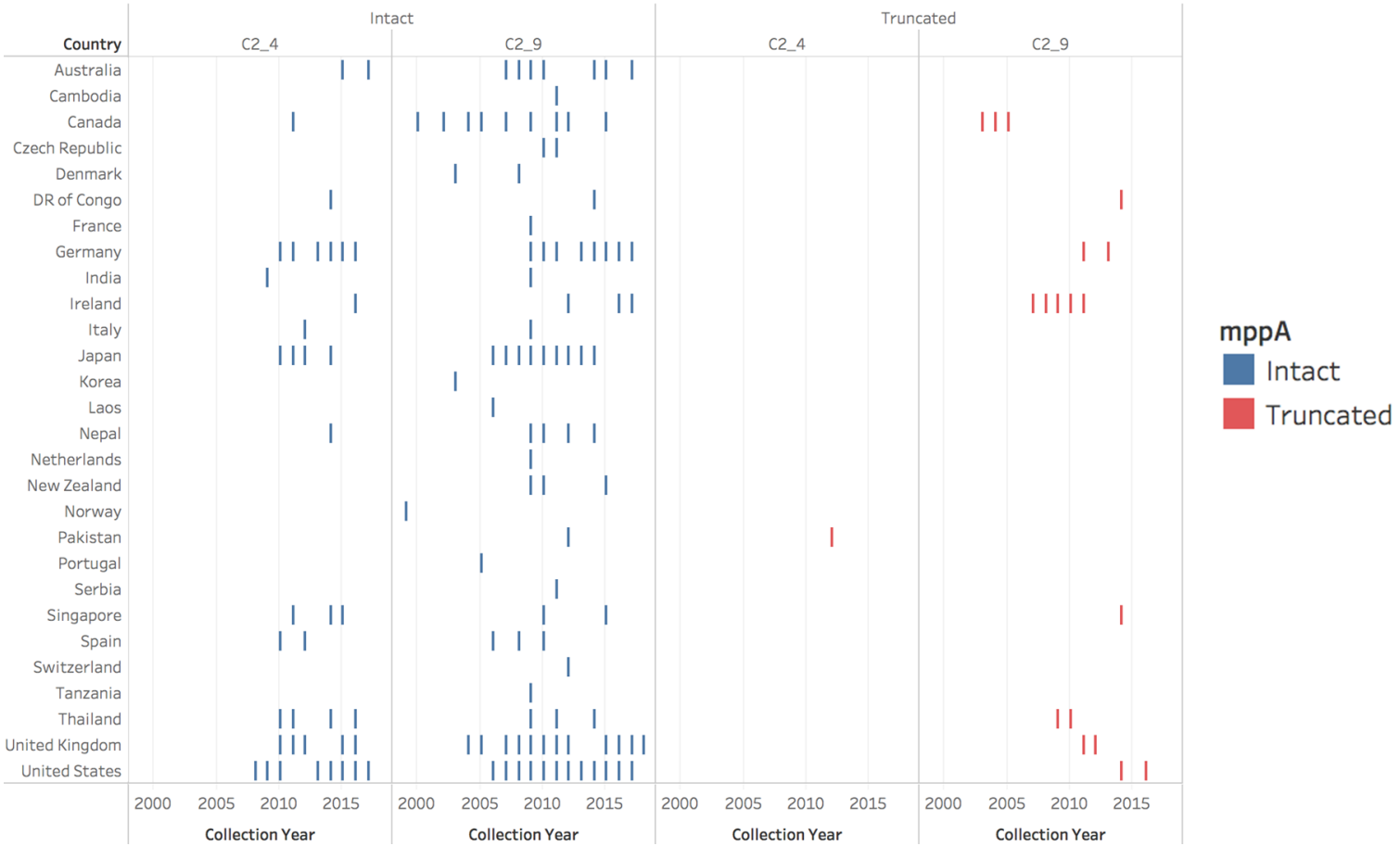

C2_9
(c)
Country
Australia
Cambodia
Canada
China
Czech Republic
Denmark
DR of Congo
France
Germany
India
Ireland
Italy
Japan
Korea
Laos
Nepal
Netherlands
New Zealand
Pakistan
Portugal
Serbia
Singapore
South Africa
Spain
Sweden
Switzerland
Tanzania
Thailand
United Kingdom
United States
Vietnam
2000
2002
2004
2006
2008
2010
2012
2014
2016
2018
Collection Year
blaCTX-M
CTX-M-14
CTX-M-15
CTX-M-27



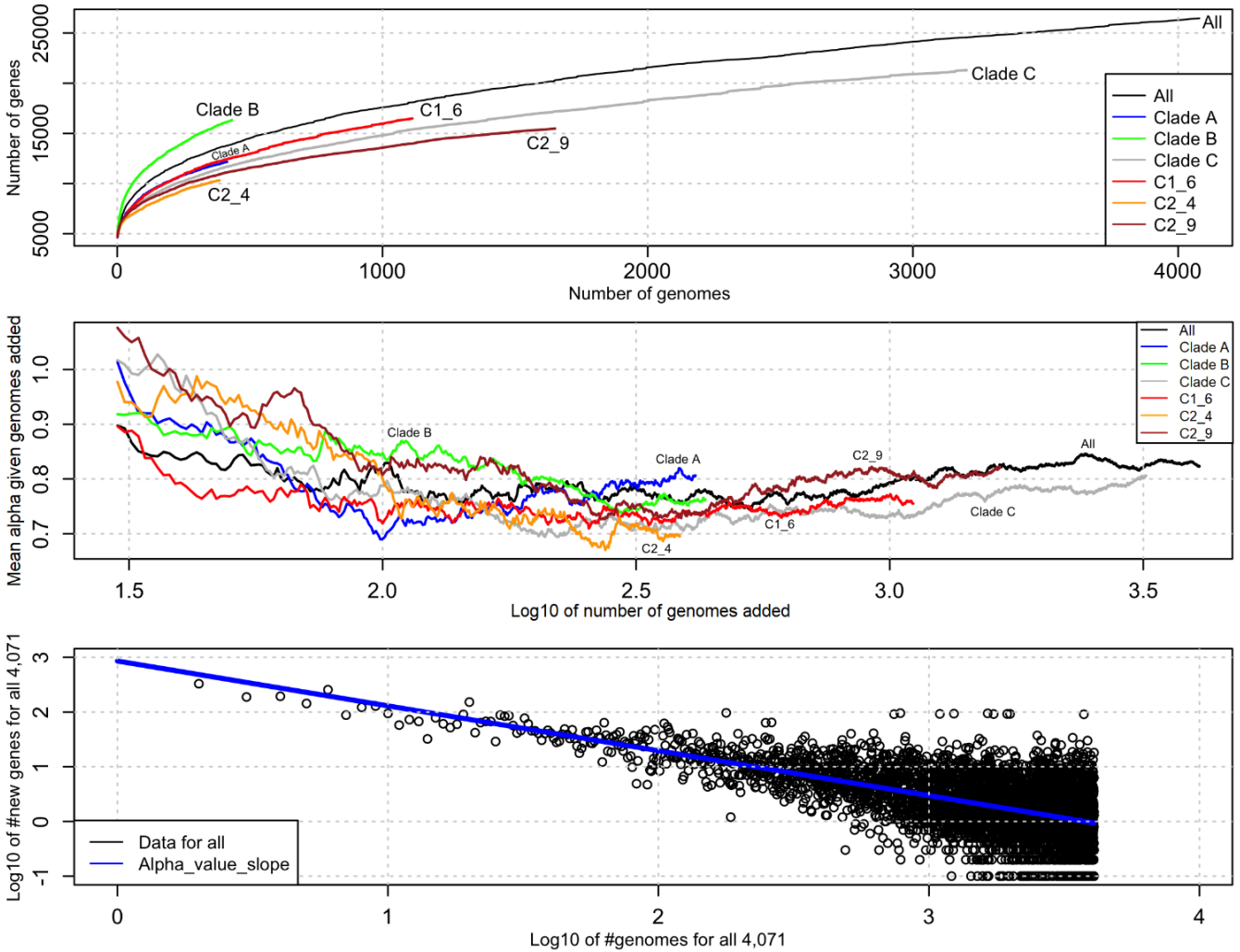
Supplementary Figure S5.7. Bubble graph of ST131 subclades C1_6, C2_4 and C2_9 and their corresponding *bla*$_{CTX\text{-}M}$ alleles (*bla*$_{CTX\text{-}M\text{-}14}$ in orange, *bla*$_{CTX\text{-}M\text{-}15}$ in turquoise, *bla*$_{CTX\text{-}M\text{-}27}$ in green, and *bla*$_{CTX\text{-}M\text{-}14/15}$ together in red). Data were plotted and drawn using Tableau v.10.1.



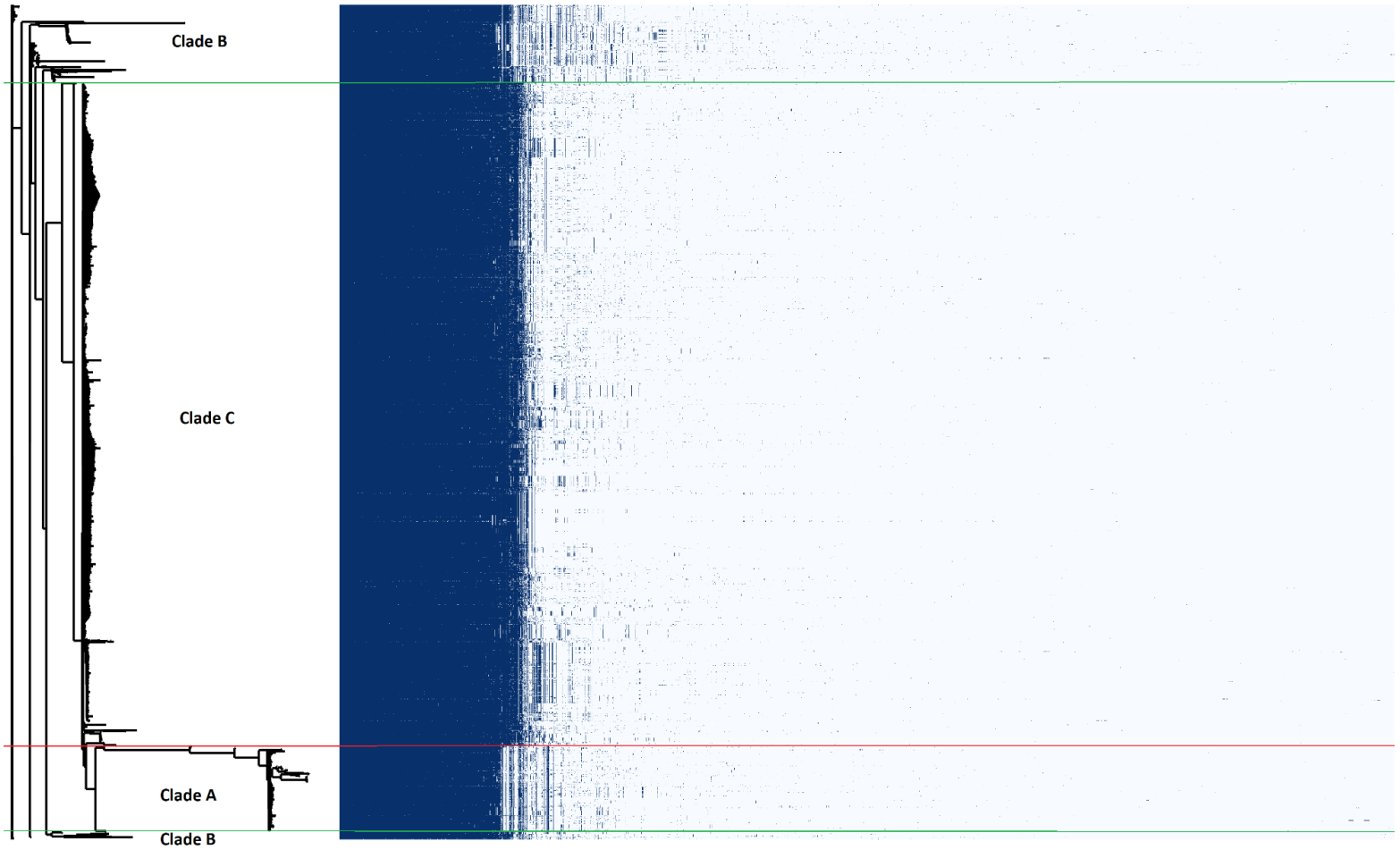
Supplementary Figure S5.8. Representative example of the *bla*CTX-M-14/15/27-positive contigs' AMR genes and MGEs annotated using Prokka and MARA. Some contigs were too short to show additional annotations, which can be estimated based on the longer contigs. (a) C1_6 isolates had the most frequent incidence of *bla*CTX-M-14-positive contigs that were typically in IS*Ecp1*-*bla*CTX-M-14-IS*903B* TUs, though with variations such as a 3' IS*903C* element instead. (b) C2 isolates tended to have a *bla*CTX-M-15 gene flanked by a 5' ISEcp1 and a Tn2 or the orf-477-Tn2 in tandem at the 3' as a 2,971 bp IS*Ecp1*-*bla*CTX-M-15-orf477Δ-*Tn2* TU. These were most likely on an IncF plasmid for the plasmid-encoded variants, but many within C2_9 had this TU chromosomally inserted at the *mppA* gene due to local sequence homology with IS*Ecp1*'s 14-bp 3' inverted repeat (IRR). (c) C1_6 isolates also had the highest incidence of *bla*CTX-M-127-positive contigs that typically had a similar IS*Ecp1*-*bla*CTX-M-14-IS*903B* as per (a).



Supplementary Figure S5.9. Chromosomal insertion of *bla*$_{CTX\text{-}M\text{-}15}$ was indicated by the truncation of the *mppA* gene and was usually observed among C2 ST131 strains. Intact *mppA* was shown in blue bars while the truncated ones were in red and were mainly observed in recent (2003-2017) samples from C2_9 group.



Supplementary Figure S5.10. Top: The average number of genes in the ST131 pangenome (y-axis) increased sub-linearly as the 4,071 genomes were added (x-axis) indicating an open pangenome for the whole collection (black) and the clades and subclades: A (blue), B (green), C (grey), C1_6 (red), C2_4 (orange) and C2_9 (brown). Middle: The *alpha* value varied with numbers of genomes sampled (shown for >30 genomes) and attained stability once the number examined exceeded 2500 (x-axis, shown on a $\log_{10}$ scale): the average *alpha* for all was 0.8123±0.024; for clade A 0.780±0.045; for clade B 0.796±0.046; for clade C 0.766±0.031; for C1_6 0.749±0.017; for C2_4 0.754±0.077; and for C2_9 0.799±0.038. Bottom: Across the 4,071 genome assemblies, *alpha* was estimated as 0.8196 such that the median number of new genes added per isolated was 2.1: the blue line indicates the regression slope *alpha*.



Supplementary Figure S5.11. A phylogeny of all ST131 (left) with their corresponding pangenome gene presence (blue) and absence (white) frequencies represented for each of the 26,479 genes detected. In the latter matrix, the 3,712 core genes are shown first, followed by the 242 soft core genes, 1,018 shell genes found 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 clade C (middle bounded by green and red lines).

5.7 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

Assefa, S., Keane, T.M., Otto, T.D., Newbold, C., Berriman, M., 2009. ABACAS: algorithm-based automatic contiguation of assembled sequences. Bioinformatics 25, 1968–1969. https://doi.org/10.1093/bioinformatics/btp347

Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data. Available at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (Accessed: 3 October 2018).

Banerjee R, Johnson JR. 2014. A new clone sweeps clean: the enigmatic emergence of Escherichia coli sequence type 131. Antimicrob Agents Chemother 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.

Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 1999 27(2):573-80
Bushnell B. BBMap short read aligner. 2016.

Chen, S. et al. (2018) 'fastp: an ultra-fast all-in-one FASTQ preprocessor', Bioinformatics, 34(17), pp. i884–i890. doi: 10.1093/bioinformatics/bty560.

Cheng, L., Connor, T.R., Sirén, J., Aanensen, D.M., Corander, J., 2013. Hierarchical and Spatially Explicit Clustering of DNA Sequences with BAPS Software. Mol Biol Evol 30, 1224–1228. https://doi.org/10.1093/molbev/mst028

Croucher, N. J. and Didelot, X. (2015) 'The application of genomics to tracing bacterial pathogen transmission', Current Opinion in Microbiology. (Host–microbe interactions: bacteria Genomics), 23, pp. 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 DOI: 10.1128/mSphere.00130-19

Downing, T. 2015 'Tackling Drug Resistant Infection Outbreaks of Global Pandemic Escherichia coli ST131 Using Evolutionary and Epidemiological Genomics', Microorganisms, 3(2), pp. 236–267. doi: 10.3390/microorganisms3020236.

Duchêne, S., Giallonardo, F.D., Holmes, E.C., Vaughan, T.G., 2019. Inferring infectious disease phylodynamics with notification data. bioRxiv. https://doi.org/10.1101/596700

Ender PT, Gajanana D, Johnston B, Clabots C, Tamarkin FJ, Johnson JR. 2009. Transmission of an extended-spectrum-β-lactamase-producing Escherichia coli (sequence type ST131) strain between a father and daughter resulting in septic shock and emphysematous pyelonephritis. J Clin Microbiol 47:3780–3782. doi:10.1128/JCM.01361-09.

Ewels, P. et al. (2016) 'MultiQC: summarize analysis results for multiple tools and samples in a single report', Bioinformatics, 32(19), pp. 3047–3048. doi: 10.1093/bioinformatics/btw354.

Gibson, M.K., Wang, B., Ahmadi, S., Burnham, C.-A.D., Tarr, P.I., Warner, B.B., Dantas, G., 2016. Developmental dynamics of the preterm infant gut microbiota and antibiotic resistome. Nat Microbiol 1, 16024. https://doi.org/10.1038/nmicrobiol.2016.24

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. et al. (2013) 'QUAST: quality assessment tool for genome assemblies', Bioinformatics, 29(8), pp. 1072–1075. doi: 10.1093/bioinformatics/btt086.
Hanage WP. Two Health or Not Two Health? That Is the Question. MBio. 2019 9;10(2). pii: e00550-19. doi: 10.1128/mBio.00550-19.

Irenge LM, Ambroise J, Bearzatto B, Durant JF, Chirimwami RB, Gala JL. Whole-genome sequences of multidrug-resistant Escherichia coli in South-Kivu Province, Democratic Republic of Congo: characterization of phylogenomic changes, virulence and resistance genes. BMC Infect Dis. 2019 19(1):137. doi: 10.1186/s12879-019-3763-3.

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.

Lawson, D.J., 2015. Populations in Statistical Genetic Modelling and Inference, in: Kreager, P., Winney, B., Ulijaszek, S., Capelli, C. (Eds.), Population in the Human Sciences. Oxford University Press, pp. 108–130. https://doi.org/10.1093/acprof:oso/9780199688203.003.0004

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

Livingstone PG, Morphew RM, Whitworth DE. Genome Sequencing and Pan-Genome Analysis of 23 Corallococcus spp. Strains Reveal Unexpected Diversity, With Particular Plasticity of Predatory Gene Sets. Front Microbiol. 2018 9:3187. doi: 10.3389/fmicb.2018.03187

Ludden C, Decano A, Jamrozy D, Zhou Z, Pickard D, Horner C, Morris D, Parkhill J, Peacock SJ, Achtman M, Dougan G, Downing T, Cormican M. Genome sequencing

confirms displacement of Escherichia coli ST131 clones and antimicrobial resistance threats. mBio (submitted).

Martin, M. (2011) 'Cutadapt removes adapter sequences from high-throughput sequencing reads', EMBnet.journal, 17(1), pp. 10–12.

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.

Nikolenko, S. I., Korobeynikov, A. I. and Alekseyev, M. A. (2013) 'BayesHammer: Bayesian clustering for error correction in single-cell sequencing', BMC Genomics, 14(1), p. S7. doi: 10.1186/1471-2164-14-S1-S7.

Page, A. J. et al. (2015) 'Roary: rapid large-scale prokaryote pan genome analysis', Bioinformatics, 31(22), pp. 3691–3693. 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 Apr 24;10:834. doi: 10.3389/fmicb.2019.00834.

Partridge, S. R. and Tsafnat, G. (2018) 'Automated annotation of mobile antibiotic resistance in Gram-negative bacteria: the Multiple Antibiotic Resistance Annotator (MARA) and database', Journal of Antimicrobial Chemotherapy, 73(4), pp. 883–890. doi: 10.1093/jac/dkx513.

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.

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.

Sarkar, S. et al. (2016) 'Biofilm formation by multidrug resistant Escherichia coli ST131 is dependent on type 1 fimbriae and assay conditions', Pathogens and Disease, 74(3). doi: 10.1093/femspd/ftw013.

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.

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.

Stork, C., Kovács, B., Rózsai, B., Putze, J., Kiel, M., Dorn, Á., Kovács, J., Melegh, S., Leimbach, A., Kovács, T., Schneider, G., Kerényi, M., Emödy, L., Dobrindt, U., 2018. Characterization of Asymptomatic Bacteriuria Escherichia coli Isolates in Search of

Alternative Strains for Efficient Bacterial Interference against Uropathogens. Front Microbiol 9, 214. https://doi.org/10.3389/fmicb.2018.00214

Tonkin-Hill, G. et al. (2018) 'Fast Hierarchical Bayesian Analysis of Population Structure', bioRxiv, p. 454355. doi: 10.1101/454355.
Tonkin-Hill, G., Lees, J.A., Bentley, S.D., Frost, S.D.W., Corander, J., 2018. Fast Hierarchical Bayesian Analysis of Population Structure. bioRxiv 454355. https://doi.org/10.1101/454355

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. et al. (2017) 'Fast and accurate de novo genome assembly from long uncorrected reads', Genome Research, 27(5), pp. 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

Wick, R. R. et al. (2017) 'Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads', PLOS Computational Biology. Edited by A. M. Phillippy, 13(6), p. 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): 28–36. 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

Chapter 6: A dynamic gene repertoire associated with the mobile resistome in the pathogen *E. coli* ST131

Abstract

The human gut microbiome includes beneficial, commensal and pathogenic bacteria that possess antimicrobial resistance (AMR) genes and exchange these predominantly through conjugative plasmids. *Escherichia coli* is a significant component of the gastrointestinal microbiome and is typically non-pathogenic in this host niche. In contrast, extra-intestinal pathogenic *E. coli* (ExPEC) including *E. coli* sequence type 131(ST131) may occupy other environments like the urinary tract or bloodstream where they express genes allowing for virulence, pathogenicity and AMR. Although ST131 isolates access extra-intestinal locations by faecal-oral routes, the extent to which they share AMR genes with other non-pathogenic gut *E. coli* has not yet been investigated at the genomic level. Here, we examined the extent of AMR gene sharing between gut *E. coli* and ST131 using a reference resistome of well-characterized preterm infant AMR genes to discover an extensive shared resistome across pathogenic and non-pathogenic isolates. In addition, individual ST131 show extensive resistome diversity highlighting that the core genome may not predict AMR phenotypes well. In addition, we show that ST131's key plasmids (pEK204, pEK499 and pEK516), which drive AMR gene transfer are highly variable within ST131 clade C. Furthermore, resolving the structures, copy numbers and locations of key ST131 transposons responsible for the mobilizing the resistome, particularly *bla*$_{\text{CTX-M}}$ genes, indicates opportunities for tracking AMR gene evolution through the prism of MGE mutation. This work shows that MGE and plasmid structure vary widely in ST131 clade C by determining *bla*$_{\text{CTX-M}}$ gene-related transposon isoforms, copy numbers and genomic locations, resistome overlap with non-pathogenic *E. coli,* and plasmid structure across diverse samples. Our results thus highlight the importance of examining the accessory genome in detail as well as the likelihood that ST131 will become more multidrug-resistant in the future.

***Publication*:** in preparation for Access Microbiology in 2019 with Buthaina AlAwadi, Cian Smyth, Genevieve Smith, Hawriya AlFoori, Louisse Mirabueno, Maddy Nelson, Zoe Vance, and Tim Downing.

6.1 Introduction

The genomes of bacterial, archaeal, viral, protozoan, fungal and helminth (worm) microbes in and on humans constitute the human microbiome. The human microbioata have been increasingly linked to human infection risk, immunity and health status and so their genomics, the microbiome, have a profound influence on the development of disease and maintaining immune and metabolic homeostasis (Clemente et al. 2012).The microbiome is involved in beneficial tasks, such as food breakdown, and interaction with the host's immune system in the gastrointestinal tract (or gut), which is packed with beneficial, commensal and pathogenic microbes. A balanced community of microorganisms is crucial to the host's health, and dysbiosis of this balance is associated with inflammatory diseases and infections (Thursby and Juge, 2017); although the dysbiotic imbalance is not yet clearly defined (Olesen & Alm 2016). Extra-intestinal pathogenic *E. coli* (ExPEC), cause disease as a result of their virulence factors which allow persistence in diverse host niches (Ben Zakour et al. 2016). Consequently, in the context of this work, dysbiosis can be associated with ExPEC infection (Bäumler & Sperandio 2016), rather than alternative interpretations (Hooks & O'Malley 2017).

E. coli are prevalent in the human microbiome. Most notably, they are one of the first bacteria colonising an infant's intestine (Penders et al. 2006). *E. coli* are a very diverse bacterial species. An estimated 500-1000 commensal strains are believed to reside in the intestine alone (Conway and Cohen, 2015) and be involved in preventing pathogens from colonising the host by producing bacteriocins (Hudault et al. 2001). Pathogenic strains, such as the *E. coli* sequence type 131 (ST131), have also been found in the human microbiome. ST131 has been subdivided into multiple clades according to the *fimH* allele (Adams-Sapper et al. 2013; Price et al. 2013; Tchesnokova et al. 2013). *FimH* encodes a type 1 fimbrin D-mannose specific adhesin protein, a virulence factor involved in the attachment to host tissue, which consequently improves colonisation in the host (Petty et al. 2014). The C clade encodes the fimH30 allele and is particularly virulent due to the clade's acquisition of a plasmid carrying a β-lactamase gene (*bla*$_{\text{CTX-M-15}}$) (Johnson et al. 2013; Ben Zakour et al. 2016). *Bla*$_{\text{CTX-M-15}}$ is of particular interest due to its increased ability to hydrolyse cephalosporins (Poirel et al., 2002).

The resistome is the set of antimicrobial resistance (AMR) genes present in a genome. ExPEC with an expansive resistome may render the treatment of serious infections with antibiotics useless, and so in this chapter, we examined the shared resistome in ST131 and selected *E. coli* from GI tract microbiome samples. It is hypothesized that frequent exposure to low doses of drugs in the environment arising from livestock growth promotion or water effluent has expanded the resistome in many microbes existing within humans. These resistomes are also likely shaped by high doses of β-lactam (and other) antibiotics to treat infections, resulting in extended-spectrum β-lactamase (ESBL) gene acquisition by pathogenic and commensal bacteria alike.

ESBLs are generally located in plasmids (Cantón et al. 2012), which makes these genes mobile and susceptible to be transferred to non-pathogenic cells. Plasmids are circular self-replicating DNA molecules that can mediate the transfer of genes allowing AMR and virulence between bacterial cells by conjugation (Hinnebusch & Tilly 1993). Some of which are lysins, toxins and adhesins that allow for attachment and colonisation in the host cell. They can be classified based on their incompatibility (Inc) group, and ST131's plasmids most commonly belong to the IncF group (Shintani et al. 2015). IncF plasmids aid transfer of AMR genes between cells and were found by Johnson et al. (2016) to have played a major role in the evolution of ST131. Although plasmids in ST131 typically encode genes for post-segregation killing and stable inheritance to ensure their propagation, they can be lost or may recombine with other plasmids in the same cell (Woodford et al. 2009, Phan et al. 2015, Nicolas-Chanoine et al. 2014). As a result of this mixing and their extensive array of mobile genetic elements (MGEs) (Frost et al. 2005), plasmids may rearrange extensively even within a clonal radiation as shown in Chapter 3.

The resistome is also defined by MGEs, particularly ISEcp1 and IS26 in *E. coli* (Smet et al. 2010), which implement a mix of replicative (copy-and-paste) and conservative (cut-and-paste) self-replicating transposition processes at genomic regions with segments homology to the IS's inverted repeats (IRs) (Naito and Pawlowska, 2015). These IS elements have three main structures: one encoding the transposition enzyme called transposase (TnpA) in the middle with two short flanking IRs (Griffiths et al. 2000). ISEcp1 belongs IS family IS1380 (Smet et al. 2010). As outlined in Chapter 3, ISEcp1,

IS903D and IS26 mediate ESBL gene transfer and thus must be explored along with the plasmid and AMR genes to better understand the mobile resistome causing acute ST131 infections.

In Chapter 5, I examined genomic diversity in a large panel (4,071) of *E. coli* ST131. This identified *bla*$_{\text{CTX-M}}$ gene and accessory genome changes as key factors differentiating ST131 subclades. In addition, it showed that such ESBL genes are highly mobile in ST131, and so exploring sequence variation at transposons, *bla*$_{\text{CTX-M}}$ genes and plasmids was an emergent question not tackled in that chapter. from large-scale genomic epidemiology to factors mobilising the ST131 resistome.

In this chapter, we sought to quantify the extent of AMR gene sharing between *E. coli* gut microbiome samples and pathogenic *E. coli* ST131. In addition, examined here are plasmid genetic structures in contigs with *bla*$_{\text{CTX-M}}$ genes using representative samples from Chapter 5 with a comparison of non-pathogenic isolates including ones from the human microbiome project.

Using functional genomics, a previous work by Gibson et al. in 2016 identified analysed seven hundred and ninety-four (N=794) previously uncharacterised contigs from draft assemblies that were associated with resistance to sixteen antibiotics. The samples were derived from preterm infant microbiomes (Gibson et al. 2016). The contigs containing AMR genes were used as reference dataset in succeeding analyses in this chapter.

6.2 Methods

***Author contributions**:* I was involved in conceptualization, genomic analysis, interpreting results, drafting the paper, editing the paper and visualization the results. Buthaina AlAwadi, Caitriona Woods, Cian Smyth, Genevieve Smith, Hawriya AlFoori, Leigh Campbell, Louisse Mirabueno, Maddy Nelson and Zoe Vance implemented bioinformatic methods, genomic analysis and interpreted results. Tim Downing helped with project design, bioinformatic processing, genomic analysis and paper writing.

6.2.1 *E. coli* genome isolate collection

We examined 12 *E. coli* isolates in this chapter: three were established ST131 reference genomes: SE15, NCTC13441 and EC958 (Table 6.1). SE15 (O150:H5, *fimH41*) was a commensal isolate lacking many virulence-associated genes, unlike the other two, and was a genetic outgroup denoting the basal ST131 clade A. As outlined in Chapter 1, it has a 122-kb plasmid (pSE15) with 150 protein-coding genes (Toh et al. 2010). SE15 shares 86% of its chromosomal genes with three uropathogenic *E. coli* genomes (CFT073, UTI89, UT536). NCTC13441 and EC958 were *bla*$_{\text{CTX-M-15}}$-positive genomes from subclade C2, with pEK499 in NCTC13441 (Woodford et al. 2009), and a larger plasmid pEC958 and a smaller one (pEC958B) in EC958 (Forde et al. 2014). pEC958A (HG941719) in isolate EC958 from a UK *bla*$_{\text{CTX-M-15}}$-positive UTI that has 85% similarity to pEK499 and is missing the latter's second *tra* region due to an IS26-mediate *bla*$_{\text{TEM-1}}$ insertion (Phan et al. 2015, Forde et al. 2014). Across its 135.6 Kb length, it has 142 genes and it belongs to Inc group F1A/F2.

Name	Sequence type	Assembly type	ST131 clade	Phenotype and source
SE15	ST131	Sanger and 454 sequencing	A	Commensal, non-pathogenic (faeces)
NCTC13441		PacBio sequencing	C2_7	Pathogenic (UTI)
EC958		PacBio sequencing	C2_7	Pathogenic (UTI)
3_2_53FAA	-	HMP assembled contigs	-	Crohn's disease (gut)
83972	ST73	HMP assembled contigs	-	Asymptomatic UTI

Table 6.1. The three *E. coli* reference genomes and two Human Microbiome Project assembled contigs used in this study. EC958 and NCTC13441 were MDR pathogenic UTIs, whereas SE15 was a commensal strain acting as a negative control.

Samples 3_2_53FAA (sometimes called EC3_2_53FAA) and 83957 (sometimes called EC83972) were from the Human Microbiome Project (HMP) (The NIH HMP Working Group et al., 2009) (Table 6.1). Strain 3_2_53FAA was a colon biopsy from a 52-year-old male Canadian diagnosed with Crohn's disease. 83972 was from the urine of a Swedish girl with a three-year history of asymptomatic bacteriuria that failed to show any symptoms and had a stable renal function (Rudick et al. 2014).

Five samples were ST131 genome assemblies representing subclades C1 and C2, all of which were FQ-R pathogenic isolates from 2005-2010 (Table 6.2). Two from C1 were *bla*$_{\text{CTX-M-14}}$-positive, and the five from C2 were *bla*$_{\text{CTX-M-15}}$-positive, bar 8289_1#24 that had a *bla*$_{\text{CTX-M-14}}$ gene too. All were isolated from urine except 8289_1#34, which was a rectal swab. All belonged to rST1503 except 8289_1#24, which was in rST1850.

ID	Accession	Name	Clade	*bla*$_{\text{CTX-M}}$ gene(s)	*mppA* gene	%Matching pV130a	%Matching pEK499	Year
8289_1#35	ERR191668	MU027534Q	C1	14	I	71	57	2008
8289_1#3	ERR191636	MU028688W	C2_7	15	I	0	0	2005
8289_1#34	ERR191667	MU005425	C2_7	15	I	74	18	2010
8289_1#24	ERR191657	MU004181Y	C2_8	14 & 15	T	68	0	2009
8289_1#27	ERR191660	MJ003268P	C2_9	15	I	100	58	2010

Table 6.2. The five *E. coli* ST131 short read genome assemblies used in this study.

6.2.2 Resistance gene sources

A resistome based on Gibson et al. 2016's study on preterm infant microbiota was used as the reference database to identify overlapping AMR genes between *E. coli* microbiome and ST131 species in the sample collection. The resistome database had 794 AMR genes from 401 stool samples longitudinally collected from 84 preterm infants throughout antibiotic therapy, and so all were associated with specific antibiotic resistances. Resistance to 16 antibiotics was functionally assessed by Gibson et al. (2016) from metagenomics expression libraries that assembled 2,004 contigs encompassing these 794 AMR genes, 79% (n=627) were classified for the first time (Gibson et al. 2016).

6.2.3 Illumina library quality control and read mapping

Similar to the methodology applied in the previous chapters, paired-end Illumina HiSeq libraries were screened for low quality (phred score < 30) and short (< 50 bp) reads using Trimmomatic v0.36 (Bolger et al., 2014) and were further corrected using BayesHammer from SPAdes v3.9 (Bolger et al., 2014). The read libraries were mapped to reference sequences with SMALT v7.6 (www.sanger.ac.uk/resources/software/smalt/), and the resulting SAM files were converted to BAM format, sorted and PCR duplicates removed using Samtools v1.19 (Li et al., 2009). The reads for each isolate were indexed using the median read length for each read library for calibrating, and then mapped to reference AMR genes using GROOT (Graphing Resistance genes Out Of meTagenomes) (Rowe & Winn 2018) to the CARD (Jia et al., 2017), ARG-ANNOT (Gupta et al., 2014) and ResFinder (Zankari et al., 2012) databases (Table 6.3).

Library	Number of reads	Read length		
		Median	Mean	SD
8289_1#34_1	2,145,896	101	99.3	6.5
8289_1#34_2		100	97.7	8.0
8289_1#3_1	1,824,030	101	99.4	6.5
8289_1#3_2		100	97.6	8.2
8289_1#35_1	2,297,624	101	99.4	6.5
8289_1#35_2		100	97.7	8.0
8289_1#24_1	20,14250	101	99.3	6.5
8289_1#24_2		100	97.7	8.1
NCTC13441	2,857,729	43	42.5	1.4
SE15	418,045	218	192.2	67.3
EC958	1,514	1,486	1,401.6	549.2

Table 6.3. Read length summary statistics for each corrected FASTQ library. SD stands for standard deviation. Paired-end read library were mapped individually as per GROOT guidelines, and are denoted as _1 and _2. The read distributions differed for NCTC13441, SE15, and EC958 because they were generated using long read approaches.

6.2.4 Homology-based resistome screening and comparison

Contig annotation and protein domain recognition was implemented using the Pfam v27.0 and ProSite databases using InterProScan v5.22-61 (Jones et al. 2014). The Comprehensive Antibiotic Resistance Database (CARD) is organized via the antibiotic resistance ontology (ARO) (McArthur et al. 2013) and its protein homolog dataset (n=2,239 genes, Jia et al. 2017) was aligned with the isolates' genomes to annotate the resistomes in detail using BLAST v2.2.31. Hits with a bit score >500 and >99% homology were considered as valid matches. Detected genes and elements were verified and visualized with Artemis (Carver et al., 2012) and the Artemis Comparison Tool (ACT) (Carver et al., 2005). The alignments were visualized using R's VennDiagram v1.6.1, Seqinr v3.4-5, UpSetR v1.4.0 and WriteXLS v5.0.0 packages.

6.2.5 Source of plasmids prevalent in ST131

Sequence and annotation files for pEK499 (NC_013122.1, EU935739), pEK516 (NC_013121.1, EU935738) and pEK204 (NC_013120.1, EU935740) were downloaded. Each of these three plasmids had similarities and differences and all three are common mediators of AMR genes in ST131 (Lanza et al. 2014). All have genes allowing for stable plasmid inheritance and post-segregation killing. Detection of replicon types in the plasmids was completed using PlasmidFinder (Carattoli et al. 2014). Each *E. coli* sequence library was mapped as above to each plasmid to verify local AMR gene and MGE genetic structure and determine copy number levels, which were visualized with Artemis (Carver et al. 2012) and R v3.4.2. Plasmid and gene sequence similarity was calculated using the Sequence Identity and Similarity (SIAS) tool (http://imed.med.ucm.es/Tools/sias.html).

6.2.6 Transposable elements common in ST131: IS*26* and IS*Ecp1*

Reference sequences for IS*Ecp1* and IS*26* were aligned with ST131 plasmids pEK499, pEK516, pEK204 and pEC958 and the NCTC13441 chromosome. The reference IS*Ecp1* and IS*26* sequences were extracted from the NCTC13441 genome with SAMtools. The transposable elements were aligned using BLAST, and their genomic locations and copy numbers were visualized using SnapGene v4.3 (from GSL Biotech, snapgene.com). The

evolutionary relationships reflected in phylogenies were constructed using T-coffee alignments in ape v5.2 in R v.1.1.463.

The IS*Ecp1* and IS*26* structures on the NCTC13441 plasmid (161,069 bp) were examined in detailed to identify the IRR, *tnpA* and IRL sequences associated with these transposons. For IS*26*, the transposase gene spans 705 bp. In IS*Ecp1*, the *tnpA* gene is longer (1,262 bp) and is bounded by 14-bp IRs with IRL 5' of it and IRR at the 3' end. IS*Ecp1* uses IRL with alternative IR-like sequences to initiate transposition (Lartigue et al. 2006).

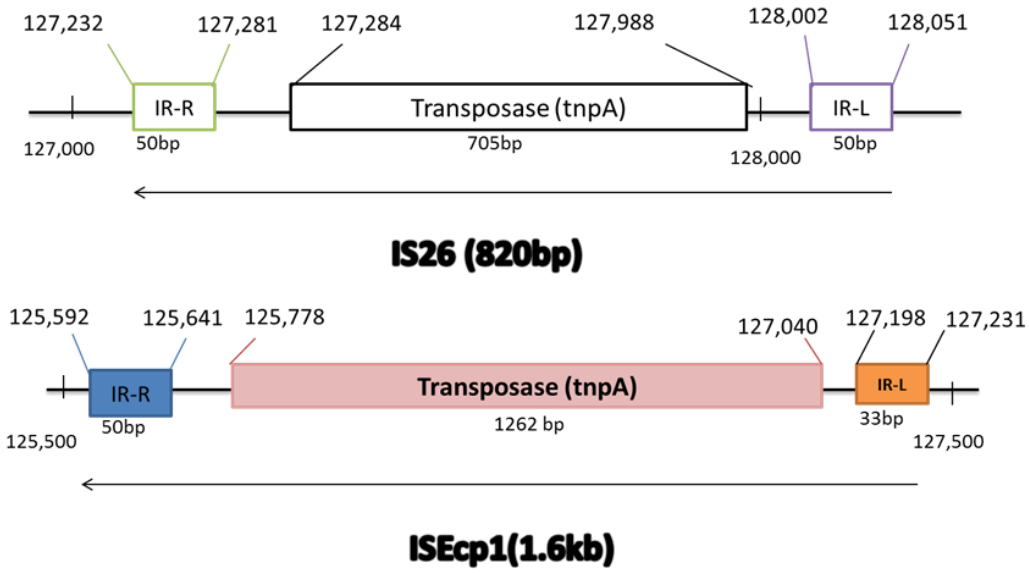


Figure 6.1. The structure of IS26 (top) and ISEcp1 (bottom) in the NCTC13441 plasmid. IS26 (820 bp) is smaller than ISEcp1 (1.6 Kb). The x-axis represents the NCTC13441 plasmid coordinates.

6.3 Results

This chapter examined the similarity in plasmid structure, transposon copy number and AMR gene content in a panel of *E. coli* genome assemblies. This linked a defined set of published AMR genes to ST131 genomes and to microbiome *E. coli* samples to inform on the resistome overlap between pathogenic and non-pathogenic isolates.

6.3.1 Variable resistome overlaps necessitate an explicit reference gene set

To relate the Gibson et al. (2016) resistome with existing AMR databases and test the robustness of a general resistome comparison, the 794 AMR genes were aligned with CARD. Using a panel of seven ST131 C1 (n=2) and C2 (n=5) libraries, the reads for each individual sequence file were mapped to the CARD, ARG-ANNOT and ResFinder databases using GROOT (Rowe & Winn 2018). This discovered considerable variation in the numbers and type of AMR genes determined by each database (Table 6.4). For comparison, faecal isolate S250 from ST131 clade B (Nicolas-Chanoine et al. 2017) had zero AMR genes in ARG-ANNOT and ResFinder, and five in CARD. This highlighted extensive potential AMR gene diversity unique to the C subclade that was highly variable within it.

Sample	Subclade	ARG-ANNOT	CARD	ResFinder
8289_1#35	C1	44	52	45
8289_1#3	C2_7	17	39	22
8289_1#34	C2_7	9	17	15
8289_1#24	C2_8	56	71	55
8289_1#27	C2_9	6	9	6

Table 6.4. Number of AMR genes identified in non-reference ST131 strains from clade by sequence similarity search against AMR databases ARG-ANNOT, CARD and ResFinder.

Consequently, we used the 794 AMR genes from (Gibson et al. 2016) as a reference set of contigs associated with antibiotic-resistance in preterm infants denoting the resistome of interest for this chapter. These came from 2,004 AMR contigs experimentally tested for resistance to 16 antibiotics, and 79% of the 794 were novel genes, highlighting an opportunity to investigate an established preterm infant resistome, rather than an undefined resistome in these established but inconsistent AMR gene databases.

6.3.2 An extensive shared resistome between ST131 and microbiome assemblies

The resistome overlap was initially compared between the three long read sequence libraries of the two ST131 C2 reference genomes (NCTC13441 and EC958) and the SE15 reference genome from ST131 clade A relative to the two HMP. This initial analysis of five isolates distinguished a specific pathogenic resistome from a shared resistome (Figure 6.2). The latter encompassed 244 AMR genes shared by all five, where the most frequent gene function was to resist penicillin (98), then chloramphenicol (59), next cephalosporins (42), followed by tetracycline (37) and then monobactams (8). This showed that non-pathogenic isolated including HMP samples 83972 and 3_2_53FAA along with commensal SE15 possessed an extensive resistome such that pathogenic ST131 genomes only had 7% additional AMR genes.

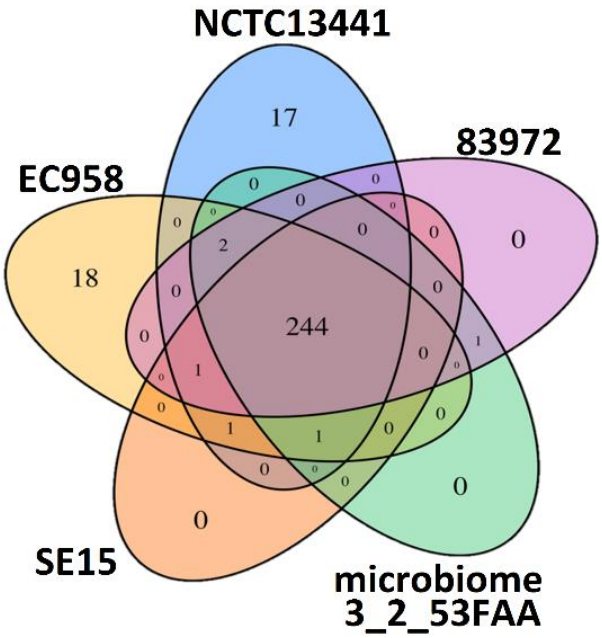


Figure 6.2. *E. coli* chromosomal AMR gene overlap showed an extensive shared resistome of 244 genes across three ST131 samples and two HMP samples (83972 and 3_2_53FAA). NCTC13441 and EC958 from subclade C2 had 17 and 18 genes (respectively) unique to their genomes, along with two genes shared with 83972.

Most chromosomal AMR genes unique to the C2 reference genomes encoded penicillins, but with notable differences within these two isolates with closely related core genomes. All 17 of the genes unique to NCTC13441 were annotated as *bla*$_{TEM}$ or *bla*$_{OXY}$ and therefore were Ambler class A β-lactamases (Table 6.5), whereas 14 out of the 18 unique to EC958 were annotated as *bla*$_{CMY}$ (AmpC) and so were Ambler class C ones. EC958 also had two genes encoding monobactam enzymes, one for tetracycline-resistance, and another for cephalosporin-resistance (Table 6.6). This difference in Ambler class AMR gene type indicated that NCTC13441 may have been exposed to more beta-lactam compounds, facilitating ESBL gene gain, but that EC958 may have encountered higher levels of cephalosporins, and so acquired more AmpC genes.

The HMP samples 3_2_53FAA and 83972 shared a ceftazidime-R gene encoding peptidoglycan glycosyltransferase *Ftsl* (contig 5_F2_CZ.13:475-2241), which could be a PBP isoform (Sun et al. 2014). Two contigs were shared between NCTC13441, EC958 and the 3_2_53FAA and 83972 microbiomes, one was homologous to *marR* (contig 4_C8_AP.19:94-609) and the other to a *Klebsiella* phosphonate C-P lyase system gene *phnK* that may act as an ABC transporter (1_E2_AP.3:1-459). The sole contig shared across SE15, NCTC13441 and EC958 alone (1_E2_PE.16:67-363) had an AraC domain: *araC* genes are in most *Enterobacteriaceae*.

AMR Contig IDs	Antibiotics	Gene/Allele
1_H5_AP.1:151-1026	AP - Ampicillin	OXY-1-1
1_F7_AX.35:206-1081 1_G7_AZ.7:473-1348	AX - Amoxicillin AZ - Aztreonam	OXY-1-2
1_A2_AXC.6:575-1444 1_A2_AZ.2:593-1126	AXCL - Amoxicillin+Clavulanate AZ - Aztreonam	OXY-2-3
4_D5_TE.10:5256-6155	TE - Tetracycline	OXY-6-2
1_B6_AP.5:907-1782 5_F2_PI.107:117-992 2_B7_PI.9:1-501	AP - Ampicillin PI - Piperacillin PI - Piperacillin	OXY-6-4
5_F2_PI.56:2439-3146	PI - Piperacillin	TEM-57
1_C6_PI.2:2496-3188	PI - Piperacillin	TEM-104
1_B6_PI.2:389-1249	PI - Piperacillin	TEM-116
5_B1_AP.7:103-963 4_B7_AX.4:127-987	AP - Ampicillin AX - Amoxicillin	TEM-215
4_C8_AP.12:3312-3698 1_F7_PE.7:2887-3369	AP - Ampicillin PE - Penicillin G	TEM-220

4_C8_PI.1:1834-2316	PI - Piperacillin

Table 6.5. The 17 AMR contigs unique to NCTC13441 were mainly Ambler class A β-lactamases.

AMR Contig IDs	Antibiotics	Gene/Allele
4_B7_PE.29:3-161	PE - Penicillin G	CMY-37
5_F2_PE.9:1100-2296	PE - Penicillin G	CMY-51
4_B7_PE.8:1042-2187	PE - Penicillin G	
1_A2_AXC.2:1124-2269	AXCL - Amoxicillin+Clavulanate	CMY-66
2_B8_PE.6:1479-2624	PE - Penicillin G	
5_F2_AP.20:1330-2040	AP - Ampicillin	CMY-67
5_F2_AP.23:2-217	AP - Ampicillin	
1_H5_AP.3:223-1368	AP - Ampicillin	CMY-85
2_B8_AX.7:1708-2418	AX - Amoxicillin	
2_B8_AX.7:2393-2854	AX - Amoxicillin	
1_H5_AXCL.2:1315-2460	AXCL - Amoxicillin+Clavulanate	
2_B8_PI.1:1082-2227	PI - Piperacillin	
5_D1_AP.16:231-722	AP - Ampicillin	CMY-98
5_D1_AXCL.7:358-1503	AXCL - Amoxicillin +Clavulanate	
5_D1_AXCL.17:2312-3022	AXCL - Amoxicillin+Clavulanate	
5_B1_CZ.6:140-1285	CZ - Ceftazidime	CMY-101
4_B7_PE.24:1-315	PE - Penicillin G	
1_H5_PI.9:561-1040	PI - Piperacillin	CMY-105

Supplementary Table 6.6. The 18 AMR contigs unique to EC958 were mainly Ambler class C β-lactamases.

To evaluate the shared and specific resistome in more detail, this resistome comparison based on the 794 AMR genes from (Gibson et al. 2016) was extended by comparing these five with four ST131 genome assemblies (8289_1#3, 8289_1#24, 8289_1#27, 8289_1#34) using a BLAST. This found 294 AMR-related contigs in total and demonstrated that the ST131 isolates had far more AMR genes (248-285) (Table 6.7). It also indicated that most isolates from C2_7 (8289_1#3, 8289_1#34 and NCTC13441) contained identical sets of AMR-related contigs (n=267), which as previously did not have 18 *bla*$_{CMY}$ (AmpC) genes present in EC958 alone, which had 285 in total. This comparison highlighted an extensive shared resistome of 207 genes where 3_2_53FAA also had an additional 39 AMR genes absent in 83972, and only two genes were associated with ST131 isolates alone (Figure 6.3). A larger number of AMR genes (n=10) were present in

all subclade C2 isolates, in addition to eight *bla*$_{OXY}$ genes present in all these, bar 8289_1#27. The sole isolated that was both *bla*$_{CTX\text{-}M\text{-}14}$-positive and *bla*$_{CTX\text{-}M\text{-}15}$-positive (8289_1#24) had nine extra *bla*$_{TEM}$ genes.

***E. coli* sample (clade)**	**Number of unique AMR contigs present**
83972 (HMP)	208
EC3_2_53FAA (HMP)	247
SE15 (A)	248
8289_1#27 (C2_9)	259
NCTC13441 (C2_7)	267
8289_1#3 (C2_7)	267
8289_1#33 (C2_7)	267
8289_1#24 (C2_8)	276
EC958 (C2_7)	285

Table 6.7. Unique NICU AMR contigs identified by BLAST where AMR contigs were classified as present where their E-value < 1e-10.

6.3.3 Extensive AMR, plasmid persistence and conjugation gene differences between common ST131 plasmids

The functional gene properties in relation to AMR, plasmid persistence and conjugation were examined by searching for homologous sequences in CARD for the three most common ST131 plasmids: pEK204, pEK499 and pEK516. These were verified in the plasmid's annotation files. This showed that each has different features: pEK204 alone is conjugative, pEK499 has the most AMR genes, and pEK516 is similar to pEK499 but has certain plasmid persistence genes present in pEK204 but absent in pEK499 (Table 6.8). Using the two HMP genome assemblies and SE15 as a commensal outgroup, we subsequently examined in more detail the genes on these plasmids present in ST131 clade C.

IncF2/F1A plasmid pEK499 lacks a *traX* gene for conjugation and was associated the historical acquisition of a *bla*$_{CTX\text{-}M\text{-}15}$ gene in ST131 that is tightly correlated with its pandemic nature (Livermore et al. 2007). It is 117.5 Kb and has 185 genes, including ones encoding *bla*$_{CTX\text{-}M\text{-}15}$, *bla*$_{TEM\text{-}1}$ and *bla*$_{OXA\text{-}1}$. This plasmid is stably inherited because it has

post-segregation killing gene *hok* and modulator *mok*, toxin-antitoxin system genes (*pemI*-*pemK*, *ccdA*-*ccdB*), and two copies of virulence-associated genes, *vagC* and *vagD*.

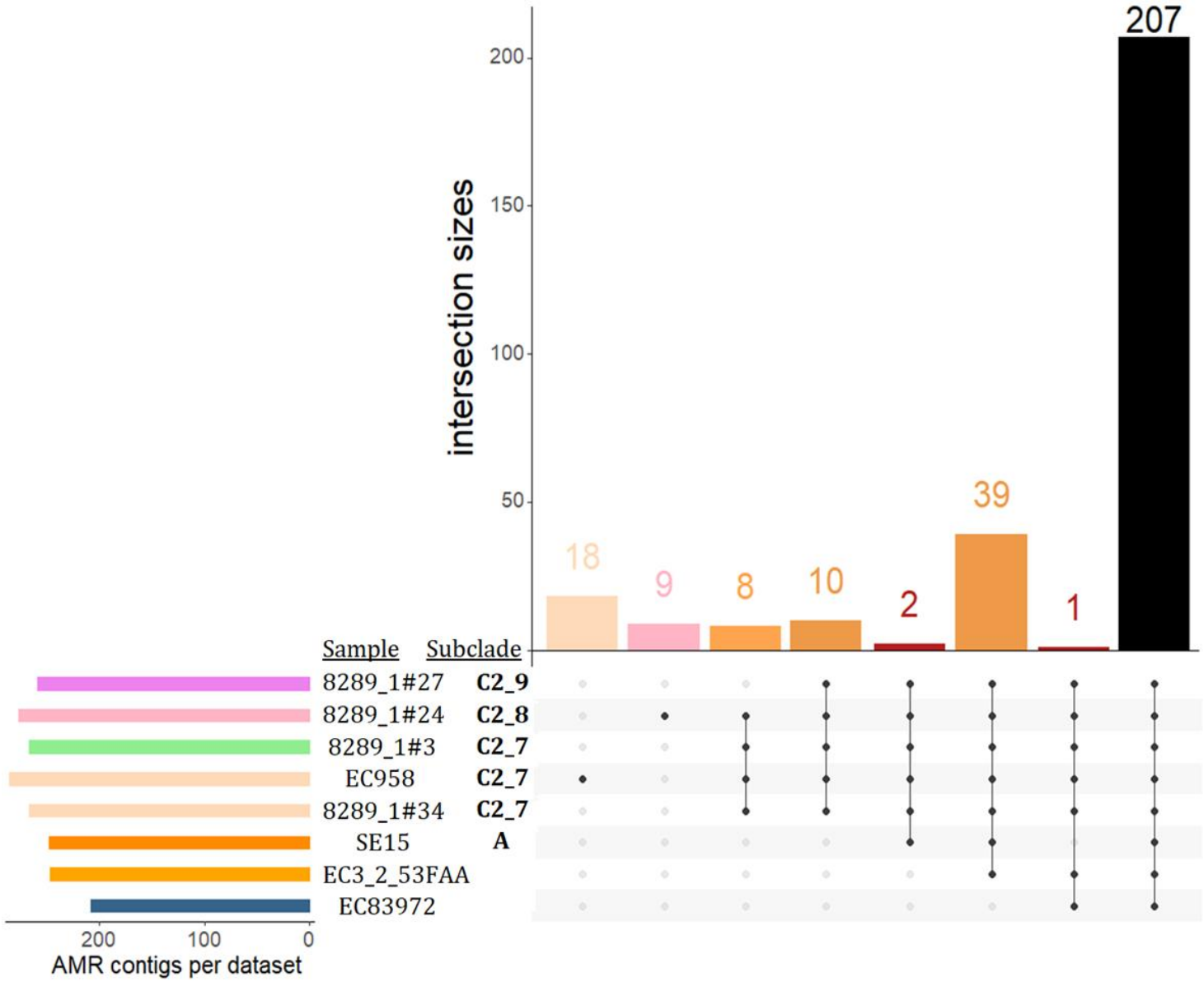


Figure 6.3. An UpsetR overlap of preterm infant AMR genes across four ST131 subclade C2 genome assemblies (8289_1#3, 8289_1#24, 8289_1#27, 8289_1#34), two ST131 reference genomes and two microbiome assemblies (EC83972 and 3_2_53FAA). The main diagram showing the intersection sizes (y-axis) and the numbers of AMR genes per category showed that the majority (207) of AMR genes were shared across isolates. The lower diagram indicates the numbers of AMR contigs per dataset, the sample name, the subclade and the corresponding categories by a filled black circle.

The other *bla*$_{CTX\text{-}M\text{-}15}$-positive plasmid (pEK516) is structurally similar to pEK499 with ~75% similarity (Woodford et al. 2009). This plasmid is shorter at 64.6 Kb, is in Inc group F2A, and has 103 genes including ones encoding *bla*$_{CTX\text{-}M\text{-}15}$, *bla*$_{TEM\text{-}1}$ and *bla*$_{OXA\text{-}1}$ like pEK499. This plasmid is stably inherited too because it has type I partitioning locus (*parM* and *stbB*) absent in pEK499, which are also in pEK204. It also has toxin-antitoxin system genes *pemI*/*K*. Although pEK516 is non-conjugative and lacks *traX* and *traC*, it retains *traA/B/D/E/K/L/M/P/V/Y*, including *traP* encoding conjugative transfer system protein TraP.

The third (pEK204) is 93.7 Kb, has 112 genes and is in Inc group I1. It is structurally similar to IncI1 plasmid R64 (Woodford et al. 2009). Notably, pEK204 has a *tra* region (transfer genes) adequate for conjugation, though some *tra* genes may be lost during culturing (Woodford et al. 2009). This plasmid is stably inherited too because it has the type I partitioning locus (*parM* and *stbB*). PEK204 encodes a 9.3 Kb region containing *bla*$_{TEM\text{-}1b}$ and also an inactive Tn*3-tnpA* transposase-encoding element with an in-frame IS*Ecp1*-*bla*$_{CTX\text{-}M\text{-}3}$ insertion with a 5' orf477-*tnpA*-*tnpR* structure. *Bla*$_{CTX\text{-}M\text{-}3}$ differs from *bla*$_{CTX\text{-}M\text{-}15}$ by a single R240G substitution, and so the *bla*$_{CTX\text{-}M\text{-}3}$ genes detected here were likely to encode *bla*$_{CTX\text{-}M\text{-}15}$. The 14 bp IRL at the 5' end of IS*Ecp1* and IRR at the distal end of the inverted orf477 element assists in mobilising *bla*$_{CTX\text{-}M\text{-}3}$, however an additional IRR at a *impB* gene 3' of the *bla*$_{TEM\text{-}1b}$ gene (7.4 Kb further away) also allows mobilisation of this whole 9.3 Kb unit (Dhanji et al. 2011), which has been found in IncFIA, IncFIA-FIB, IncN and IncY plasmids and arose on pCOL1b-P9-like plasmid (Woodford et al. 2009). ImpB is an error-prone DNA subunit functioning (with ImpA) like the UmuDC error-prone DNA repair system to combat mutations from UV radiation (Runyen-Janecky et al. 1999).

Group	Name	Gene product function	pEK499	pEK516	pEK204
AMR genes	*aac(3)-II*	aminoglycoside N(3')-acetyltransferase III; resistance to gentamicin, netilmicin, tobramycin, sisomicin		1	
	aac(6')-Ib-cr	aminoglycoside N(6')-acetyltransferase type Ib-cr; quinolone resistance	1	1	
	aadA5	aminoglycoside resistance protein	1		
	catB4	chloramphenicol acetyltransferase; inactivates chloramphenicol	1	1	
	ctx-m-15	extended spectrum beta-lactamase	1	1	
	ctx-m-14	extended spectrum beta-lactamase			
	ctx-m-3	extended spectrum beta-lactamase			1
	dfrA7	dihydrofolate reductase type VII; trimethoprim resistance	1		
	mph(A)	macrolide 2-phosphotransferase; inactivates erythromycin	1		
	oxa-1	beta-lactamase OXA-1 precursor	1	1	
	sulI	dihydropteroate synthase; sulfonamide resistance protein	1		
	tem-1	beta-lactamase	1	1	1
	tetA	tetracycline resistance protein class A	1	1	
Segregation genes	*ccdA and ccdB*	plasmid maintenance protein CcdA and CcdB; Antitoxin component of a type II toxin-antitoxin (TA) system which inhibits the post-segregational killing (PSK) of plasmid-free cells	1		
	hok	post-segregation killing protein (small toxic polypeptide)	1	1	
	mok	modulator of post-segregation killing protein	1	1	
	parM	plasmid segregation protein; stable plasmid inheritance protein A synonym StbA in pC15-1a		1	1
	pemI	stable plasmid inheritance transcriptional regulator/antitoxin	1	1	
	pemK	stable plasmid inheritance protein toxin-antitoxin system pemI-pemK/toxin	1	1	
	stbB	similar to stable plasmid inheritance protein B [plasmid R100]		1	1
	vagC	virulence-associated protein vagC (1/2); toxin addiction system; antitoxin	1		
	vagD	virulence-associated protein vagD (1/2); toxin addiction system; toxin	1		
Conjugation genes	*traC*	conjugal transfer ATP-binding protein; associated with conjugal transfer	1		1
	traX	responsible for the N-terminal acetylation of F pilin; involved in F pilus assembly			1

Table 6.8. AMR, segregation and conjugation genes in pEK499, pEK516 and pEK204 (top). 1 indicates presence. The $bla_{\text{TEM-1b}}$ gene was 860 bp in pEK204 and 728 bp in pEK499 and pEK516. Genes *parM* and *stbB* were 980 bp in pEK204 but were 962 bp in pEK499 and pEK516.

6.3.4 Variable pEK499 and pEK516 gene differences within ST131 distinct from microbiome samples

Plasmid pEK499 was aligned with the two HMP samples, SE15 and five ST131 clade C genome assembles: 8289_1#3, 8289_1#24, 8289_1#27, 8289_1#34 and 8289_1#35. In addition, the reads for these five ST131 clade C libraries were mapped to pEK499 to infer their local copy numbers. SE15's pECSF1 (aka pSE15) plasmid was from a commensal isolate to serve as a genetically distinct outgroup (Toh et al 2010). This plasmid is 122.3 Kb long and is in Inc group F2A/F1B and has 150 genes, none of which are associated with AMR (Toh et al. 2010).

This showed that the Clade C samples had larger segments homologous to pEK499 relative to SE15 and the HMP samples, but that plasmid structure remained highly variable with these five clade C isolates (Figure 6.4). The segregation genes at 16-18 Kb were partially absent in some clade C samples except for 8289_1#24 and 8289_1#27. The (inactive) conjugation (*tra*) genes at 22-36 Kb were largely conserved in all ST131 and SE15, except for 8289_1#34. Genes encoding bla_{TEM}, $bla_{\text{OXA-1}}$ and $bla_{\text{CTX-M-15}}$ were at 40, 58 and 63 Kb (respectively). This indicated that bla_{TEM} was in 8289_1#3, 8289_1#24 and 8289_1#35, and that bla_{OXA} was in 8289_1#24, 8289_1#27, 8289_1#34 and 8289_1#35, such that 8289_1#35 from subclade C1 and 8289_1#24 from subclade C2 had the most diverse complement of ESBL genes.

Like pEK499, Clade C samples had extended regions similar to pEK516 relative to SE15 and the HMP samples, particularly for 8289_1#3 (Figure 6.5). The pEK516 region at 22-61 Kb is inverted but very similar to pEK499 and was largely present in the clade C isolates and SE15 here, though 8289_1#34 had notably less homology at the (inactive) conjugation (*tra*) genes at 33-40 Kb and segregation genes at 45-48 Kb. Regions adjacent to the genes encoding $bla_{\text{OXA-1}}$, $bla_{\text{CTX-M-15}}$ and bla_{TEM} were at 12, 20 and 24 Kb (respectively) had higher levels of copy number variation. As above, 8289_1#3, 8289_1#24 and 8289_1#35 had bla_{TEM}, and 8289_1#24, 8289_1#27, 8289_1#34 and 8289_1#35 had bla_{OXA}.

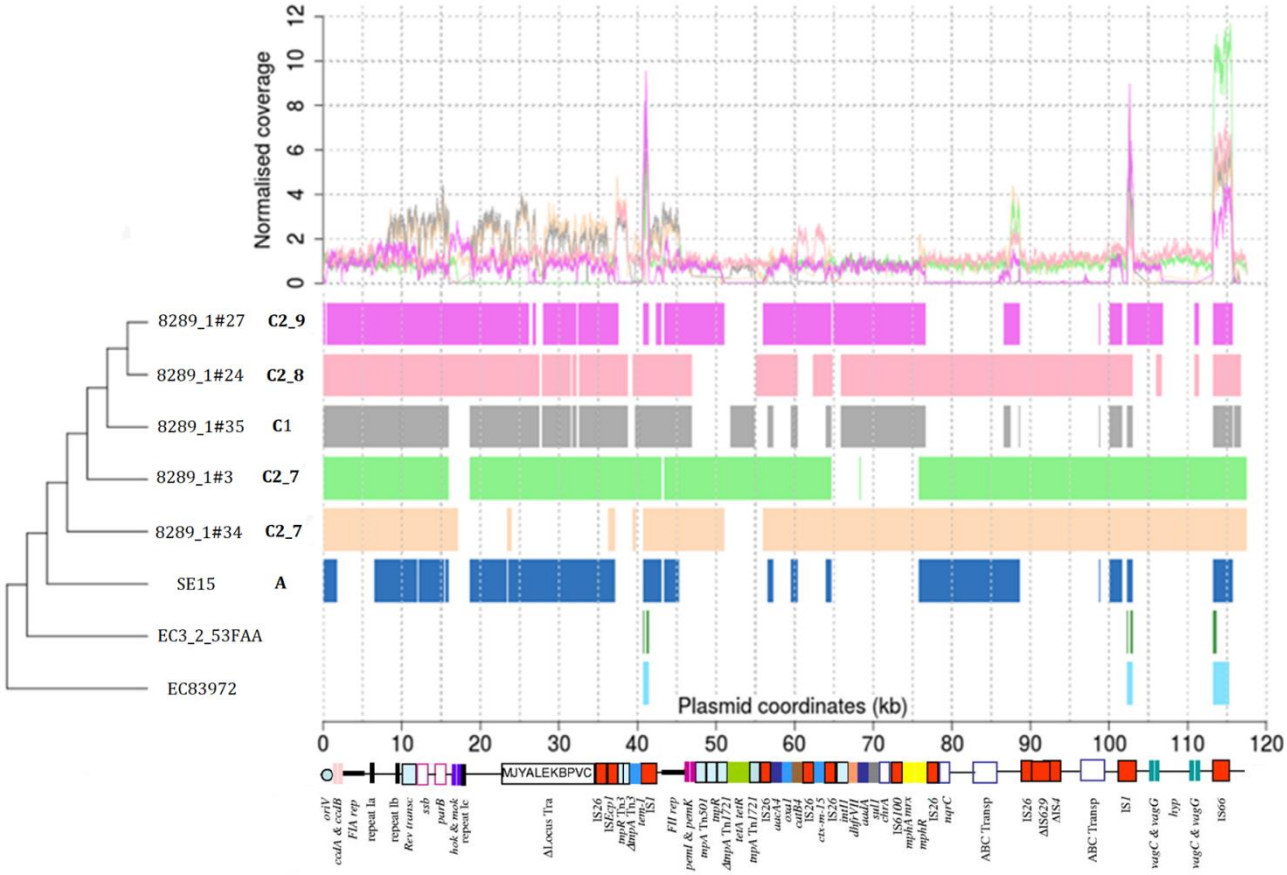


Figure 6.4. Regions of similarity and normalized copy number variation for pEK499 (117,536 bp). A coloured cladogram (left) shows the genetic relationships of the isolates examined where the two most divergent ones (3_2_53FAA in dark green and 83972 in cyan) did not have pEK499, and the commensal SE15 (navy) had limited regions of similarity. The ST131 isolates BLAST alignments did show higher levels of similarity for 8289_1#34 in beige, 8289_1#3 in light green, 8289_1#35 in grey, 8289_1#24 in pink and 8289_1#27 in mauve, though different regions were homologous across the samples. Genes encoding *bla*$_{TEM}$, *bla*$_{OXA-1}$ and *bla*$_{CTX-M-15}$ were at 40, 58 and 63 Kb (respectively). The regions of high copy number were IS1 (at 41 and 103 Kb) and IS66 (at 113 Kb) elements. The annotation is modified from Woodford et al. (2009).

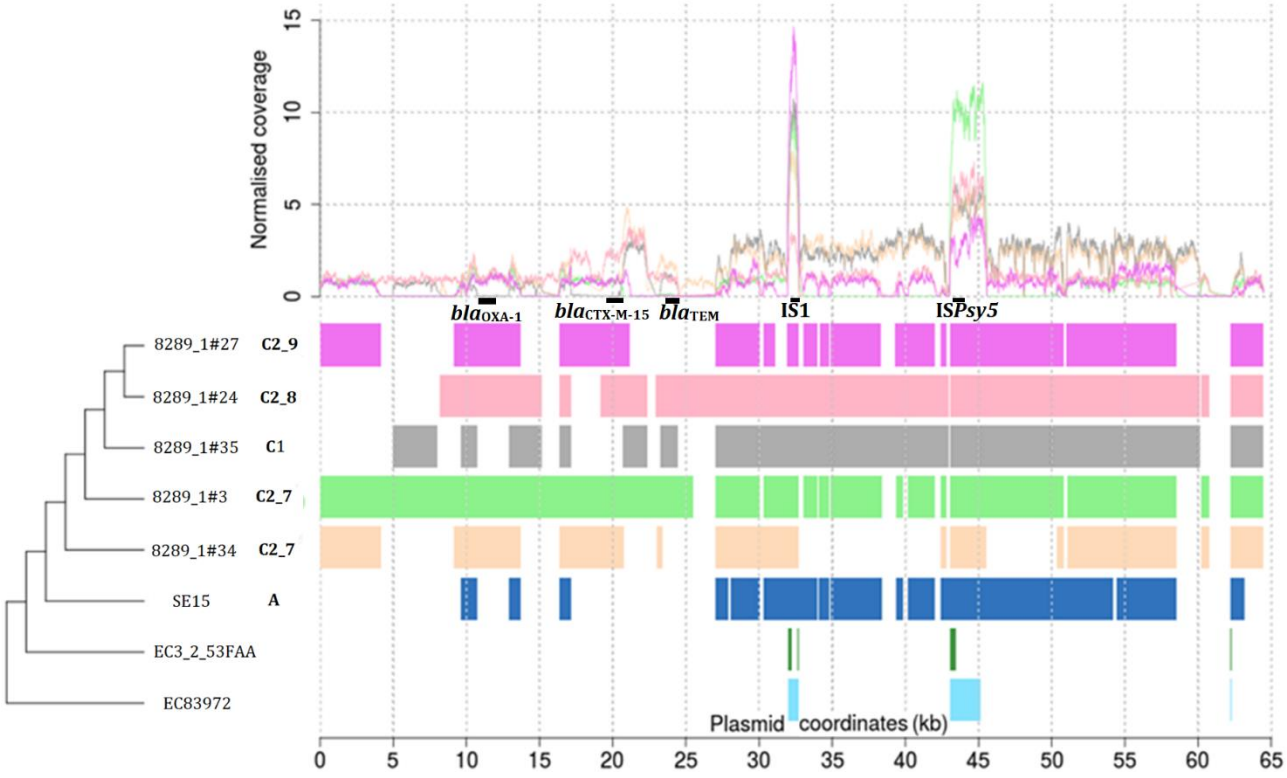


Figure 6.5. Regions of similarity and normalized copy number variation for pEK516 (64,471 bp). A coloured cladogram (left) shows the genetic relationships of the isolates examined where the two most divergent ones (3_2_53FAA in dark green and 83972 in cyan) did not have pEK516, and the commensal SE15 (navy) had reduced regions of similarity. The ST131 isolates BLAST alignments had higher levels of similarity for 8289_1#34 in beige, 8289_1#3 in light green, 8289_1#35 in grey, 8289_1#24 in pink and 8289_1#27 in mauve, though variable regions were homologous across the samples. Genes encoding *bla*$_{\text{OXA-1}}$, *bla*$_{\text{CTX-M-15}}$ and *bla*$_{\text{TEM}}$ were at 12, 20 and 24 Kb (respectively) highlighted with black boxes. The regions of high copy number were IS*1* (32 Kb) and IS*Psy5* (with unannotated genes at 43-45 Kb) elements shown by black boxes.

6.3.5 Extensive pEK204 homology to a single ST131 clade C isolate

Long regions homologous to pEK204 were found in 8289_1#27 among the isolated examined here (Figure 6.6). This sample had the complete I1 replicon, *oriT* region, *tra* region, shufflons subject to site-specific recombinase activity. 8289_1#27 also had a *pilL-pilV* cluster encoding a second major pilin subunit of the type IV family associated with

enhanced cell adherence and biofilm formation in enteroaggregative and Shiga toxin-producing *E. coli* (Dudley et al. 2006). The regions absent from 8289_1#27 encoded polymerase and UV protection proteins (at 13-15 Kb) and segregation genes *parM* and *stbB* (at 18-20 Kb, along with hypothetical genes).

The longer complete IS*Ecp1*-*bla*$_{\text{CTX-M-15}}$-orf477-*tnpA*-*tnpR*-*bla*$_{\text{TEM}}$ structure was present in 8289_1#3 (C2_7) only, whereas inferred further transposition by IS*Ecp1* meant that 8289_1#34 in the same genetic group (C2_7) had lost it and the *bla*$_{\text{TEM}}$ gene, as did 8289_1#27 (C2_9) even though it had most of the rest of pEK204. 8289_1#35 (C1) had IS*Ecp1* and *bla*$_{\text{TEM}}$ only, unlike 8289_1#24 (C2_8) that had two copies of the *bla*$_{\text{CTX-M-15}}$ gene, three IS*Ecp1* copies in total, and part of the *tnpR*. The conserved regions shared across the clade C set encoded a DNA polymerase V subunit and UV protection and mutation gene (at 17 Kb), a methylase gene (at 22 Kb), and essential maintenance gene and segregation genes (at 27 Kb).

6.3.6 Mobilization of AMR genes driven by ISEcp1, IS26 and IS903D

IS26 had the highest copy number of the three transposons examined here: it had 19 copies in pEC958 (n=8), pEK516 (n=5) and pEK499 (n=6), but none in pEK204 nor any on the NCTC13441 chromosome. Two main IS26 isoforms were present (Figure 6.7): one with nine copies across pEC958 (n=4), pEK516 (n=2) and pEK499 (n=3); and the other with nine copies across pEC958 (n=3), pEK516 (n=3) and pEK499 (n=3); along with a single slightly divergent IS26 in pEC958. Most were 820 bp, though the most divergent one in pEC958 in green was longer due to the insertion of ISEcp1 (1,332 bp), splitting IS26 at its IRs. IS26 was also found in the 3_2_53FAA HMP accessory genome, which had an IRR 4 Kb from a *tnpA* gene, which was 26.5 Kb from another IRR copy.

8289_1#3 had four extensively rearranged plasmid-bourne IS26-like structures, indicating previous activity. The first was a pair of adjacent IRRs, like the second that also had a *tnpA* and a 3' IRR. The third was a single IRR, and the fourth was a *tnpA* 8 Kb from another *tnpA* copy that in turn was 23 Kb from a IRR copy. 8289_1#3 also had a remnant of ISEcp1's 50 bp IRR 42.5 Kb from the main copy, symptomatic of an excised ISEcp1.

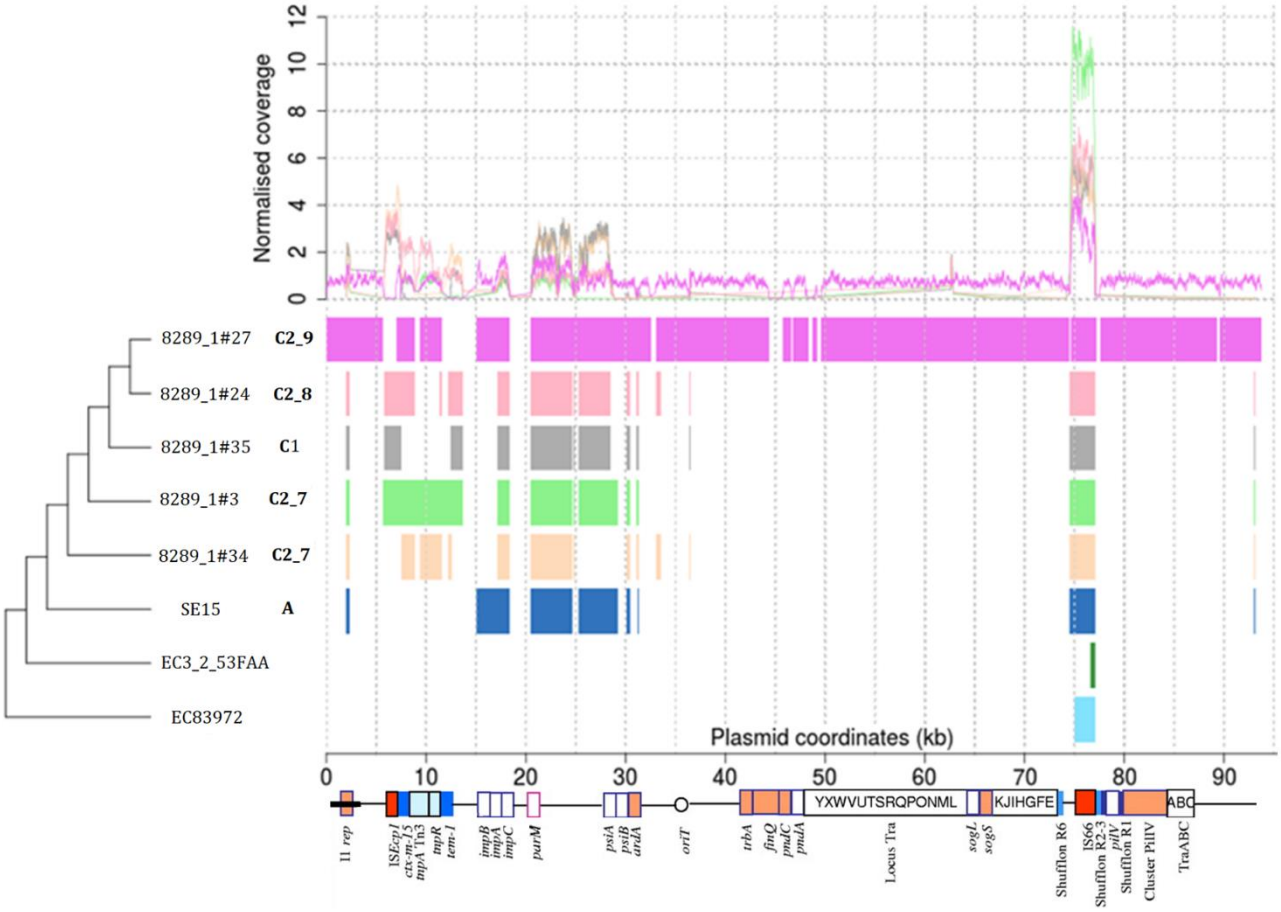


Figure 6.6. Regions of similarity and normalized copy number variation for pEK204 (93,732 bp). A coloured cladogram (left) shows the genetic relationships of the isolates examined where the two most divergent ones (3_2_53FAA in dark green and 83972 in cyan) did not have pEK516, and the commensal SE15 (navy) had only small regions of similarity. The ST131 isolates had small regions of similarity based on BLAST alignments, bar 8289_1#27 that possessed plasmid regions similar to pEK204. The *bla*$_{\text{CTX-M}}$ gene was at 8 Kb, *bla*$_{\text{TEM}}$ was at 13 Kb, followed by mixed conjugation and segregation genes at 36-70 Kb, before a high copy number region at 75-77 Kb encoding an IS66 element and unannotated genes. The annotation is modified from Woodford et al. (2009).

Two main isoforms of ISEcp1 (lengths 1,556-1,655 bp) were also present as single copies in the NCTC13441 chromosome, pEK204, pEK499 and pEK516 (Figure 6.8). The chromosomal and pEK204 ones were closely related and were distinct from the other isoform on the pEK499 and pEK516 plasmids. A single ISEcp1 was located at bases 37,109-38,748 on pEC958 .

In the ST131 clade C isolates, all had at least one ISEcp1 copy. 8289_1#24 was inferred to have undergone multiple arrangements because it's plasmid copy was re-structured with the tnpA gene 93 Kb from an IRR that was 89 Kb 5' from an IRL with another 8 Kb to another IRL copy (Figure 6.9). The chromosomal copy had a duplicated *tnpA* and two IRL copies, presumably due to independent recombination events.

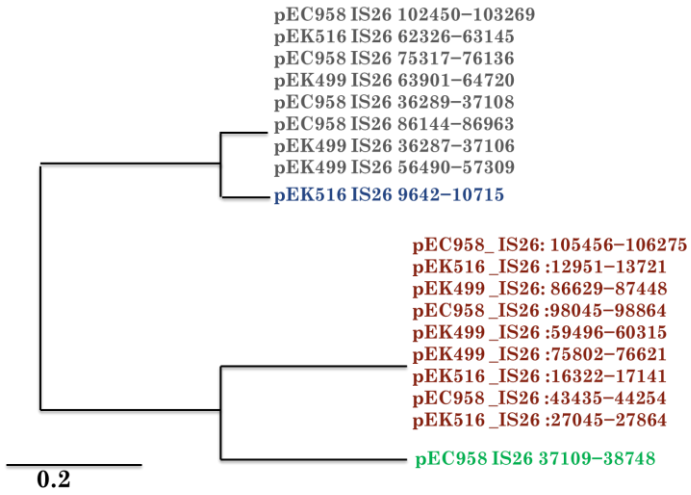


Figure 6.7. Phylogenetic relationship of 19 copies of IS26 element obtained from four ST131 plasmids (pEK499, pEC958, pEK516, pEC958). This showed two main clusters of isoforms. The plasmid coordinates of the copies are shown (start-end).

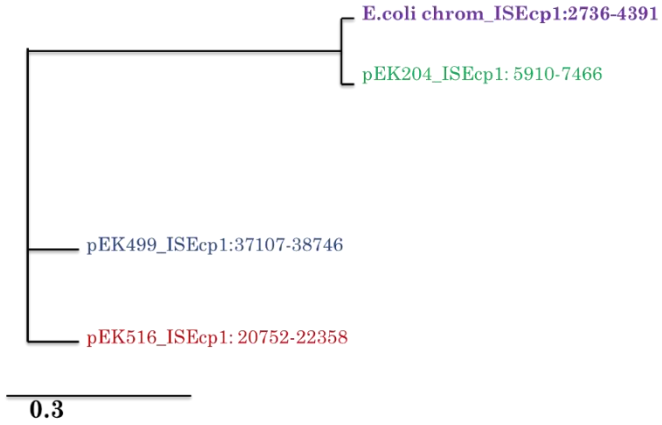


Figure 6.8. Phylogenetic relationship of four ISEcp1 copies in three ST131 plasmids (pEK499, pEK516, pEC958) and the NCTC13441 chromsome. This showed two main clusters of isoforms.

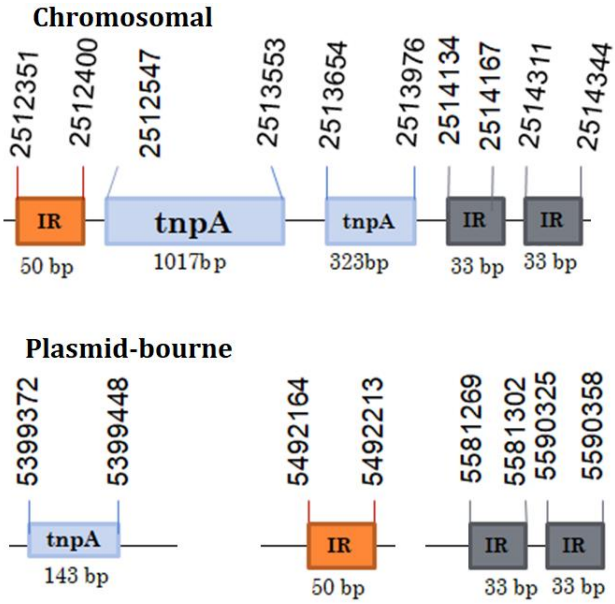


Figure 6.9. ISEcp1 TE structure of 8289_1#24 had two copies, one on the chromosome (top) and one on the plasmid (bottom). Both had been rearranged by multiple recombination events.

6.4 Discussion

There was an extensive shared resistome between ExPEC *E. coli* ST131 compared to non-pathogenic and microbiome *E. coli*, indicating likely gene transfer between commensal and pathogenic bacteria inhabiting the human gut and urinary tract (Ben Zakour et al. 2016). This resistome was based on a set of functionally characterised contigs isolated from preterm infant faeces after antibiotic-resistance profiling, thus providing a direct link between genotype and phenotype indicating resistance to penicillin, chloramphenicol, cephalosporin, tetracycline and monobactam compounds. Asymptomatic *E. coli* HMP 83972 is used for therapeutic urinary bladder colonization in patients because it is protective against super-infection with more virulent strains (Sundén et al. 2006, Sundén et al. 2010) and here had both the fewest AMR genes and has no large plasmids, putatively due to virulence gene loss during adaptation to commensalism (Zdziarski et al. 2010).

Horizontal DNA transfer and subsequent recombination is a common mechanism of genome evolution and adaptation in bacteria generally (Robinson and Enright 2004, Brochet et al. 2008, Chen et al. 2014) and in *E. coli* specifically (Milkman 1997, Cooper 2007, Tenaillon et al. 2010, Didelot et al. 2012, Dixit et al. 2015, Tchesnokova et al. 2019) as well as the human microbiome (Smillie et al. 2011, Lloyd-Price et al. 2017). Notably in *E. coli* generalized transduction via phages may contributed significantly to transfer of smaller DNA segments (Dixit et al. 2015, Didelot et al. 2012) because *E. coli* is not naturally component for DNA uptake via transformation.

There is extensive of *E. coli* gene transfer in the gut and of gene exchange with spreading to the bloodstream (Tamburini et al. 2018). A *bla*$_{\text{CTX-M-1}}$-positive Incl1 plasmid spread between *E. coli* subtypes in the gut of a cystic fibrosis patient (Knudsen et al. 2018). Additionally, a *bla*$_{\text{OXA-48}}$-positive IncL/M-type plasmid from K. pneumoniae was received by *E. coli* (Gottig et al. 2015, Willemsen et al. 2016). And a range of sulphonamide- and ampicillin-resistance genes (*sul2*, *bla*$_{\text{TEM-1b}}$) were transferred on a pNK29 plasmid within *E. coli* subtypes in the human gut (Trobos et al. 2009, Karami et al. 2007). MGEs also play a role in this: for instance, IS*Kpn19* initiated replicative transposition at a known IR between IS*3000* and IS*Aba125* elements in *E. coli* producing *bla*$_{\text{NDM-5}}$ (New Delhi Metallo-

beta-lactamase) (Xie et al. 2018). This gene can be encoded by IncF2 plasmids and uses IS*26*-mediated recombination for mobilisation and rearrangement (Pitart C et al 2015).

The ST131 clade C isolates investigated here had notably higher levels of AMR genes than the HMP samples and commensal SE15. Some of these were transferred from other species, such as *bla*$_{OXY}$ from *Klebsiella oxytoca* and *bla*$_{CTX-M}$ from *Kluyvera* (Lartigue et al. 2006). These clade C genome assemblies possessed highly conserved core genomes, but nonetheless showed numerous differences in AMR gene content, which correlated with their putative mobilizers, MGE transposition, plasmid recombination and conjugation. Reference NCTC13441 had numerous *bla*$_{TEM}$ or *bla*$_{OXY}$ β-lactamase genes, contrasting with EC958's greater concentration of cephalosporin-resistance *bla*$_{CMY}$ (AmpC) genes, even though these isolates' core genomes had high identity (Decano et al. 2019, Chapter 4).

This was also illustrated by the acquisition of an IncI pEK204-like plasmid in one clade C isolate (8289_1#27), which contained a type IV pilus biosynthesis locus (*pil*) allowing better epithelial cell adhesion and superior biofilm formation (Dudley et al. 2006). Given that clade C's fitness advantage has been tightly associated with the fimH30 fimbrial adhesin profile (Petty et al. 2014), the emergence of additional conjugation machinery on a conjugative plasmid in ST131 is a concern.

The rearranged structures found for IS*Ecp1* and IS*26* indicated that chromosomally or plasmid-bourne AMR genes in transposable elements can vary in ST131 further. This has implications for clinical typing if this relies on archaic techniques like MLST or rST that will have limited information on AMR phenotypes. This was illustrated by *bla*$_{CTX-M-14}$-positive 8289_1#35 from subclade C1 and *bla*$_{CTX-M-14/15}$-positive 8289_1#24 from subclade C2, which had duplicated *tra* regions and a higher diversity of distinct ESBL genes compared to the other ST131 clade C isolates, which may be driven by transposition-mediated recombination in 8289_1#24.

This chapter emphasized the potential relevance of accessory genome screening with scope for developing diagnostics tracking plasmid and MGE types in addition to those existing for AMR genes (Durrant et al. 2019). Tracking transmission patterns within

closely related infection outbreaks may be better informed by linking plasmid replicon data with transposon types in addition to AMR profiles.

In addition, given that the jumbled nature of *E. coli* genomes is hypothesized to be a product of type I restriction/modification system enzyme fragmentation of horizontally transferred DNA (McKane and Milkman 1995, Loenen et al. 2014, Dixit et al. 2015), further work can explore further conjugated DNA processing in the cell, how recombinant DNA is regulated, and what the fitness effects result.

6.5 References

Adams-Sapper S, Diep BA, Perdreau-Remington F, Riley LW. Clonal composition and community clustering of drug-susceptible and -resistant *Escherichia coli* isolates from bloodstream infections. Antimicrob Agents Chemother. 2013 57(1):490-7. doi: 10.1128/AAC.01025-12

Bäumler AJ, Sperandio V. Interactions between the microbiota and pathogenic bacteria in the gut. Nature 2016 535(7610):85-93. doi: 10.1038/nature18849

Ben Zakour NL, Alsheikh-Hussain AS, Ashcroft MM, Khanh Nhu NT, Roberts LW, Stanton-Cook M, Schembri MA, Beatson SA. Sequential Acquisition of Virulence and Fluoroquinolone Resistance Has Shaped the Evolution of *Escherichia coli* ST131. MBio. 2016 7(2):e00347-16. doi: 10.1128/mBio.00347-16

Brochet M, Rusniok C, Couvé E, Dramsi S, Poyart C, Trieu-Cuot P, Kunst F, Glaser P. Shaping a bacterial genome by large chromosomal replacements, the evolutionary history of Streptococcus agalactiae. Proc Natl Acad Sci U S A. 2008 105(41):15961-6. doi: 10.1073/pnas.0803654105

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

Carattoli A, Zankari E, García-Fernández A, Voldby Larsen M, Lund O, Villa L, Møller Aarestrup F, Hasman H. *In silico* detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing. Antimicrob Agents Chemother. 2014 58(7):3895-903. doi: 10.1128/AAC.02412-14

Carver T, Harris SR, Berriman M, Parkhill J, McQuillan JA. Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. Bioinformatics. 2012 28(4):464-9. doi: 10.1093/bioinformatics/btr703

Carver, T.J., Rutherford, K.M., Berriman, M., Rajandream, M.-A., Barrell, B.G., Parkhill, J., 2005. ACT: the Artemis Comparison Tool. Bioinformatics 21, 3422–3423. https://doi.org/10.1093/bioinformatics/bti553

Chen L, Mathema B, Pitout JD, DeLeo FR, Kreiswirth BN. Epidemic Klebsiella pneumoniae ST258 is a hybrid strain. MBio. 2014 5(3):e01355-14. doi: 10.1128/mBio.01355-14

Clemente JC, Ursell LK, Parfrey LW, Knight R. The impact of the gut microbiota on human health: an integrative view. Cell. 2012 148(6):1258-70. doi: 10.1016/j.cell.2012.01.035

Conway T, Cohen PS. Commensal and Pathogenic *Escherichia coli* Metabolism in the Gut. Microbiol Spectr. 2015 Jun;3(3). doi: 10.1128/microbiolspec.MBP-0006-2014

Cooper TF. Recombination speeds adaptation by reducing competition between beneficial mutations in populations of Escherichia coli. PLoS Biol. 2007 5(9):e225

Dhanji H, Doumith M, Hope R, Livermore DM, Woodford N. ISEcp1-mediated transposition of linked blaCTX-M-3 and blaTEM-1b from the IncI1 plasmid pEK204 found in clinical isolates of *Escherichia coli* from Belfast, UK. J Antimicrob Chemother. 2011 66(10):2263-5. doi: 10.1093/jac/dkr310

Didelot X, Méric G, Falush D, Darling AE. Impact of homologous and non-homologous recombination in the genomic evolution of Escherichia coli. BMC Genomics. 2012 13:256. doi: 10.1186/1471-2164-13-256

Dixit PD, Pang TY, Studier FW, Maslov S. Recombinant transfer in the basic genome of Escherichia coli. Proc Natl Acad Sci U S A. 2015 112(29):9070-5. doi: 10.1073/pnas.1510839112

Dudley EG, Abe C, Ghigo JM, Latour-Lambert P, Hormazabal JC, Nataro JP. An IncI1 plasmid contributes to the adherence of the atypical enteroaggregative *Escherichia*

coli strain C1096 to cultured cells and abiotic surfaces. Infect Immun. 2006 74(4):2102-14

Durrant MG, Li MM, Siranosian B, Bhatt AS. Mobile genetic element insertions drive antibiotic resistance across pathogens. BioRxiv 2019 doi: http://dx.doi.org/10.1101/527788.

Forde BM, Ben Zakour NL, Stanton-Cook M, Phan MD, Totsika M, Peters KM, Chan KG, Schembri MA, Upton M, Beatson SA. The complete genome sequence of Escherichia coli EC958: a high quality reference sequence for the globally disseminated multidrug resistant *E. coli* O25b:H4-ST131 clone. PLoS One. 2014 9(8):e104400. doi: 10.1371/journal.pone.0104400

Frost LS, Leplae R, Summers AO, Toussaint A. Mobile genetic elements: the agents of open source evolution. Nat Rev Microbiol. 2005 3(9):722-32

Gibson MK, Wang B, Ahmadi S, Burnham CA, Tarr PI, Warner BB, Dantas G. Developmental dynamics of the preterm infant gut microbiota and antibiotic resistome. Nat Microbiol. 2016 1:16024. doi: 10.1038/nmicrobiol.2016.24

Göttig S, Gruber TM, Stecher B, Wichelhaus TA, Kempf VA. In vivo horizontal gene transfer of the carbapenemase OXA-48 during a nosocomial outbreak. Clin Infect Dis. 2015 60(12):1808-15. doi: 10.1093/cid/civ191

Gupta SK, Padmanabhan BR, Diene SM, Lopez-Rojas R, Kempf M, Landraud L, Rolain JM. ARG-ANNOT, a new bioinformatic tool to discover antibiotic resistance genes in bacterial genomes. Antimicrob Agents Chemother. 2014 58(1):212-20. doi: 10.1128/AAC.01310-13

Hinnebusch J, Tilly K. Linear plasmids and chromosomes in bacteria. Mol Microbiol. 1993 10(5):917-22.

Hooks KB, O'Malley MA. Dysbiosis and Its Discontents. mBio 2017 8(5):e01492-17 DOI: 10.1128/mBio.01492-17

Hudault S, Guignot J, Servin AL. *Escherichia coli* strains colonising the gastrointestinal tract protect germfree mice against *Salmonella typhimurium* infection. Gut. 2001 49(1):47-55

Jia B, Raphenya AR, Alcock B, Waglechner N, Guo P, Tsang KK, Lago BA, Dave BM, Pereira S, Sharma AN, Doshi S, Courtot M, Lo R, Williams LE, Frye JG, Elsayegh T, Sardar D, Westman EL, Pawlowski AC, Johnson TA, Brinkman FS, Wright GD, McArthur AG. CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. Nucleic Acids Res. 2017 45(D1):D566-D573

Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, Pesseat S, Quinn AF, Sangrador-Vegas A, Scheremetjew M, Yong SY, Lopez R, Hunter S. InterProScan 5: genome-scale protein function classification. Bioinformatics. 2014 30(9):1236-40. doi: 10.1093/bioinformatics/btu031

Karami N, Martner A, Enne VI, Swerkersson S, Adlerberth I, Wold AE. Transfer of an ampicillin resistance gene between two Escherichia coli strains in the bowel microbiota of an infant treated with antibiotics. J Antimicrob Chemother. 2007 60(5):1142-5

Knudsen PK, Gammelsrud KW, Alfsnes K, Steinbakk M, Abrahamsen TG, Müller F, Bohlin J. Transfer of a bla (CTX-M-1)-carrying plasmid between different Escherichia coli strains within the human gut explored by whole genome sequencing analyses. Sci Rep. 2018 8(1):280. doi: 10.1038/s41598-017-18659-2

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

Lartigue MF, Poirel L, Aubert D, Nordmann P. In vitro analysis of ISEcp1B-mediated mobilization of naturally occurring beta-lactamase gene blaCTX-M of Kluyvera ascorbata. Antimicrob Agents Chemother. 2006 50(4):1282-6

Livermore DM, Canton R, Gniadkowski M, Nordmann P, Rossolini GM, Arlet G, Ayala J, Coque TM, Kern-Zdanowicz I, Luzzaro F, Poirel L, Woodford N. CTX-M: changing the face of ESBLs in Europe. J Antimicrob Chemother. 2007 59(2):165-74

Lloyd-Price J, Mahurkar A, Rahnavard G, Crabtree J, Orvis J, Hall AB, Brady A, Creasy HH, McCracken C, Giglio MG, McDonald D, Franzosa EA, Knight R, White O, Huttenhower C. Strains, functions and dynamics in the expanded Human Microbiome Project. Nature 2017 550(7674):61-66. doi: 10.1038/nature23889

Loenen WA, Dryden DT, Raleigh EA, Wilson GG, Murray NE. Highlights of the DNA cutters: a short history of the restriction enzymes. Nucleic Acids Res. 2014 42(1):3-19. doi: 10.1093/nar/gkt990

McArthur AG, Waglechner N, Nizam F, Yan A, Azad MA, Baylay AJ, Bhullar K, Canova MJ, De Pascale G, Ejim L, Kalan L, King AM, Koteva K, Morar M, Mulvey MR, O'Brien JS, Pawlowski AC, Piddock LJ, Spanogiannopoulos P, Sutherland AD, Tang I, Taylor PL, Thaker M, Wang W, Yan M, Yu T, Wright GD. The comprehensive antibiotic resistance database. Antimicrob Agents Chemother. 2013 57(7):3348-57. doi: 10.1128/AAC.00419-13

McKane M, Milkman R. Transduction, restriction and recombination patterns in Escherichia coli. Genetics. 1995 139(1):35-43
Milkman R. Recombination and population structure in Escherichia coli. Genetics. 1997 146(3):745-50

Naito M, Pawlowska TE. The role of mobile genetic elements in evolutionary longevity of heritable endobacteria. Mob Genet Elements. 2015 6(1):e1136375

Nicolas-Chanoine MH, Bertrand X, Madec JY. Escherichia coli ST131, an intriguing clonal group. Clin Microbiol Rev. 2014 27(3):543-74. doi: 10.1128/CMR.00125-13

Nicolas-Chanoine MH, Petitjean M, Mora A, Mayer N, Lavigne JP, Boulet O, Leflon-Guibout V, Blanco J, Hocquet D. The ST131 *Escherichia coli* H22 subclone from human intestinal microbiota: Comparison of genomic and phenotypic traits with those of the globally successful H30 subclone. BMC Microbiol. 2017 17(1):71. doi: 10.1186/s12866-017-0984-8

Olesen SW, Alm EJ. Dysbiosis is not an answer. Nat Microbiol. 2016 1:16228

Penders J, Thijs C, Vink C, Stelma FF, Snijders B, Kummeling I, van den Brandt PA, Stobberingh EE. Factors influencing the composition of the intestinal microbiota in early infancy. Pediatrics. 2006 118(2):511-21

Petty NK, Ben Zakour NL, 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. Global dissemination of a multidrug resistant *Escherichia coli* clone. Proc Natl Acad Sci U S A. 2014 111(15):5694-9. doi: 10.1073/pnas.1322678111

Phan MD, Forde BM, Peters KM, Sarkar S, Hancock S, Stanton-Cook M, Ben Zakour NL, Upton M, Beatson SA, Schembri MA. Molecular characterization of a multidrug resistance IncF plasmid from the globally disseminated *Escherichia coli* ST131 clone. PLoS One. 2015 10(4):e0122369. doi: 10.1371/journal.pone.0122369

Pitart C, Solé M, Roca I, Román A, Moreno A, Vila J, Marco F. Molecular characterization of blaNDM-5 carried on an IncFII plasmid in an Escherichia coli isolate from a nontraveler patient in Spain. Antimicrob Agents Chemother. 2015 59(1):659-62. doi: 10.1128/AAC.04040-14

Poirel L, Gniadkowski M, Nordmann P. Biochemical analysis of the ceftazidime-hydrolysing extended-spectrum beta-lactamase CTX-M-15 and of its structurally related beta-lactamase CTX-M-3. J Antimicrob Chemother. 2002 50(6):1031-4

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. The epidemic of extended-spectrum-β-lactamase-producing *Escherichia coli* ST131 is driven by a single highly pathogenic subclone, H30-Rx. MBio. 2013 4(6):e00377-13. doi: 10.1128/mBio.00377-13.

Robinson DA, Enright MC. Evolution of Staphylococcus aureus by large chromosomal replacements. J Bacteriol. 2004 186(4):1060-4

Rowe WPM, Winn MD. Indexed variation graphs for efficient and accurate resistome profiling. Bioinformatics. 2018 34(21):3601-3608. doi: 10.1093/bioinformatics/bty387

Rudick CN, Taylor AK, Yaggie RE, Schaeffer AJ, Klumpp DJ. Asymptomatic bacteriuria *Escherichia coli* are live biotherapeutics for UTI. PLoS One. 2014 9(11):e109321. doi: 10.1371/journal.pone.0109321

Runyen-Janecky LJ, Hong M, Payne SM. The virulence plasmid-encoded impCAB operon enhances survival and induced mutagenesis in *Shigella flexneri* after exposure to UV radiation. Infect Immun. 1999 67(3):1415-23

Shintani M, Sanchez ZK, Kimbara K. Genomics of microbial plasmids: classification and identification based on replication and transfer systems and host taxonomy. Front Microbiol. 2015 6:242. doi: 10.3389/fmicb.2015.00242.

Smet A, Van Nieuwerburgh F, Vandekerckhove TT, Martel A, Deforce D, Butaye P, Haesebrouck F. Complete nucleotide sequence of CTX-M-15-plasmids from clinical *Escherichia coli* isolates: insertional events of transposons and insertion sequences. PLoS One. 2010 5(6):e11202. doi: 10.1371/journal.pone.0011202

Smillie CS, Smith MB, Friedman J, Cordero OX, David LA, Alm EJ. Ecology drives a global network of gene exchange connecting the human microbiome. Nature 2011 480(7376):241-4. doi: 10.1038/nature10571

Sundén F, Håkansson L, Ljunggren E, Wullt B. Bacterial interference--is deliberate colonization with Escherichia coli 83972 an alternative treatment for patients with recurrent urinary tract infection? Int J Antimicrob Agents. 2006 28 Suppl 1:S26-9.

Sundén F, Håkansson L, Ljunggren E, Wullt B. Escherichia coli 83972 bacteriuria protects against recurrent lower urinary tract infections in patients with incomplete bladder emptying. J Urol. 2010 184(1):179-85. doi: 10.1016/j.juro.2010.03.024

Tamburini FB, Andermann TM, Tkachenko E, Senchyna F, Banaei N, Bhatt AS. Precision identification of diverse bloodstream pathogens in the gut microbiome. Nat Med. 2018 24(12):1809-1814. doi: 10.1038/s41591-018-0202-8

Tchesnokova V, Billig M, Chattopadhyay S, Linardopoulou E, Aprikian P, Roberts PL, Skrivankova V, Johnston B, Gileva A, Igusheva I, Toland A, Riddell K, Rogers P, Qin X, Butler-Wu S, Cookson BT, Fang FC, Kahl B, Price LB, Weissman SJ, Limaye A, Scholes D, Johnson JR, Sokurenko EV. Predictive diagnostics for *Escherichia coli* infections based on the clonal association of antimicrobial resistance and clinical outcome. J Clin Microbiol. 2013 Sep;51(9):2991-9. doi: 10.1128/JCM.00984-13

Tchesnokova V, Radey M, Chattopadhyay S, Larson L, Weaver JL, Kisiela D, Sokurenko EV. Pandemic fluoroquinolone resistant Escherichia coli clone ST1193 emerged via simultaneous homologous recombinations in 11 gene loci. PNAS 2019 https://doi.org/10.1073/pnas.1903002116

Tenaillon O, Skurnik D, Picard B, Denamur E. The population genetics of commensal Escherichia coli. Nat Rev Microbiol. 2010 8(3):207-17. doi: 10.1038/nrmicro2298

Thursby E, Juge N. Introduction to the human gut microbiota. Biochem J. 2017 474(11):1823-1836. doi: 10.1042/BCJ20160510

Toh H, Oshima K, Toyoda A, Ogura Y, Ooka T, Sasamoto H, Park SH, Iyoda S, Kurokawa K, Morita H, Itoh K, Taylor TD, Hayashi T, Hattori M. Complete genome sequence of

the wild-type commensal *Escherichia coli* strain SE15, belonging to phylogenetic group B2. J Bacteriol. 2010 192(4):1165-6. doi: 10.1128/JB.01543-09

Trobos M, Lester CH, Olsen JE, Frimodt-Møller N, Hammerum AM. Natural transfer of sulphonamide and ampicillin resistance between Escherichia coli residing in the human intestine. J Antimicrob Chemother. 2009 63(1):80-6. doi: 10.1093/jac/dkn437

Willemsen I, van Esser J, Kluytmans-van den Bergh M, Zhou K, Rossen JW, Verhulst C, Verduin K, Kluytmans J. Retrospective identification of a previously undetected clinical case of OXA-48-producing K. pneumoniae and E. coli: the importance of adequate detection guidelines. Infection. 2016 44(1):107-10. doi: 10.1007/s15010-015-0805-7

Woodford N, Carattoli A, Karisik E, Underwood A, Ellington MJ, Livermore DM. Complete nucleotide sequences of plasmids pEK204, pEK499, and pEK516, encoding CTX-M enzymes in three major *Escherichia coli* lineages from the United Kingdom, all belonging to the international O25:H4-ST131 clone. Antimicrob Agents Chemother. 2009 53(10):4472-82. doi: 10.1128/AAC.00688-09.

Xie M, Li R, Liu Z, Chan EWC, Chen S. Recombination of plasmids in a carbapenem-resistant NDM-5-producing clinical Escherichia coli isolate. J Antimicrob Chemother. 2018 73(5):1230-1234. doi: 10.1093/jac/dkx540

Zankari E, Hasman H, Cosentino S, Vestergaard M, Rasmussen S, Lund O, Aarestrup FM, Larsen MV. Identification of acquired antimicrobial resistance genes. J Antimicrob Chemother. 2012 67(11):2640-4. doi: 10.1093/jac/dks261

Zdziarski J, Brzuszkiewicz E, Wullt B, Liesegang H, Biran D, Voigt B, Grönberg-Hernandez J, Ragnarsdottir B, Hecker M, Ron EZ, Daniel R, Gottschalk G, Hacker J, Svanborg C, Dobrindt U. Host imprints on bacterial genomes--rapid, divergent evolution in individual patients. PLoS Pathog. 2010 6(8):e1001078. doi: 10.1371/journal.ppat.1001078

Chapter 7: Thesis Summary, Conclusions and Future Work

7.1 Thesis Summary

In chapter 1, I conducted a comprehensive background study on ST131 evolution and the strategies used for phylogenomic analyses of ST131 populations. This was followed by a pilot study involving 100 ST131 strains in Chapter 2 to optimise genomic methods that were planned to apply to the subsequent analyses in the project.

In chapter 3, a UTI outbreak in Irish nursing homes was investigated by performing a deep evaluation of the samples' evolutionary history. Using short read data, we attempted to identify the origin and date of emergence of the strains that form clade with genetetic characteristics different from the rest of previously samples global ST131 isolates.

In chapter 4, the architecture of pathogenic ST131 carrying *bla* genes was resolved using long read sequences. In association with my collaborators at the Sanger Institute, I applied new high-resolution Oxford Nanopore DNA sequencing technology to six ST131 samples from infected patients and compared the output to Illumina short reads. This was to tackle infections more effectively by improving our understanding of what plasmids are being exchanged and their exact antibiotic resistance gene contents. A combination of methods showed that drug-resistance genes on plasmids were highly mobile in this chapter because they can jump into ST131's chromosomes. We found that the plasmids are very elastic and undergo extensive rearrangements even in closely related samples. We also confirmed the chromosomal insertion of *bla*CTX-M-15 gene and its flanking elements. This application of DNA sequencing technologies illustrated at a new level the highly dynamic nature of ST131 genomes.

In chapter 5, the largest whole genome collection of ST131 by far was analysed by identifying their origin, evolution and population structure. We provided a deep

resolution of the epidemiology and genomic context of key ESBL genes. We also showed that the core genome is highly stable in contrast to its hypervariable accessory genome.

In chapter 6, we quantified the degree of AMR gene sharing between gut *E. coli* and ST131. We also showed that ST131's key plasmids that play a major role in AMR gene transfer are highly variable within ST131 clade C. The results in this chapter indicate that resolving the structures, copy numbers and locations of key ST131 provide opportunities for better surveillance of AMR in ST131 populations.

Overall, this thesis provided a robust evidence of inter-clonal diversity using whole genomes and emphasized that long read assembly resolved plasmid transposition, chromosomal insertion of AMR genes and the complete genome architecture of ST131. Moreover, using the strategies I developed throughout my PhD study, I was able to resolved the population structure of ST131 using the largest collection thus far.

7.2 Avenues for Future Work

This project can be further extended by applying the methods developed for clinical metagenomics and by adapting culture-independent sequencing approaches of clinical samples for rapid diagnostics, informing antibiotic treatment, and studies of bacterial evolution. This entails the application of [1] culture-independent sequence-based approach, [2] the use of Oxford nanopore long read sequencing across a larger sample size (as PacBio is 20-fold more expensive) (Kim et al. 2019), [3] bacterial genome assembly where taxonomic binning is not required (Nicholls et al. 2019) and [4] determining antibiotic sensitivity of all 4,000 or more ST131 samples.

Future work could examine what accessory gene (Ben Zakour et al. 2016) and core gene regulatory (McNally et al. 2016) changes may result in new clonal expansions in the descendant lineages and identify molecular adaptations to gastrointestinal or urinary tract environments (McNally et al. 2019). It should also reducing ST131 sampling bias by expanding numbers of non-human isolates and diversity of geographic region sampled, which can help resolve potential sources of *E. coli*'s ESBL genes like *bla*$_{\text{CTX-M-15}}$, for which

there was no evidence of retail meat (Randall et al. 2016) or livestock (Ludden et al. 2019) as sources for BSIs thus far, though transfer of bacteria may occur (Roer et al. 2019).

Another aspect where the results of this thesis could be useful is drug discovery. I have drafted a project proposal to collaborate with Dr. Zamin Iqbal's group and Dr. Nassos Typas to explore this area.

The potential of using antibacterial drug combinations to enhance drug effectivity has been recently modelled using Gram-negative *Escherichia coli*, *Salmonella enterica* serovar Typhimurium and Pseudomonas *aeruginosa* (Brochado et al. 2018). In this proposed project, I plan to collaborate with the group that pioneered these experiments and apply their strategies to identify the optimal drug combinations and antibiotic effectivity against ST131 uropathogens. The work is divided into two stages (Figure 7.1) under a strict timeline to achieve optimal delivery of results. The first stage, which involves sample collection to antibiotic resistance profiling will be performed at EMBL Heidelberg while the 2nd stage that comprises sample sequencing and further downstream data analyses will be performed at EBI in Hinxton, UK.

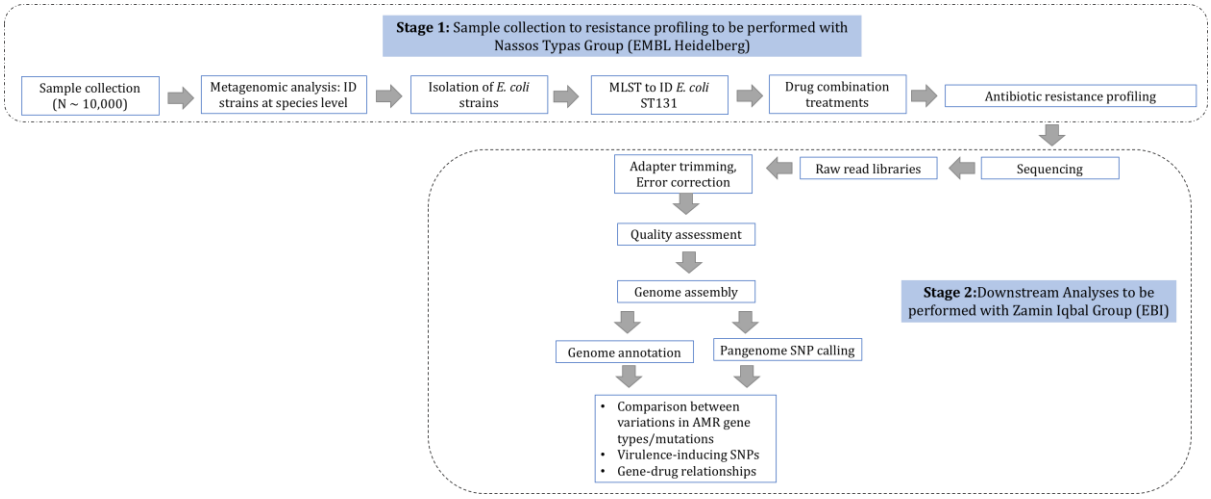


Figure 7.1. Proposed workflow of the upstream (Stage 1) and downstream (Stage 2) processes for investigating the gene-drug (genotype-phenotype) relationships in *E. coli* sequence type (ST) 131. The 1st stage will be performed with Dr. Nassos Typas Group and will involve sample collection to antibiotic resistance profiling of strains that were both untreated and treated (wild type) with drug (antibiotics) combinations. The 2nd

stage that includes sequencing to downstream data analyses will be implemented with Dr. Zamin Iqbal's Group. Clinical and environmental bacterial strains will be collected from clinical and environmental sources. These isolates will be classified at the species level using metagenomics strategies. Multi-locus sequence typing of all recovered *E. coli* will then be performed to distinguish *E. coli* ST131 from the other sequence types followed by the treatment of several drug combinations. Treated and non-treated samples will be sequenced using Illumina HiSeq; the resulting raw read libraries will be rid of sequencing the adapters and their base errors corrected. All "clean" read libraries that will pass the quality assessment will then be assembled. Draft genomes will be annotated to determine the variation in the AMR gene type/mutations as well as SNPs in virulence-associated elements in both the wild type and the treated isolates.

7.3 Conclusions and Final thoughts

My entire PhD work strongly advocates that improved treatments of antibiotic-resistant bacterial infections require massive investment in high-throughput genomics. The sustained use of ciprofloxacin and third generation cephalosporins will continue to enrich for virulent clades such as ST131 Clade C2 lineages. The collective data and results presented in this thesis highlight the global need to reduce the selective pressure from these antimicrobials. The diversity of ST131 lineages and resistance elements indicates a need for surveillance strategies to identify ST131 subclones, plasmids and transposable elements. The characterisation of those specific properties that make specific lineages successful in particular contexts remains one of the key challenges in understanding the dynamics of emergence and spread of new variants of common bacterial species. Focused attention to successful strains could help to explore these interactions and control the epidemic of antimicrobial resistance.

7.4 References

Charalampous, T., Kay, G.L., Richardson, H., Aydin, A., Baldan, R., Jeanes, C., Rae, D., Grundy, S., Turner, D.J., Wain, J., Leggett, R.M., Livermore, D.M., O'Grady, J., 2019. Nanopore metagenomics enables rapid clinical diagnosis of bacterial lower respiratory infection. Nature Biotechnology 37, 783. https://doi.org/10.1038/s41587-019-0156-5.

Kim HS, Jeon S, Kim C, Kim YK, Cho YS, Kim J, Blazyte A, Manica A, Lee S, Bhak J. Chromosome-scale assembly comparison of the Korean Reference Genome KOREF from PromethION and PacBio with Hi-C mapping information. BioRxiv 2019 doi: https://doi.org/10.1101/674804.

Nicholls SM, Quick JC, Tang S, Loman NJ. Ultra-deep, long-read nanopore sequencing of mock microbial community standards. Gigascience 2019 8(5): giz043 doi: 10.1093/gigascience/giz043.

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.

Randall, L.P., Lodge, M.P., Elviss, N.C., Lemma, F.L., Hopkins, K.L., Teale, C.J., Woodford, N., 2016. 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. 241, 283–290.

Ludden, C., Raven, K.E., Jamrozy, D., Gouliouris, T., Blane, B., Coll, F., et al. (2019) One health genomic surveillance of Escherichia coli demonstrates distinct lineages and mobile genetic elements in isolates fromhumans versus livestock. MBio 10, e02693-18. http://doi.org/10.1128/mBio.02693-18.

Roer L, Hansen F, Thomsen MCF, Knudsen JD, Hansen DS, Wang M, Samulioniené J, Justesen US, Røder BL, Schumacher H, Østergaard C, Andersen LP, Dzajic E, Søndergaard

TS, Stegger M, Hammerum AM, Hasman H. 2017. WGS-based surveillance of third-generation cephalosporin-resistant Escherichia coli from bloodstream infections in Denmark. J Antimicrob Chemother 72:1922–1929. https://doi.org/10.1093/jac/dkx092.

Brochado, A. R. et al. Species-specific activity of antibacterial drug combinations. Nature 559, 259–263 (2018).

Mateus A, Bobonis J, Kurzawa N, Stein F, Helm D, Hevler J, Typas A, Savitski MM. (2018). Thermal proteome profiling in bacteria: probing protein state in vivo. Mol. Syst. Biol. 14(7) doi: 10.15252/msb.20188242.

Appendix

The phylogenetic trees in Chapters 3, 4 and 5 that were generated using iTol v.4 (https://itol.embl.de/) can be accessed via my account username **"aedecano"**.

The scripts for processing the data in this thesis are publicly available on my Github (https://github.com/) account, "**aedecano"**.

The Supplementary information i.e Supplementary tables of this thesis can be viewed and downloaded from my Figshare account, **https://figshare.com/authors/Arun_Decano/5793341**. The files are labelled with "Ch" and the relevant chapter number they belong to (e.g. Ch2 for the data belonging to Chapter 2).