Applications of whole genome sequencing to understanding the mechanisms, evolution and transmission of antibiotic resistance in *Escherichia coli* and *Klebsiella pneumoniae*

Nicole Stoesser

Brasenose College

Nuffield Department of Medicine/Medical Sciences Doctoral Training Centre

A thesis submitted for the degree of Doctor of Philosophy

Michaelmas Term 2014

ABSTRACT

Whole genome sequencing (WGS) has transformed molecular infectious diseases epidemiology in the last five years, and represents a high resolution means by which to catalogue the genetic content and variation in bacterial pathogens.

This thesis utilises WGS to enhance our understanding of antimicrobial resistance in two clinically important members of the *Enterobacteriaceae* family of bacteria, namely *Escherichia coli* and *Klebsiella pneumoniae*. These organisms cause a range of clinical infections globally, and are increasing in incidence. The rapid emergence of multi-drug resistance in association with infections caused by them represents a major threat to the effective management of a range of clinical conditions.

The reliability of sequencing and bioinformatic methods in the analysis of *E. coli* and *K. pneumoniae* sequence data is assessed in chapter 4, and provides a context for the subsequent study chapters, investigating resistance genotype prediction, outbreak epidemiology in two different contexts, and population structure of an important global drug-resistant *E. coli* lineage, ST131 (5-8). In these, the advantages (and limitations) of short-read, high-throughput, WGS in defining resistance gene content, associated mobile genetic elements and host bacterial strains, and the relationships between them, are discussed. The overarching conclusion is that the dynamic between all the components of the genetic hierarchy involved in the transmission of important antimicrobial resistance elements is extremely complicated, and encompasses almost every imaginable scenario. Complete/near-complete assessment of the genetic content

of both chromosomal and episomal components will be a prerequisite to understanding the evolution and spread of antimicrobial resistance in these organisms.

TABLE OF CONTENTS

PREFACE

1.1. ACKNOWLEDGEMENTS

I would like to profoundly thank the many colleagues who have made the work in this thesis possible, and have been the main reason my PhD years have been both hugely educational and fun. I am very grateful for the support and mentorship of my three supervisors, Derrick Crook, Tim Peto and Peter Donnelly in this project, and to Sarah Walker, who acts as an "unofficial" supervisor to many, including me, and is an inspiration to us all. I would also like to give enormous thanks to my family and friends, who are an unstinting source of constant and patient support.

1.2. FUNDING

The work in this thesis was supported by a Wellcome-Trust/University of Oxford Medical Sciences Doctoral Training Centre Doctoral Research Fellowship (099423/Z/12/Z; start date 01/Apr/2012).

Additional funding support was provided by the National Institute for Health Research (NIHR) Biomedical Research Centre Infection Theme funds; the Health Innovation Challenge Fund (HICF-T5-358); the UKCRC Modernising Medical Microbiology Consortium, funded under the UKCRC Translational Infection Research Initiative supported by the Medical Research Council, Biotechnology and Biological Sciences Research Council and the National Institute for Health Research on behalf of the Department of Health Grant (Grant G0800778) and the Wellcome Trust (Grant 087646/Z/08/Z).

Research and sample collection in South-East Asia was supported as part of Wellcome Trust Major Overseas Programme funding for the Mahidol Oxford Tropical Medicine Research Unit, based in Bangkok, and its affiliated units in Siem Reap, Cambodia; Mae Sot, Thailand; and Vientiane, Lao People's Democratic Republic (Laos).

1.3. DECLARATION OF CONTRIBUTION TO THESIS

I, Nicole Stoesser, designed and conducted all the analyses presented in this thesis, with the support of my supervisors and colleagues.

Specific assistance from others in relation to work in this thesis is outlined below.

1.4. ATTRIBUTIONS

1.4.1. CHAPTER 4 – "Thesis methods – Evaluation"

The Sequencing Hub at the Wellcome Trust Centre for Human Genetics (WTCHG) undertook library preparation and sequencing on the Illumina HiSeq. All raw read data were routinely processed through the Crook research group's bioinformatic pipeline, as for all sequenced samples. I designed the studies and conducted the laboratory work, DNA extractions, *de novo* assemblies by non-pipeline methods (SPAdes, MaSuRCa, A5, and unscaffolded Velvet) and comparative analysis of results.

1.4.2. CHAPTER 5 – "Resistance gene identification in *Escherichia coli* and *Klebsiella pneumoniae*, and prediction of antimicrobial susceptibility phenotype from whole genome sequencing data"

I researched the area and designed the study. I created the reference resistance gene database. I selected strains for the study on the basis of a data extract of all *E. coli* and *K. pneumoniae* bloodstream infections supplied by my colleague Sarah Walker, retrieved these isolates from the diagnostic laboratory storage facility, sub-cultured and then extracted DNA from them. Phoenix-based phenotyping was done partly by my diagnostic microbiology laboratory colleagues and partly by me; I carried out all

agar-based phenotyping to investigate isolates with genotype-phenotype discrepancies. The WTCHG Sequencing Centre performed sequencing as previously described. Sequence data were processed through the Crook group's pipeline as previously described. Resistance gene identification and classification were carried out by me, by using and modifying a python script that partly automated the BLASTn algorithm, originally written by David Eyre. Advice in relation to the statistical analyses was provided by Sarah Walker and Tim Peto; I undertook the data analyses. I conceived of and wrote the published manuscript in JAC, which was modified in line with comments from contributing colleagues.

1.4.3. CHAPTER 6 – "Outbreak epidemiology – use of WGS to investigate an extended series of NDM-1-positive *Klebsiella pneumoniae* infections in a single institution in Nepal"

Isolates and raw epidemiological data collected in relation to the epidemiologically defined case clusters were supplied by colleagues at Patan Hospital; Prof Derrick Crook advised on the sampling frame for sequencing outside this cluster. I cultured and extracted DNA from samples. I conducted some of the phenotyping using the BD Phoenix; I was also assisted in this by routine diagnostic microbiology staff at the John Radcliffe Hospital Microbiology Laboratory. Illumina sequencing was carried out at the WTCHG sequencing hub; Illumina reads were processed through the Crook group's pipeline. PacBio sequencing of PMK1 and preliminary assembly into contigs was carried out at the Department of Genomics and Genetic Sciences, Mt Sinai, New York, by Drs Robert Sebra, Ali Bashir and Andrew Kasarskis. Adam Giess was responsible for finishing the PacBio genome assembly, determining single nucleotide variant level differences amongst the core genome of the outbreak strains, and

comparing PMK1 with the whole dataset at the chromosomal and plasmid level to generate Figure 6.12. I undertook the maximum-likelihood, BEAST, and Outbreaker analyses, with advice from Danny Wilson and Xavier Didelot. I interpreted all data in this chapter and generated all figures, except Figure 6.12.. I conceived of and drafted the manuscript published in AAC, which was modified in line with comments from contributing colleagues.

1.4.4. CHAPTER 7 – "Outbreak epidemiology – use of WGS to investigate an extended series of KPC-*Klebsiella pneumoniae* infections in a single institution in the USA"

Isolates and raw epidemiological data were collected by my colleagues, Drs Amy Mathers and Costi Sifri, at the University of Virginia, Charlottesville, USA; this collaboration was facilitated by Prof Derrick Crook. The study design and analyses was conceived of jointly by Dr Mathers and myself with advice from my supervisors. I cultured and extracted the DNA for the study. Illumina sequencing was carried out at the WTCHG sequencing hub; Illumina reads were processed through the Crook group's pipeline. Plasmid assemblies for pUVA01 and pUVA02 were created by Dr Anna Sheppard, who also generated the BLAST percentage comparisons represented in Table 7.7. I conducted all the other data analyses and created all the other figures in this chapter, with advice from Dr Xavier Didelot, who also assisted with some of the annotations for Figures 7.3. and 7.4.. The draft manuscript published in AAC was jointly written by Dr Mathers, Dr Anna Sheppard, and myself, and modified in line with comments from other contributing authors.

1.4.5. CHAPTER 8 – "Population genetics of CTX-M-associated resistance with a common global *Escherichia coli* lineage, sequence type (ST) 131"

My access to this global dataset of isolates for sequencing was made possible through collaborations with Prof Nick Day, director of the Mahidol-Oxford Tropical Medicine Research Unit (MORU), and directors of the South-East Asia units in Mae Sot (Prof Francois Nosten, Dr Paul Turner); Vientiane (Dr Paul Newton); and Siem Reap (me – Aug 2011-Feb 2012, when the unit was small in size); Profs James Johnson, Lance Price and Evgeni Sokurenko (USA); Veronica Kos and Bob McLoughlin (AstraZeneca research and development); and Dr Amee Manges (Vancouver). I cultured isolates and extracted DNA for the samples – the ST131 collection was identified as the result of sequencing and typing of a wider collection of over 1200 isolates. Sequencing and preliminary read data processing were as previously described; for the US dataset the raw read data were provided by our collaborators and was processed through the Crook group's bioinformatic pipeline as for newly sequenced isolates. I conceived of and carried out all of the analyses, with advice from my supervisors. I am specifically grateful to Dr Laura Mataseje (Winnipeg, Canada) for sharing her protocol for plasmid electroporation; to Dr Anna Sheppard, for writing a script to automate the BLASTn/heatmap approach to making comparisons with a large number of plasmid references, the graphical representations of mapping coverage for references, and providing general assistance with the bioinformatic analyses; to Prof Peto for suggesting the approach plotting pairwise comparisons of plasmid transformants versus TMRCA; and to Drs Xavier Didelot and Daniel Wilson for their suggestions and advice on the tree-building approaches used.

1.5. PUBLICATIONS

1.5.1. IN RELATION TO THIS THESIS

Mathers AJ/Stoesser N, Sheppard AE, Giess A, Yeh AJ, Didelot X, Turner SD, Peto TEA, Crook DW, Sifri CD. *Klebsiella pneumoniae* carbapenemase (KPC) producing *K. pneumoniae* at a Single Institution: Insights into Endemicity from Whole Genome Sequencing. Antimicrob Agents Chemother. 2015 Mar;59(3):1656-63.

Buchanan R, Stoesser N, Crook DW, Bowler IC. Multi drug resistant *Escherichia coli* soft tissue infection investigated with bacterial whole genome sequencing. BMJ Case Rep. 2014 Oct 19;2014. pii: bcr2014207200.

Stoesser N/Sheppard AE, Shakya M, Sthapit B, Thorson S, Giess A, Kelly D, Pollard AJ, Peto TEA, Walker AS, Crook DW. Dynamics of multiple drug-resistant *Enterobacter cloacae* outbreaks in a neonatal unit in Nepal: Insights using wider sampling frames and next generation sequencing. J Antimicrob Chemother. 2015 Jan 3. pii: dku521. Epub ahead of print.

Stoesser N/ Xayaheuang S, Vongsouvath M, Phommaosne K, Elliott I, del Ojo Elias C, Crook DW, Newton PN, Buisson Y, Lee S. Carriage of *Enterobacteriaceae* producing extended spectrum beta-lactamases in kindergarten children in the Lao People's Democratic Republic. Accepted in J Antimicrob Chemother, Feb 2015.

Stoesser N/Giess A, Batty EM, Sheppard AE, Walker AS, Wilson DJ, Didelot X, Bashir A, Sebra R, Kasarskis A, Sthapit B, Shakya M, Kelly D, Pollard AJ, Peto

TEA, Crook DW, Donnelly P, Thorson S, Amatya P, Joshi S. Genome Sequencing of an Extended Series of NDM-*Klebsiella pneumoniae* Neonatal Infections in a Nepali Hospital Characterizes the Extent of Community Versus Hospital-associated Transmission in an Endemic Setting. Antimicrob Agents Chemother. 2014 Dec;58(12):7347-57.

Stoesser N, Batty EM, Eyre DW, Morgan M, Wyllie DH, Del Ojo Elias C, Johnson JR, Walker AS, Peto TE, Crook DW. Predicting antimicrobial susceptibilities for *Escherichia coli* and *Klebsiella pneumoniae* isolates using whole genomic sequence data. J Antimicrob Chemother. 2013 Oct;68(10):2234-44.

Schlackow I, Stoesser N, Walker AS, Crook DW, Peto TE, Wyllie DH; Infections in Oxfordshire Research Database Team. Increasing incidence of *Escherichia coli* bacteraemia is driven by an increase in antibiotic-resistant isolates: electronic database study in Oxfordshire 1999-2011. J Antimicrob Chemother. 2012 Jun;67(6):1514-24.

1.5.2. IN RELATION TO OTHER WORK DURING THE COURSE OF THIS THESIS

Moore CE, Soeng S, Sar P, Patchhat H, Kumar V, Sopheary S, Stoesser N, Bousfield R, NPJ Day, CM Parry. Antimicrobial susceptibility of uropathogens isolates from children attending a pediatric hospital in Siem Reap, North West Cambodia, 2007-2011. Accepted in Paediatr Int Child Health, Sept 2014.

Khauv P, Turner P, Channy S, Sona S, Moore CE, Bousfield R, Stoesser N, Emary K, Thanh DP, Baker S, Thi V, Hang T, Van Doorn R, Day NPJ, Parry CM. Ophthalmic infections in children presenting to Angkor Hospital for Children Siem Reap Cambodia. BMC Research Notes, Sept 2014 (in press).

Carter MJ, Emary KRW, Moore CE, Parry CM, Soeng S, Putchhat H, Reaksmey S, Chanpheaktra N, Stoesser N, Dobson ADM, Day NPJ, Kumar V, Blacksell SD. Rapid diagnostic tests for dengue virus infection in febrile Cambodian children: Diagnostic accuracy and incorporation into diagnostic algorithms. PLoS Neglected Tropical Diseases, Sept 2014 (in press).

Pocock JM, Khun PA, Moore CE, Vuthy S, Stoesser N, Parry CM. Septic arthritis of the hip in a Cambodian child caused by multidrug-resistant *Salmonella enterica* serovar Typhi with intermediate susceptibility to ciprofloxacin treated with ceftriaxone and azithromycin. Paediatr Int Child Health. 2014 Apr 21:2046905514Y0000000123.

Jaroensuk J, Stoesser N, Leimanis ML, Jittamala P, White NJ, Nosten FH, McGready R. Treatment of suspected hyper-reactive malarial splenomegaly (HMS) in pregnancy with mefloquine. Am J Trop Med Hyg. 2014 Apr;90(4):609-11.

Dingle KE, Elliott B, Robinson E, Griffiths D, Eyre DW, Stoesser N, Vaughan A, Golubchik T, Fawley WN, Wilcox MH, Peto TE, Walker AS, Riley TV, Crook DW, Didelot X. Evolutionary history of the *Clostridium difficile* pathogenicity locus. Genome Biol Evol. 2014 Jan;6(1):36-52.

Eyre DW, Fawley WN, Best EL, Griffiths D, Stoesser NE, Crook DW, Peto TE, Walker AS, Wilcox MH. Comparison of multilocus variable-number tandem-repeat analysis and whole-genome sequencing for investigation of *Clostridium difficile* transmission. J Clin Microbiol. 2013 Dec;51(12):4141-9.

Stoesser N, Moore CE, Pocock JM, An KP, Emary K, Carter M, Sona S, Poda S, Day N, Kumar V, Parry CM. Pediatric bloodstream infections in Cambodia, 2007 to 2011. Pediatr Infect Dis J. 2013 Jul;32(7):e272-6.

Chheng K, Carter MJ, Emary K, Chanpheaktra N, Moore CE, Stoesser N, Putchhat H, Sona S, Reaksmey S, Kitsutani P, Sar B, van Doorn HR, Uyen NH, Van Tan L, Paris D, Blacksell SD, Amornchai P, Wuthiekanun V, Parry CM, Day NP, Kumar V. A prospective study of the causes of febrile illness requiring hospitalization in children in Cambodia. PLoS One. 2013 Apr 9;8(4):e60634

Stoesser NE, Martin J, Mawer D, Eyre DW, Walker AS, Peto TE, Crook DW, Wilcox MH. Risk factors for *Clostridium difficile* acquisition in infants: importance of study design. Clin Infect Dis. 2013 Jun;56(11):1680-1.

Stoesser N, Emary K, Soklin S, Peng An K, Sophal S, Chhomrath S, Day NP, Limmathurotsakul D, Nget P, Pangnarith Y, Sona S, Kumar V, Moore CE, Chanpheaktra N, Parry CM. The value of intermittent point-prevalence surveys of healthcare-associated infections for evaluating infection control interventions at

Angkor Hospital for Children, Siem Reap, Cambodia. Trans R Soc Trop Med Hyg. 2013 Apr;107(4):248-53.

Emary K, Moore CE, Chanpheaktra N, An KP, Chheng K, Sona S, Duy PT, Nga TV, Wuthiekanun V, Amornchai P, Kumar V, Wijedoru L, Stoesser NE, Carter MJ, Baker S, Day NP, Parry CM. Enteric fever in Cambodian children is dominated by multidrug-resistant H58 *Salmonella enterica* serovar Typhi with intermediate susceptibility to ciprofloxacin. Trans R Soc Trop Med Hyg. 2012 Dec;106(12):718-24.

Stoesser N, Pocock J, Moore CE, Soeng S, Hor P, Sar P, Limmathurotsakul D, Day N, Kumar V, Khan S, Sar V, Parry CM. The epidemiology of pediatric bone and joint infections in Cambodia, 2007-11. J Trop Pediatr. 2013 Feb;59(1):36-42.

Crook DW, Walker AS, Kean Y, Weiss K, Cornely OA, Miller MA, Esposito R, Louie TJ, Stoesser NE, Young BC, Angus BJ, Gorbach SL, Peto TE; Study 003/004 Teams. Fidaxomicin versus vancomycin for *Clostridium difficile* infection: meta-analysis of pivotal randomized controlled trials. Clin Infect Dis. 2012 Aug;55 Suppl 2:S93-103.

Ke L, An KP, Heng S, Riley M, Sona S, Moore CE, Parry CM, Stoesser N, Chanpheaktra N. Paediatric *Chromobacterium violaceum* in Cambodia: the first documented case. Trop Doct. 2012 Jul;42(3):178-9.

Moore CE, Hor PC, Soeng S, Sun S, Lee SJ, Parry CM, Day NP, Stoesser N.

Changing patterns of gastrointestinal parasite infections in Cambodian children: 2006-2011. J Trop Pediatr. 2012 Dec;58(6):509-12.

Stoesser N, Pocock J, Moore CE, Soeng S, Chhat HP, Sar P, Limmathurotsakul D, Day N, Thy V, Sar V, Parry CM. Pediatric suppurative parotitis in Cambodia between 2007 and 2011. Pediatr Infect Dis J. 2012 Aug;31(8):865-8.

1.6. ABBREVIATIONS (listed alphabetically)

AP-PCR	Arbitrarily-primed polymerase chain reaction
bla$_{\text{CTX-M}}$	Cefotaximase beta-lactamase (CTX-M) gene/enzyme
bla$_{\text{KPC}}$	*Klebsiella pneumoniae* carbapenemase (KPC) gene/enzyme
bla$_{\text{NDM}}$	New Delhi metallo-beta-lactamase (NDM) gene/enzyme
bla$_{\text{OXA}}$	Oxacillinase beta-lactamase (OXA) gene/enzyme
bla$_{\text{SHV}}$	Sulfhydryl variable beta-lactamase (SHV) gene/enzyme
bla$_{\text{TEM}}$	Temoneira beta-lactamase (TEM) gene/enzyme
bp	base-pair
BLAST/BLASTn	Basic Local Alignment Search Tool/nucleotide BLAST
BSI	Bloodstream infection
Contig	A contiguous sequence, for example one generated from a *de novo* assembly
CPE	Carbapenemase-producing *Enterobacteriaceae*
DNA	Deoxyribonucleic acid
EARSS/EARS-Net	European Antimicrobial Resistance Surveillance System/European Antimcrobial Resistance Surveillance Network
ECOR	*E. coli* Reference Collection
ESBL	Extended-spectrum beta-lactamase
ESC-R/S	Extended-spectrum cephalosporin resistant/susceptible
EUCAST	The European Committee on Antimicrobial Susceptibility Testing
ExPEC	Extraintestinal pathogenic *E. coli*
FDA	Federal Drug Administration
Gb	Gigabase

GI	Gastrointestinal
Inc	Plasmid incompatibility group
IRL	Inverted repeat, left
IRR	Inverted repeat, right
IS<*number*>	Insertion sequence, typically characterised with a numeric
Kb/Kbp	Kilobase/Kilobase-pair
k-mer	a sub-sequence (of length *k*) of a sequencing read
KPC	*Klebsiella pneumoniae* carbapenemase
MBL	Metallo-beta-lactamase
Mb/Mbp	Megabase/megabase-pair
MIC	Minimum inhibitory concentration
MLEE	Multi-locus enzyme electrophoresis
MLST	Multi-locus sequence typing
NCBI	National Centre for Biotechnology Information
NICU/PICU	Neonatal/paediatric intensive care
PCR	Polymerase chain reaction
PFGE	Pulsed field gel electrophoresis
RAPD	Random amplified polymorphic DNA
RFLP	Restriction fragment length polymorphism
rRNA	ribosomal RNA (ribonucleic acid)
SNV	Single nucleotide variant
ST	Sequence type (by multi-locus sequence typing)
Tn<*number*>	Nomenclature denoting defined transposon
tRNA	transfer RNA (ribonucleic acid)

tBLASTx	Search of translated nucleotide databases using a translated nucleotide query
UTI	Urinary tract infection
US	United States
UVaMC	University of Virginia Medical Centre
VLBW	Very low birthweight (<1500g)
WGS	Whole genome sequencing
WHO	World Health Organisation

CHAPTER 2: BACKGROUND

2.1. INTRODUCTION

Escherichia coli and *Klebsiella pneumoniae* are two species within the *Enterobacteriaceae* family of bacteria, which has over 40 genera and 160 species, many of which are common clinical pathogens(1). *E. coli* and *K. pneumoniae* are capable of causing a wide range of invasive disease in individuals of all age groups, including urinary tract, bloodstream, abdominal, respiratory and central nervous system infections, amongst others. Broad-spectrum antimicrobial resistance has emerged rapidly within both species in the last 20 years, and has been labelled as one of the key threats to human health by a number of international and national public health bodies, including the World Health Organisation (WHO), the US Centre for Communicable Disease Control (CDC) and Public Health England (PHE).

Beta-lactam antibiotics have been commonly used for treatment of infections caused by both species, and broad-spectrum beta-lactam resistance mediated by either extended-spectrum beta-lactamases (ESBL) or carbapenemases represents a critical threat. Various genetic elements have been specifically associated with the evolution and successful spread of beta-lactam resistance mechanisms, across a hierarchy ranging from different allelic variants of gene families, transposons, plasmids, and certain host bacterial strains, but unravelling the relative contribution of each of these has been difficult. The recent rapid development of next generation sequencing methods and decrease in sequencing costs have made high-resolution characterisation on larger datasets at each of these levels feasible(2).

Understanding the evolution and spread of genetic elements contributing to antimicrobial resistance is a necessary prerequisite to targeting interventions and modelling the likely evolution of successive resistance events. The broader aim of this thesis was therefore to use the higher resolution afforded by whole genome sequencing (WGS) methods to develop a more detailed understanding of key aspects of the molecular epidemiology of antimicrobial resistance in *E. coli* and *K. pneumoniae*. The main resistance mechanisms of interest in this thesis were those causing resistance to the beta-lactam class of antibiotics, although other resistance mechanisms of particular relevance are included in some of the analyses. The objectives related to developing approaches to identify all components included in the hierarchy of genetic elements described above, and applying them to clinically orientated sub-studies covering diagnostics (Chapter 5); infectious diseases epidemiology at the level of observed clinical outbreaks (Chapters 6 and 7); and then on a wider scale in the context of an epidemic, disease-causing lineage (Chapter 8).

2.2. A BRIEF NOTE ON SPECIES TAXONOMY AND NOMENCLATURE

2.2.1. *ESCHERICHIA COLI*

Escherichia coli is a facultatively anaerobic Gram-negative bacillus of the genus *Escherichia*, named after Theodor Escherich (1857-1911). Escherich, a German paediatrician, had discovered *E. coli* in the faeces of newborns as part of his work investigating the relationship between enteric bacteria and digestion in infants(3), initially describing it as “bacteria coli commune”, in reference to its commonality in the colon. Its potential role in clinical disease subsequently became swiftly apparent to clinical scientists, with the first description of associated pathology published by Laruette in 1889, and involving two cases of peritonitis(4). By 1896, it had been

described in a range of conditions, including enteritis, peritonitis, urinary tract and gall bladder infections, meningitis and puerperal sepsis, amongst others(5).

E. coli strains have been broadly classified into three major groups of relevance to humans, including (i) commensal strains, (ii) extra-intestinal pathogenic strains (extra-intestinal pathogenic *E. coli*; ExPEC) and (iii) intestinal or enteropathogenic (enteric/diarrhoea-causing) strains. Enteropathogenic strains, which are essentially obligate pathogens, can be further subdivided into at least seven pathotypes based on pathophysiological/genetic properties: (a) enterotoxigenic (ETEC), (b) Shiga-toxin-producing/enterohaemorrhagic (STEC/EHEC), (c) enteropathogenic (EPEC), (d) enteroinvasive (EIEC), (e) enteroaggregative (EAEC), (f) diffusely adherent (DAEC) and (g) *Shigella* spp. – the latter being recognised as a separate species, but clearly genetically associated with EIEC(6, 7). Whilst there is evidence of exchange of genetic material amongst all members of the species, the molecular epidemiology of the enteropathogenic group is distinct and beyond this particular thesis, and will therefore not be discussed further here.

Non-enteropathogenic *E. coli*, making up categories (i) and (ii) above, are the most numerous component of the facultative anaerobic faecal flora of most healthy human hosts, and cause disease only in a limited number of individuals, typically in response to precipitating factors such as impaired host immunity, the introduction of the organism to otherwise sterile sites or the presence of an indwelling foreign body. Within this group, the presence of some genetic features, such as genes encoding for adhesins, iron uptake mechanisms, cellular toxins and host-defense avoidance mechanisms, has historically been thought to confer increased clinical virulence, leading to the assignation of the aforementioned designation of ExPEC to these

strains(8). More recently, however, in light of an increasingly large body of genetic studies, has come the recognition that the overlap between strains traditionally though of as ExPEC and commensal strains is substantial, and that factors likely contributing to increased clinical virulence also confer an advantage with respect to gastrointestinal colonisation(9, 10). The distinction between the two is therefore not straightforward, and will require further studies of virulence.

2.2.2. *KLEBSIELLA PNEUMONIAE*

The genus *Klebsiella* was named after the German-Swiss pathologist, Edwin Klebs, (1834-1913) who was a prolific scientist and did much to associate pathogens with specific diseases, although he did not work directly on *Klebsiella* species. *Klebsiella pneumoniae* is one of up to twelve species in the genus, with an on-going debate concerning the exact nature of its taxonomic relationships, including those with respect to other genera, such as *Raoultella*. *K. pneumoniae* itself is widely considered to be made up of three subspecies: *pneumoniae*, *ozaenae* and *rhinoscleromatis*, with the sub-species designation historically dependent on clinical pathogenesis and differences in biochemical profile(11). *K. pneumoniae* subsp. *pneumoniae* was historically associated with respiratory infection and was first identified from the lungs of a patient with pneumonia; subsp. *rhinoscleromatis* with rhinoscleroma, a chronic granulomatous disease of the nose endemic in Africa, areas of Europe and Latin America; and subsp. *ozaenae* with ozena, a malodorous atrophic rhinitis. Recent genetic work has shown that subsp. *ozaenae* and subsp. *rhinoscleromatis* are likely monomorphic clones within the *K. pneumoniae* subsp. *pneumoniae* group(12). This work focuses on an investigation of resistance in *K. pneumoniae* subsp. *pneumoniae*, with none of the sampling frames structured to collect isolates from patients with rhinoscleroma or ozena as clinical sub-groups.

2.3. MOLECULAR TYPING METHODS FOR NON-ENTEROPATHOGENIC *ESCHERICHIA COLI* AND *KLEBISELLA PNEUMONIAE*

The epidemiological literature on antimicrobial resistance in non-enteropathogenic *E. coli* and *K. pneumoniae* is predicated on a broad range of typing methodologies, which reflect access to and evolution of typing technologies, the particular questions for a given study, and user preference. This section focuses on summarising the major traditional typing methods, in order to contextualise the analytical method used in this thesis, namely whole genome sequencing (WGS), which is described in greater detail in Chapter 2 – "Thesis Methods".

2.3.1. SEROTYPING

The development of serotyping as a technique for classifying non-enteropathogenic *E. coli* and *K. pneumoniae*, as well as a number of other *Enterobacteriaeceae*, was principally driven by the work of Ida Orskov, Fritz Orskov and Fritz Kauffmann at the Danish Statens Serum Institut (SSI) between 1964-1992. The principle behind the technique was to use agglutination with specific antisera to determine the presence of certain antigens; each isolate could then be given a numeric profile. For *E. coli*, this involves the detection of somatic lipopolysaccharides (O), capsular (K) and flagellar (H) antigens; for *K. pneumoniae*, the O or K antigens. Certain associations with disease phenotypes have been identified in both species, such as O157:H7 or O26:H11 for enterohaemorrhagic infections or K1 strains in meningitis with *E. coli*(13), or K1 strains of *K. pneumoniae* with community-associated disease and primary liver abscess(14, 15). Hundreds of antigen types have been characterised(16), but the technique is slow, expensive, labour-intensive and essentially only done in reference laboratories. In addition, its resolution is too limited to use effectively for fine-scale analyses of outbreaks and population genetic comparisons.

2.3.2. MULTI-LOCUS ENZYME ELECTROPHORESIS (MLEE)

This approach is based on an assessment of the variability of electrophoretic mobility of enzymes in bacterial strains, and enabled the first quantitative assessment of differences between strains and within a population, by correlating protein electrophoresis patterns for 12 enzyme loci with the presence of different allozymes, and correspondingly, underlying nucleotide sequence for different isolates(17).

In brief, cells are grown up in culture and then lysed by a variety of methods. Crude protein extracts are separated out from lysed debris by high-speed centrifugation, and then run out and stained on a gel(18). The variability in mobility comes about as a result of changes in charge within the molecule that are generated by amino acid substitutions. In some cases however, different allozymes will have the same charge, and charge will not be affected by synonymous substitutions. As a typing method it has therefore largely been superseded by direct sequencing-based methods, which are also more effective in terms of generating robust data for inter-laboratory comparisons.

2.3.3. PHYLOTYPING/PHYLOGROUPING

This is a technique specific to *E. coli* and is based on the discrete clustering of strains of *E. coli* into distinct genetic clusters, termed phylogroups. Initially, four of these phylogroups were observed, and were named A, B1, B2 and D. A triplex PCR based on the presence/absence of two genes (*chuA*, *yjaA*) and a DNA fragment (TSPE4.C2) was developed to enable rapid classification of strains into these groups, and formed the basic technique for subsequent 'phylogrouping/phylo-typing'(19). Since then, as a larger number of strains have been processed, this system has been revised to include a new gene target *arpA*, leading to a more refined classification scheme encompassing

additional phylogroups C, E and F(20). Phylogrouping is a robust method for broad classification of *E. coli* into genetically related clusters, but again has limited resolution to make any detailed comparisons.

2.3.4. DNA FINGERPRINTING METHODS

A number of these methods have been developed as a means to type organisms in a variety of contexts, ranging from describing population structure to defining transmission events within outbreaks. Essentially they fall into two categories: (i) those based on restriction analysis of bacterial DNA (restriction fragment length polymorphism [RFLP] analyses); (ii) those that depend on the amplification of certain genetic targets by PCR. All of the approaches in these categories are highly dependent on experimental conditions and variability in results can be observed between runs on the same DNA extract, between operators, and between laboratories. They also commonly require a subjective interpretation of banding patterns, affording additional scope for error, even with the use of software packages such as PyElph or E-Gel Imager to undertake this assessment in a more formalised way. Different methods are considered more appropriate depending on the context of the research question, making it difficult to use one technique to answer a number of different research questions, and to make comparisons with other datasets which have been used to answer different research questions(21).

Pulsed field gel electrophoresis (PFGE) involves the enzymatic cutting or restriction of extracted bacterial chromosomes with a set of restriction enzymes, and the staining and visualisation of the cut fragments on a gel. Typically rare cutting enzymes are used, so that the fragments generated are larger and less numerous. Resulting patterns are then compared for similarity either by visual inspection, or by using software that

calculates the similarity of banding patterns according to a distance algorithm. Standardised protocols have been developed to try and overcome methodological variability, and this technique has been widely used for outbreak investigation in both non-enteropathogenic *E. coli* and *K. pneumoniae*. It is currently considered one of the gold standard typing methods for this sort of investigation because of its high discriminatory power. PFGE is however labour-intensive and requires several days to process the sample and analyse the results(21). It is also less well-suited to an analysis of population structure over the longer and medium-term, as a number of mechanisms could be disproportionately associated with wide restriction site divergence and different banding patterns, including: nucleotide substitutions, insertion sequence migration, gain or loss of plasmids or pathogenicity islands and rearrangements. The main difficulty is that these events cannot be interpreted from the underlying banding pattern changes, or distinguished from each other(22).

In random amplified polymorphic DNA (RAPD) analysis, segments of template DNA are amplified at random using a single 10-mer primer that will bind to many different loci and initiate amplification of a range of random segments from a single DNA template. This process is carried out under low-stringency conditions, which facilitates primer binding in the absence of complete homology. The amplification products are then separated by gel electrophoresis and banding patterns compared across isolates. Although this technique offers some advantages, such as a low input DNA requirements and no need for a knowledge of the underlying genetic sequence, it is one of the techniques most prone to inter-experimental variability(23). A similar technique has been developed using primers of 15 nucleotides, with different amplification and electrophoretic conditions – this is known as arbitrarily-primed PCR (AP-PCR), but the principles are the same(24), and the same pitfalls remain. For

non-enteropathogenic *E. coli*, PFGE is considered to provide greater typing discrimination than RAPD-PCR(21).

2.3.5. MULTI-LOCUS SEQUENCE TYPING (MLST)

Multi-locus sequence typing is a nucleotide sequence-based method of typing isolates that relies on determining the sequence for a given number of housekeeping genes, which are thought to reflect core genomic evolution but are not much affected by recombination or horizontal gene transfer. Typically, DNA templates are amplified with locus-specific PCR reactions for around 7 genes in a scheme, sequenced, and then assigned allelic profiles for each housekeeping gene on the basis of nucleotide sequence. Allelic profiles are then assigned a sequence type, or ST.

MLST has been widely used in the typing of non-enteropathogenic *E. coli* and *K. pneumoniae*, with the two most commonly used schemes hosted by the University of Warwick (formerly at the University of Cork, Ireland) for non-enteropathogenic *E. coli* (http://mlst.warwick.ac.uk/mlst/dbs/Ecoli (25)) and the Institut Pasteur in Paris for *K. pneumoniae* (http://www.pasteur.fr/recherche/genopole/PF8/mlst/Kpneumoniae.html (26). These websites catalogue a large number of profiles – 1801 for *K. pneumoniae* (Figure 2.1.) and 4949 for *E. coli* – indicative of the phenomenal diversity represented in these species. MLST is useful in that it is a directly comparable representation of underlying genetic sequence, and is comparable between studies. Nevertheless, it represents only a fraction of the genomic content present in a non-enteropathogenic *E. coli* and *K. pneumoniae* strain (~0.05-0.10%), and cannot typically be used for comparative analyses over short evolutionary timescales, or for detailed strain comparisons.

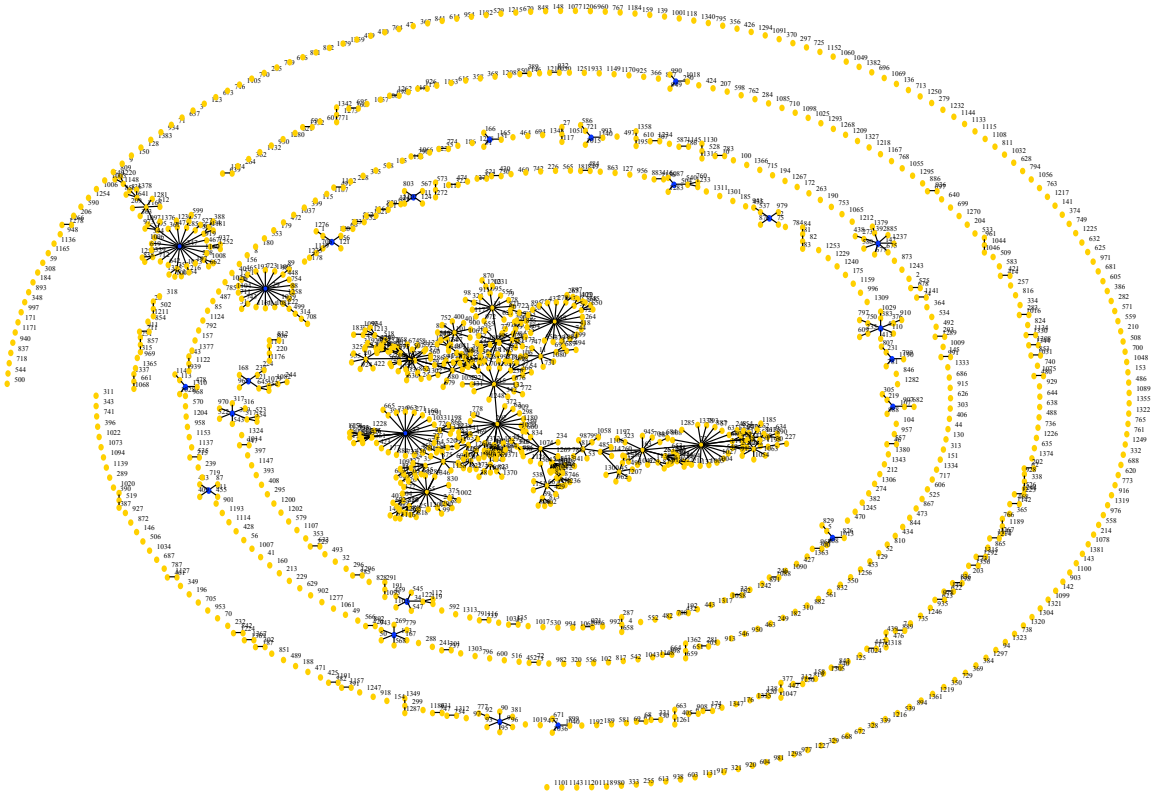


Figure 2.1. Eburst depiction of *K. pneumoniae* ST profiles depicting singleton STs and clonal complexes (generated June 2012).

2.4. EPIDEMIOLOGY OF NON-ENTEROPATHOGENIC *ESCHERICHIA COLI* AND *KLEBSIELLA PNEUMONIAE*

Non-enteropathogenic *E. coli* and *K. pneumoniae* strains both have the capacity to cause disease, but are also found as asymptomatic colonisers of the gastrointestinal tract of a large number of species and exist in the environment. This section gives a brief overview of the relevance of both species to human disease, and of the epidemiology of their existence as commensal and environmental organisms.

2.4.1. CLINICAL DISEASE

2.4.1.1. NON-ENTEROPATHOGENIC *ESCHERICHIA COLI*

Non-enteropathogenic *E. coli* are a common cause of a wide range of clinical infections amongst individuals of all age groups, and can involve almost any

anatomical site(8). Typical sites of infection include the urinary tract, the central nervous system (especially in neonates) and intra-abdominal locations such as the biliary tract, with less common involvement of the musculoskeletal or integumentary systems. Intra-vascular device-associated/catheter-associated and pulmonary infections are more typically prevalent within hospital/institutional settings(27). All of these infections can be associated with concomitant bloodstream infection (BSI), which may also occur without any obvious source.

The urinary tract is the most frequently affected clinical site of infection, and non-enteropathogenic *E. coli* are responsible for approximately 80% of the 150 million urinary tract infections (UTIs) that occur annually, at a cost of approximately six billion US$(28). Almost half of all women will have had at least one episode of UTI in their lifetime, with additional risk groups represented by infants, pregnant women, patients with diabetes or immunodeficiencies, and patients with underlying urological abnormalities(29). UTIs also have a particularly high prevalence in the elderly, and represent either the most common or second most common infections in elderly women living in long-term care facilities or the community, respectively(30). This has major healthcare implications for high-income countries in particular, which are experiencing an on-going demographic shift towards an ageing population.

Non-enteropathogenic *E. coli* are implicated as the commonest cause of bloodstream infections in a number of epidemiological surveys. In the UK, for example, non-enteropathogenic *E. coli* caused 23% of all BSI reported through a voluntary surveillance scheme run by the Health Protection Agency (HPA, now Public Health England [PHE]), with an increase of 33% over the 2004-2008 study period(31). In addition, rising rates of infection were attributable to increasing numbers of

organisms resistant to the antimicrobials most frequently used for clinical management(32). Concerns about the rising rates of non-enteropathogenic *E. coli* -BSI led to the implementation of mandatory surveillance for these infections in June 2011; this has since demonstrated there are approximately 32,300 *E. coli*-BSI/year nationally, which equates to a rate of 60.8/100,000 population.

A similar epidemiological picture was observed in 33 European/Mediterranean countries, using data from the European Antimicrobial Resistance Surveillance System (EARSS, now European Antimicrobial Resistance Surveillance, EARS-Net). Here, non-enteropathogenic *E. coli* were again noted to be the most commonly reported pathogen in BSI, with year-on-year increases observed to 2008; these increases were again driven by rising numbers of antimicrobial-resistant pathogens(33). In the EARSS 2013 Surveillance Report the percentage of multi-drug resistant *E. coli* isolates (defined as showing combined resistance to third generation cephalosporins, fluoroquinolones and aminoglycosides) significantly increased in 13 of 28 reporting countries, with only two countries (Bulgaria and Malta) reporting decreasing trends(34). In this report, the percentage of *E. coli* resistant to third generation cephalosporins ranged from 3% (Sweden)-36% (Cyprus).

In resource-limited settings, non-enteropathogenic *E. coli* are also one of the commonest community-associated bacterial causes of BSI. In two recent systematic reviews investigating bacterial causes of BSI in both adults and children, non-enteropathogenic *E. coli* accounted for 240/2132 (6.8%) infections in South and South-East Asia and 412/2331(7.3%) of infections in Africa, making it the second most common Gram-negative cause of BSI after *Salmonella enterica* (*S. Typhi* in Asia, non-typhoidal salmonellae in Africa)(35, 36).

Bacterial infections are a leading cause of the 3 million global neonatal deaths per annum, with sepsis/meningitis representing the third most common cause (5.2% of deaths; 393000 cases)(37). *Streptococcus agalactiae* (Group B streptococcus [GBS]) and non-enteropathogenic *E. coli* are the organisms most commonly associated with early-onset infection (variably defined as within 72 hrs-7 days of delivery), and are responsible for approximately 70% of these infections. Non-enteropathogenic *E. coli* disproportionately affects pre-term/very low-birthweight (VLBW; <1500g) neonates and is potentially rising in incidence in the US, given the frequency of maternal prophylaxis for GBS and the larger numbers of pre-term VLBW that can be supported with improved healthcare(38).

2.4.1.2. *KLEBSIELLA PNUEMONIAE*

K. pneumoniae has dominated as one of the leading causes of nosocomial Gram-negative infection, and like non-enteropathogenic *E. coli*, can be a causative pathogen in a number of clinical syndromes, including urinary tract, pulmonary, device-associated, soft tissue/surgical site and bloodstream infections(39). In the UK, surveillance data for *Klebsiella* spp. are voluntary, but rates of bloodstream infection show small year-on-year increases, with an overall rate of around 11 bacteraemias/100000 population/year. Increasing rates of resistance to broad-spectrum beta-lactams such as extended-spectrum cephalosporins and carbapenems are also statistically significant, although these are low overall for the carbapenems (<1%)(40). In recent European surveillance data, *K. pneumoniae* was one of the top ten pathogens isolated in infections in long-term care facilities(41), and one of the commonest pathogens isolated in healthcare-associated infections, although still currently at lower rates than non-enteropathogenic *E. coli* in most Western European

contexts(42). It is recognised as a highly significant pathogen in the paediatric critical care setting(39).

The impact of *K. pneumoniae* is not, however, limited to the nosocomial setting - it also presents a major community-associated threat, most significantly in resource-limited contexts. In the two large systematic reviews of bloodstream infection in Africa and South-East Asia discussed earlier, it is one of the top five Gram-negative causes in both adults and children(35, 36). It is recognised as a particular contributor to infant and neonatal sepsis(43). In countries of the Asian Pacific Rim, such as Taiwan, certain strains with a hypermucoviscous phenotype and specific capsular serotypes have also been associated with a community-acquired syndrome encompassing pneumonia, liver abscess and metastatic disease(39). This clinical entity is found in younger and less comorbid patients, and may be emerging in Europe(44).

2.4.2. CARRIAGE AND RESERVOIRS

Non-enteropathogenic *E. coli* and *K. pneumoniae* are not obligate pathogens, and strains causing clinical disease are also carried asymptomatically in humans and animals, as well as being distributed in the environment. The wide niche in which these organisms can evolve and replicate has important implications for the development of resistance and for possible transmission networks. Although a comprehensive investigation of environmental and carriage isolates was beyond the scope of this thesis, it is of relevance to the studies presented, and so a brief review of the available data on their presence in these compartments is considered below.

2.4.2.1. HUMAN CARRIAGE

The predominant reservoir of non-enteropathogenic *E. coli* for human hosts is the gastrointestinal (GI) tract. At birth, the GI tract is sterile, becoming rapidly colonised by a number of different bacterial genera, the pattern of which seems to be heavily influenced by environmental exposures, including maternal flora, diet and healthcare exposures, with progression to a more stable microbiome somewhere between the ages of one and four years(45, 46). Non-enteropathogenic *E. coli* are some of the earliest colonisers, and typically become the most dominant aerobic gut organism, despite the fact that they are outnumbered by 100-10000:1 by obligately anaerobic bacteria, which make up 99.9% of the 10^{11} organisms/g of faeces in the human colon(47). Carriage prevalence of non-enteropathogenic *E. coli* amongst humans is estimated at greater than 90%(48), and is probably likely to be closer to 100%, invariably occurring by one-two months of age(49). Non-enteropathogenic *E. coli* reside in the mucus layer of the colonic epithelium, which provides them with the appropriate nutritional environment for their metabolic needs(50).

Studies specifically investigating the prevalence of faecal carriage of non-resistant *K. pneumoniae* in healthy, asymptomatic individuals are less common. Early *K. pneumoniae* colonisation of the neonate appears to be influenced by a number of different factors, such as birth in a developing country, pre-term birth/low birthweight, early-formula feeding as opposed to breastfeeding, birth by caesarean section, hospital exposure and antibiotic use(51). In a longitudinal study of 22 home-delivered infants in Pakistan, most neonates were initially colonised by *Enterobacteriaceae* other than non-enteropathogenic *E. coli*, including *K. pneumoniae*, between day 2 and 7 of life, with the hypothesis that these were acquired from environmental sources such as cow's or buffalo's milk, often given in addition

to breastfeeding(49). In hospital environments, particularly critical care environments, colonisation by *K. pneumoniae* may occur in more than 60% of neonates(52). Rates of carriage in children <1 year of age appear highest in early infancy (~70%), tailing off between the ages of 1-5 years(53); estimates of carriage rates in adults are approximately 10-20%(53, 54).

Older studies give us some insight into the prevalence of carriage of *K. pneumoniae*, but may have been limited by the use of culture-based approaches. Recent studies have used metagenomic data, such as those obtained through microarrays. One such study demonstrated that *K. pneumoniae* represented 0.01% of bacterial faecal flora in healthy adults (21-60 years) versus 0.02% in young children between the ages of 1 and 4 years, suggesting that these organisms may be present in the gastrointestinal tract but remain unsampled in culture-based surveys due to, relatively speaking, low numbers. They may however remain important for clinical transmission and the spread of resistance genes.

There are a number of studies of faecal carriage prevalence of ESBL- and/or carbapenemase-positive non-enteropathogenic *E. coli*/*K. pneumoniae* from a variety of global settings, including both nosocomial and community-sampling frames amongst adults and children. The first reports of community-based carriage were published in 2000 and 2001-2, from studies in Poland and Spain respectively(55, 56); rates before 2008 were rarely above 10%. Figures 2.2. and 2.3. show the evolution of community ESBL carriage rates over time, and an estimation of the human reservoir by WHO geographical region(57). What is strikingly obvious are the marked differences between regions, which may reflect a number of factors including antibiotic selection pressures, population density and levels of sanitation; and the fact

that the asymptomatic reservoir, and the potential for human-human, and wider, transmission is enormous.

The diversity of non-pathogenic *E. coli* strains carried within humans has been shown to be different depending on the population sampled, and this may be influenced by a complex interplay between geography and climate, diet, and levels of sanitation. In one study, the degree of diversity in three separate populations was investigated, namely French residents in France, French expatriates in Guyana, and an indigenous group of Guyanese, the Wayampi Indians. The degree of diversity was shown to be significantly higher in the Wayampi than in the French residing in France (3.1 versus 1.9 strains; p=0.03). The expatriates fell between these two values, with a mean of 2.3 strains, although this was not significantly different from either of the other two populations(58). A further observation made in this study was that the distribution of strain types, as defined by phylogroup, was different in the Wayampi than in the French, with a larger proportion of A and B1 strains represented in the Wayampi, and a larger proportion of B2 strains observed in the French residents in France (the B2 phylogroup strains are more commonly associated with clinical disease). An extension of this study featured the inclusion of additional groups of individuals, and confirmed the finding that geography and climate seem to play an important role in the structuring of the within-host *E. coli* population, although the contribution of socioeconomic factors to these differences was acknowledged but not investigated(59). The diversity of carried *K. pneumoniae* strains has not, to my knowledge, been determined in any published study. Differences in the

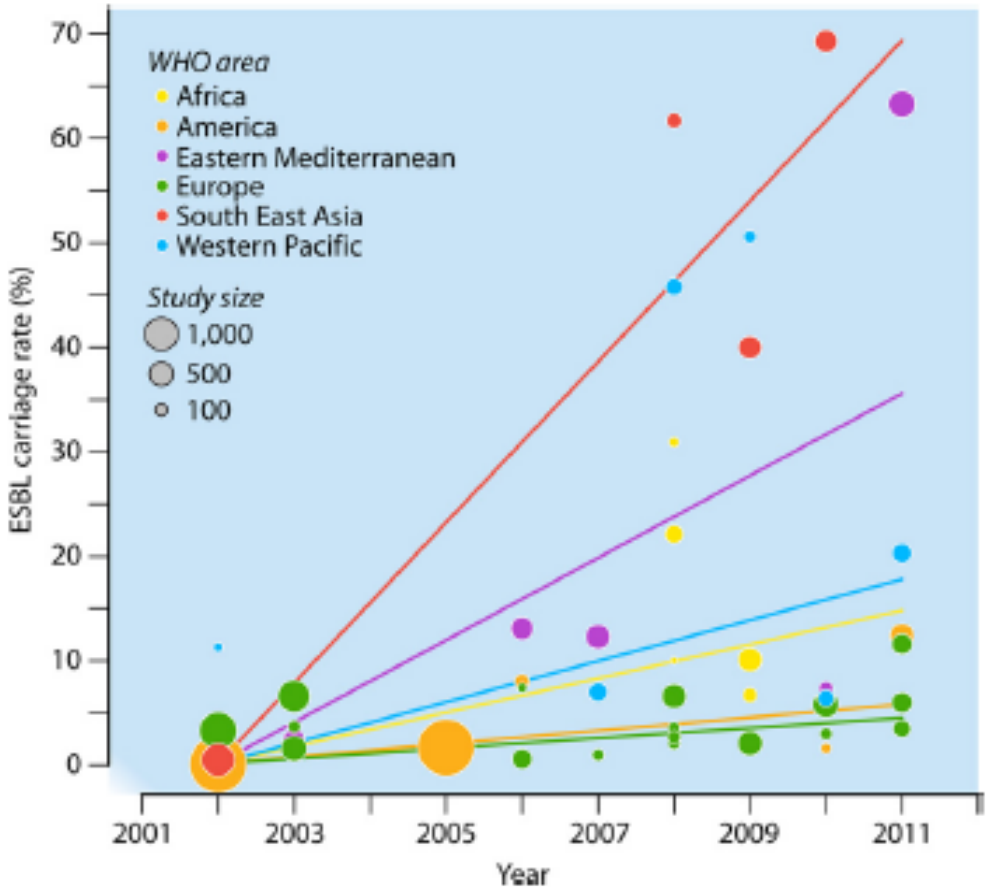


Figure 2.2. ESBL-*Enterobacteriaceae* carriage rates in the community, according to their geographical and temporal distribution. Each bubble is proportional to the size of the corresponding study. The lines represent the evolution of ESBL-*Enterobacteriaceae* carriage rates over time for each geographical area, as established by a weighted linear regression model. Taken from (57).

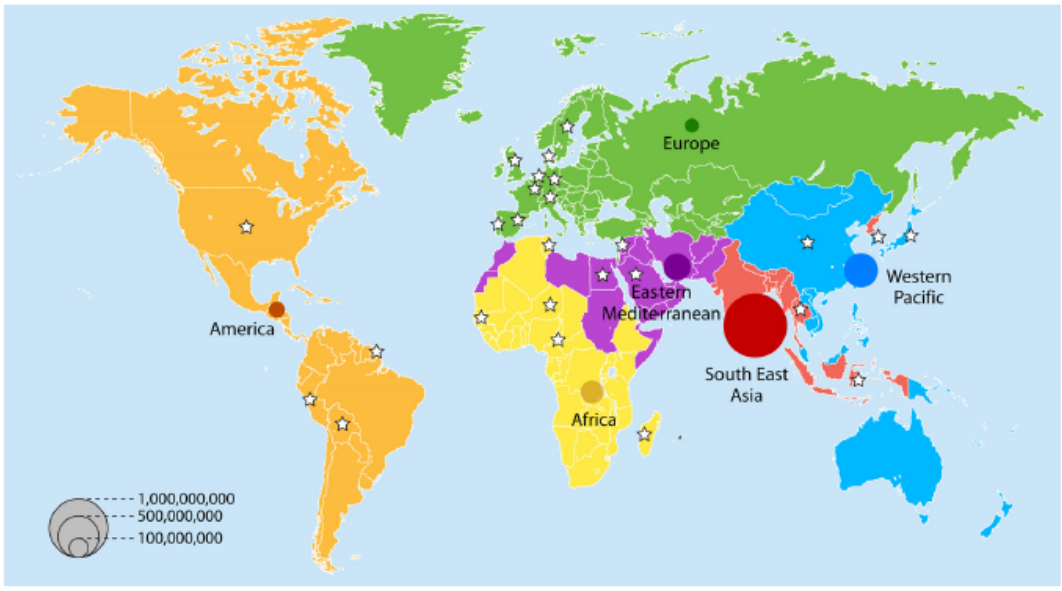


Figure 2.3. Number of ESBL carriers in the community in 2010, according to WHO region grouping (represented by colours). Stars represent countries with data available for modelling. Each bubble is proportional to the estimated number of ESBL carriers for that region. Taken from (57).

gastrointestinal niche may have important implications for the emergence of resistance, and for the selection of clinically "successful" lineages.

The diversity of carried ESBL or carbapenemase-producing strains within hosts has been investigated only in a limited fashion, and typically in terms of assessing diversity at a species level where it has been noted that individuals can carry more than one ESBL/carbapenemase-positive species of *Enterobacteriaceae*(60-62). One study in Spain typed two colonies in 7/105 ESBL-positive non-enteropathogenic *E. coli* isolates, but these were not investigated further as they shared the same antimicrobial resistance phenotypes and had evidence of the same ESBL gene. In terms of transmission epidemiology, in high-income settings where resistance prevalence is still considered relatively low, the prevailing idea is that resistant strains are representatives of single clones, and can be tracked on this basis. This idea may be appropriate when resistance is rare and the mechanisms causing it are homogenous, but is likely to become less so with increasing prevalence and diversification, and the increased likelihood of repeated introductions/exposures from high prevalence areas given the "global village".

2.4.2.2. ANIMAL CARRIAGE

Animals also provide an important reservoir of both susceptible and drug-resistant non-enteropathogenic *E. coli*. Non-enteropathogenic *E. coli* can be commensals and/or pathogens in a wide range of vertebrate species, with most study data coming from mammals and/or birds. One of the largest animal studies undertaken investigated 1898 non-enteropathogenic *E. coli* from 35 mammalian and 11 bird species, sampled in the 1980s and 2000s, identifying sampling year, the extent of domestication and diet all having an impact upon the degree of intra-host diversity, number of virulence

genes and distribution of phylogroups present(63). A second large study carried out in Australian vertebrates identified body mass, diet, climate and taxonomic rank as being important factors in both prevalence rates of non-enteropathogenic *E. coli* carriage and distribution of phylogroups – non-enteropathogenic *E. coli* were most commonly found in mammals (56%), and birds (23%), followed by relatively low carriage rates in reptiles (12%) and fish (10%). However, prevalence varied from 0-100% in both mammal and bird groups, suggesting that there may also be a less clearly understood impact of physiology and exposure on rates of carriage amongst the lower taxonomic orders(64).

ESBL-producing-non-enteropathogenic *E. coli* have also been identified in wild animals, having been first observed in 2006 in birds of prey and deer in Portugal(65). Since then ESBL- non-enteropathogenic *E. coli* strains have been isolated from: wild birds(66, 67), red foxes(68), deer(69), stray dogs(70), rodents(71), wild boar(72), and gorillas(73), amongst many others.

The presence of non-enteropathogenic *E. coli* in companion animals such as dogs and cats has been investigated in a number of studies, and both of these pets are commonly colonised or infected by isolates also thought to be causing disease in humans(74-76). Studies of farm animals have shown that carriage of non-enteropathogenic *E. coli* is common in a number of different farmed species, including cattle, swine, horses and broiler chickens(77-82). Antimicrobial resistance, including ESBL-type resistance, is frequent, and may be increasing, with carriage prevalence rates of 3-33%(80, 83), 2-63%(80, 81, 83), and 0%(83) for cattle, pigs and chickens reported respectively, depending on geographic location and year of sampling. In many cases, as in humans, non-enteropathogenic *E. coli* are acquired

within days of birth(77), and this can include colonisation with drug-resistant strains. In a multi-site study of pig farms in Denmark, rates of colonisation with ESBL- non-enteropathogenic *E. coli* were shown to be associated with various features of the production cycle and aspects of husbandry, including the historic usage of third and fourth generation cephalosporins (banned since 2010 in Denmark), intensive cleaning of farm sections, and possibly with feeding practices during the stages of weaning and finishing(84). A number of recent studies have attempted to demonstrate the overlap in isolates obtained from livestock, retail meat, and clinical cases of human disease, and have shown that there is apparent transmission between these compartments(85, 86).

There are very limited data specifically related to the asymptomatic carriage of *K. pneumoniae* in animals. There is one report of high rates of carriage (20-40%) of hypermucoid strains of *K. pneumoniae* in vervet monkeys on St. Kitts(87), and of drug-resistant strains in household cockroaches and houseflies in Libya(88, 89). There have however been a number of studies in which *K. pneumoniae,* including ESBL- and carbapenemase-producers, have been cultured from samples taken from companion animals and livestock, with the implication that these are involved in transmission networks(90, 91).

2.5 GENETIC POPULATION STRUCTURE OF THE SPECIES: EVOLUTION AND DIVERSITY

The background diversity of a species has important implications for sequence data analysis and interpretation. This section summarises aspects of genetic diversity relevant to the data presented in the thesis.

2.5.1. *ESCHERICHIA COLI*

The analysis of diversity in non-enteropathogenic *E. coli* predates the availability of nucleotide data. Early analyses of diversity were based on using serotyping to distinguish different non-enteropathogenic *E. coli* strains, revealing a diverse range of antigen types (Section 2.3.1) (92). DNA hybridisation studies confirmed the diversity present within the species, with hybridisation values of between 36-100% for experiments involving a number of non-enteropathogenic *E. coli* strains in combination with the *E. coli* reference K-12 strain(93). Subsequently MLEE (Section 2.3.2) (17) became widely used as a technique to study *E. coli* diversity, an example being an analysis of a large collection of 1608 isolates from clinical/carriage, human/animal, adult/paediatric contexts from a number of geographical locations categorised into three groups (I, II and III), with groups II and III more closely genetically related(94, 95). A sub-group of 72 of these strains thought to be representative of the genetic diversity in the species as a whole was subsequently set up as the ECOR (*E*. *CO*li *R*eference) collection(96).

The results of MLEE-based studies identified substantial diversity within the species, and whilst in some cases similar or identical enzyme profiles were present in unrelated isolates, suggesting that recombination may play an important role, the thought was that effectively *E. coli* was evolving in a clonal manner. These studies also demonstrated that the O/K/H serotyping scheme gave limited insight into the true population structure, given that genetically distant strains could exhibit the same serotype, and closely related strains could have different profiles. MLEE profiles could also be used to generate proxy phylogenies, which resolved the species into distinct genetic clusters, which were termed phylogroups. These form the genetic subsets that are distinguished by the PCR-based phylogrouping method described

earlier (and initially included A, B1, B2 (thought to be sister taxa), D, and a fifth, smaller group, E(97, 98)). The simple A-E phylogroup classification has however been well-supported in subsequent studies using a whole range of techniques, including restriction fragment length polymorphism (RFLP), random amplified polymorphic DNA (RAPD)(99) and more recently nucleotide sequence data, in the form of both multi-locus sequence typing (MLST) and whole genome sequencing (WGS)(98).

MLEE is a limited approach to defining phylogenetic relationships, and enzymes with different underlying sequence structure may have similar electrophoretic capacity. Multi-locus sequence typing (MLST)-based studies superseded MLEE methods, and in the first instance used different sets of housekeeping genes to characterise the relationships between a range of *E. coli* strains (both enteropathogenic and non-enteropathogenic). However, the use of different sets of housekeeping genes led to the determination of different topologies and conclusions, with some studies highlighting the clonal structure of the species(25, 100, 101), and some suggesting that recombination was frequent. The result is therefore dependent upon the catalogue of genes selected, and with the small fraction of the genome analysed (~0.07%) the outcome could be easily biased, if for example, one of the genes chosen for a scheme was in a "hot-spot" of recombination, or under selection(98).

Analysis of whole genomes gives greater insight into the phenomenal diversity within the species. In one of the first of these studies, an analysis of 20 fully sequenced genomes demonstrated that the average number of genes was 4,721, with only 1,976 (42%) of these represented in all 20 strains. The pan-genome (i.e. the full set of non-orthologous genes amongst all the genomes) on the other hand contained 17,838

genes, with the core genome therefore composed of only 11% of all of these genes (Figure 2.4.).

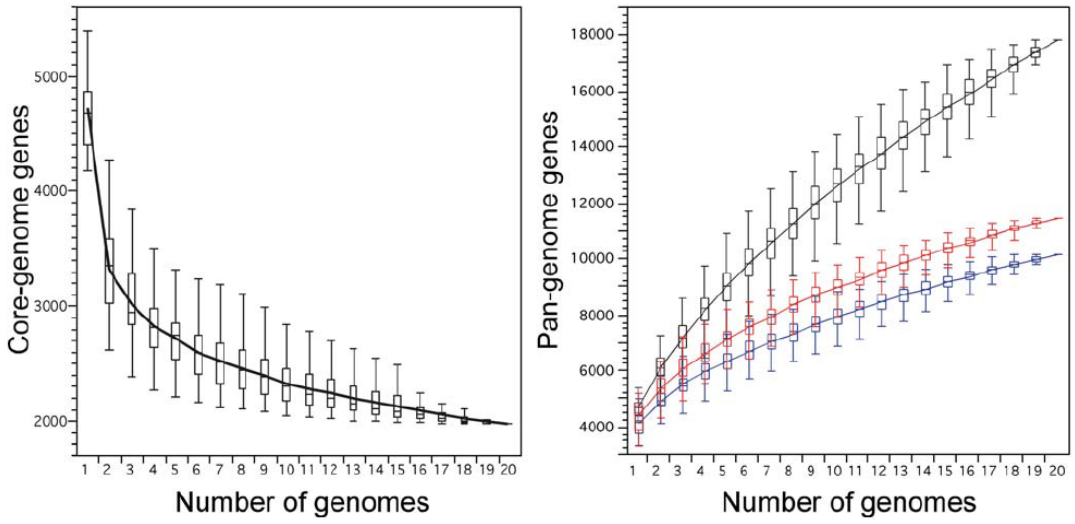


Figure 2.4. A representation of *E. coli* core and pangenomes in relation to the number of sequences included in the comparison. The panel on the right includes comparisons of all genes (black), all genes with <80% sequence homology (red), and all genes excluding insertions sequence and prophage-like elements (blue). Taken from (102).

As can be seen, the curve for the pangenome shows no sign of flattening, a phenomenon which has been described as an "open pan-genome"; in other words, the pool of genes that can be gained or lost by members of the species through horizontal gene transfer is vast. The figure also demonstrates the contribution made by mobile and repetitive elements to diversity within the species.

The huge genetic diversity present in *E. coli* was also in evidence in a second study, which went further in defining features of the recombination events amongst 27 fully sequenced members of the A, B1, B2 and E phylogroups(103). Here the authors used a program called progressiveMauve(104) to align the genomes, and then extracted core regions of at least 500bp found in all 27 genomes. These stretches of sequence ranged in size form 501bp to 27,115bp, with a mean of 4,322bp and a concatenated length of 3.3Mbp. They subsequently applied ClonalFrame to the core alignment,

which is a Bayesian method that infers a phylogeny under an evolutionary model and accounts for the effects of homologous recombination(105). This gave rise to a fully resolved clonal genealogy, in which A and B1 were most closely related, and B2 the most divergent clade. They estimated the average length of fragments involved in homologous recombination as being 542bp, with a relative recombination rate of 1.024 (recombination happened just as frequently as mutation). Although this rate was mostly constant throughout the genome, there were three "hot-spots" of recombination in which rates were significantly higher – this included two known regions encompassing regions around the *rfb* operon (involved in O-antigen synthesis) and the *fimA* gene. A third, smaller region containing a number of other genes including *aroC* was also identified, with the suggestion that this may be associated with virulence and there be under a selection pressure imposed by the immune system. The authors also used these estimates of recombination to identify the number of recombination events within and between the phylogroups analysed, finding evidence to support greater within phylogroup recombination that between phylogroups, with the exception of the more closely related A and B1. They hypothesised that this would be consistent with concept of on-going speciation (or sub-speciation) in *E. coli*, where species appear when the impact of recombination between sub-groups is reduced, and that this may be in response to subtle differences in ecological niches or life cycle to the extent that the genetic flux between lineages diminishes and they gradually diverge from each other.

The enormous genomic plasticity and diversity observed within the species, and the fact that *Shigella* spp., which cluster genetically within the *E. coli* tree but have been named differently mostly for historical and clinical reasons, both beg the question as to how the *E. coli* species should be defined, and where its boundaries lie. It also

makes genomic analysis difficult, as such a large amount of variation and mobility pushes the currently available software and hardware resources to their limits.

2.5.2. *KLEBSIELLA PNEUMONIAE*

For *Klebsiella* spp., the available data on population structure are more limited, although research interest is growing as the clinical significance of this pathogen in resource-rich settings increases, particularly in the context of its association with carbapenemases such as *Klebsiella pneumoniae* Carbapenemase (bla_{KPC}). One of the earliest detailed studies published used RAPD combined with sequencing of *gyrA* and *parC* genes to analyse a set of 120 *Klebsiella* spp. clinical and reference strains obtained from a number of hospitals across Europe and the Mediterranean region, and stratified by ciprofloxacin resistance. The results reflected significant diversity, with 70 genotypes identified amongst 86 *K. pneumoniae* isolates, and all of the profiles for the 30 *K. oxytoca* and 4 *K. planticola* isolates being distinct. Conversion of the data into a distance matrix and clustering analysis revealed that the *K. pneumoniae* strains fell into three major clusters: KpI, KpII and KpIII. KpI contained the largest number of isolates. The *K. oxytoca* isolates fell into two clusters, and the *K. planticola* isolates into a single group. Two isolates remained disassociated from any of the major clusters, and there was no correlation between the clusters and geographical location or clinical origin of the sample(106). The associated sequence-based analysis of the two genes determining quinolone resistance (*gyrA* and *parC*) was consistent with the RAPD clustering, and there was statistical support for the phylogeny even when the codons associated with acquisition of resistance were removed from the analysis.

A subsequent similar analysis (RAPD + partial sequence-based typing) of a larger (n=420) dataset of clinical *K. pneumoniae* only again revealed the three-group

structure and lack of geographical structuring, but did find significant differences in relation to clinical source (KpIII never being found in urinary tract isolates), and noted a significant association with ceftazidime (possibly ESBL) resistance with KpI versus KpIII (22% versus 9%; p=0.01). A third study investigating 100 animal isolates again found similar distributions of strains and clustering in the same groups, but noted that broad-spectrum antimicrobial resistance appeared less common than in human strains(107).

An MLST scheme developed for *K. pneumoniae*(26) and hosted at the Institut Pasteur, was subsequently used to describe the population structure in greater detail, with a particular focus on identifying whether there might be specific clones associated with virulence traits(12). Despite a low proportion of sites across the 7 MLST genes being polymorphic (129/3012 positions, 4%), this analysis of 235 isolates identified 117 STs, with a network structure based on concatenated sequence suggestive of high-levels of recombination, and evidence that this had even occurred within two of the seven MLST housekeeping genes examined (*tonB* and *phoE*). Additional support for high levels of recombination was provided by the observation that ST profiles with only a single locus difference frequently had a large number of mutations present in that locus (e.g. 4-6), which would be more than would be expected by introduction through mutation, and that the same capsular type (*cps* operon) was shared amongst isolates in genetically divergent lineages. Nonetheless, they did observe the presence of several clones with similar features (e.g. capsular serotype, virulence factor content and metabolic profile); the extreme cases being represented by *K. rhinoscleromatis* and *K. ozaenae*, which were shown to be clones of *K. pneumoniae* subsp. *K. pneumoniae*.

More recent attempts at defining population structure have involved the use of WGS data, but have focused specifically on descriptions of a KPC-carbapenemase associated clone, ST258, in the US. Of note are two studies, one in which 83 ST258 genomes were sequenced, and two distinct clades within ST258 identified, mostly divergent as the result of a specific 215Kb region encompassing the capsular *cps* operon. The authors hypothesised that this region was acquired on two different occasions given the different distribution of non-synonymous:synonymous SNVs within this region in the different clades(108). The analysis also partially characterised the KPC-variants and plasmid content of the strains, presenting a complicated, mixed picture, which suggested both clonal expansion in conjunction with certain plasmids, and horizontal transfer of other plasmids/parts of plasmid.

A second analysis of a similar dataset of 57 US ST258 strains corroborated the presence of these two clades within ST258, and identified the same region of divergence by a different method(109). The authors also carried out a more detailed analysis of KPC, Tn*4401* and plasmid structure, although the conclusions were similar – various KPC alleles and Tn*4401* isoforms were strongly associated with clade and plasmid sub-structure, although variability in some of the plasmids was also evident.

No equivalent assessment of comparisons of core and pan-genome to that for *E. coli* (Figure 2.4.) (102) have been published for *K. pneumoniae*, however early, unpublished work predating this thesis suggests that a pattern similar to *E. coli* will be evident. An analysis of 14 sequenced genomes using a non-scalable method identified a decreasing core size with the addition of each new sample in the analysis, and a similarly "open" structure to the pan-genome, as observed for *E. coli.*

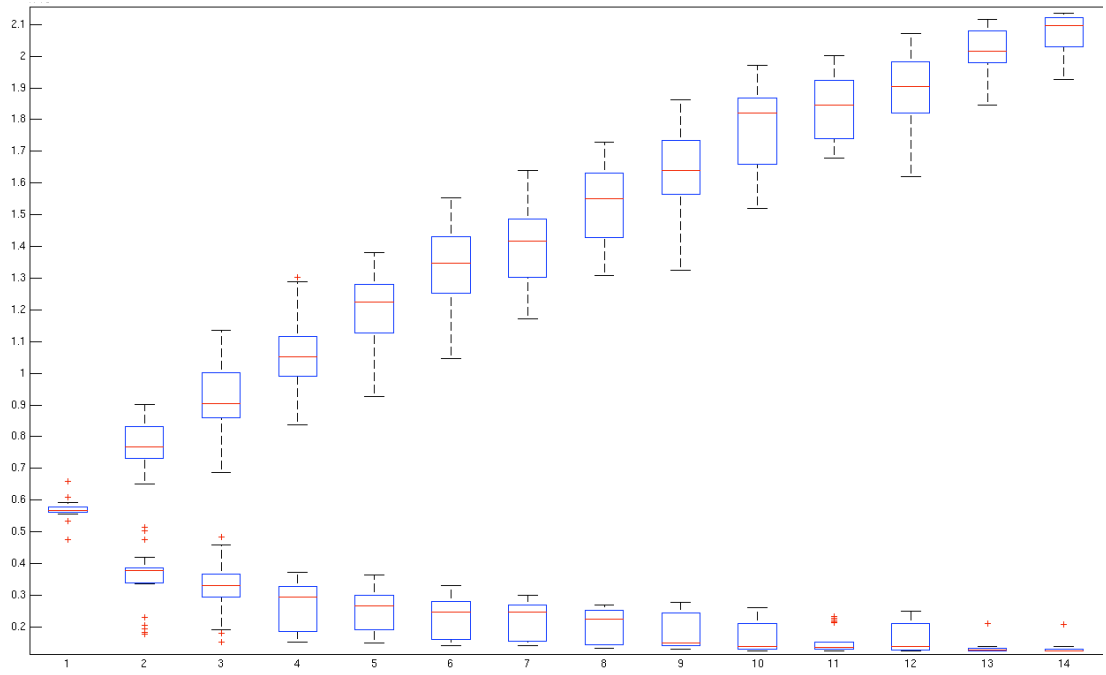


Figure 2.5. A representation of *K. pnuemoniae* core (lower set of boxplots) and pangenome (upper set of boxplots) in relation to the number of sequences added to the comparison. I am grateful to Dr Azim Ansari for the figure (unpublished work).

2.6. OVERVIEW OF ANTIMICROBIAL RESISTANCE IN NON-ENTEROPATHOEGNIC *ESCHERICHIA COLI* AND *KLEBSIELLA PNEUMONIAE*, WITH A FOCUS ON BROAD-SPECTRUM BETA-LACTAM RESISTANCE

Genetic mechanisms encoding for antimicrobial resistance are known to predate the clinical use of antibiotics(110). Many of the major antimicrobial drug classes in use today are derived from microbiota (e.g. penicillin, streptomycin), and are produced by these organisms as defence mechanisms against other species. Rates of resistance in clinical isolates however, have increased dramatically over the last 40 years, typically in a step-wise fashion in response to the sequential introduction of novel antimicrobial classes (Figure 2.6.).

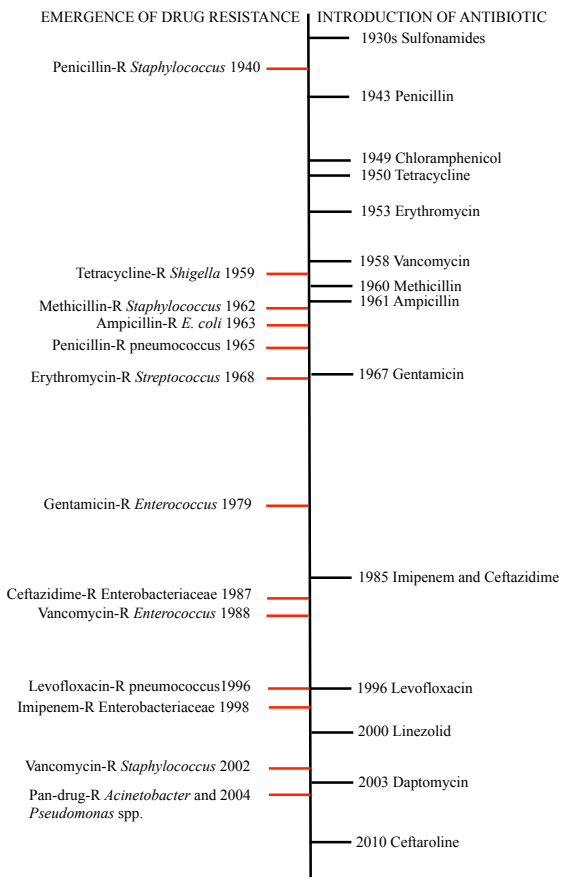


FIGURE 2.6. Timeline of clinical introductions of antibiotics and emergence of resistance.

The discovery and development of novel classes of antibiotics has however slowed since the 1960s, mostly attributable to the perceived lack of economic benefit identified in the development of these drugs(110). Recently, given the rapid emergence of antimicrobial resistance in common clinical pathogens and concern that we may be facing a return to the pre-antibiotic era in infectious disease management, there has been a big drive to encourage the development of new agents. An example of this is the "10x '20" (10 new antibiotics by 2020) initiative led by the Infectious Diseases Society of America. Unfortunately, although some progress has been made, their 2013 progress report highlights the continuing slow progress in this field(111).

A number of mechanisms responsible for causing antibiotic resistance exist, and include (i) enzymatic inactivation or modification of antibiotics, (ii) alteration in the

bacterial target site for a particular drug class, (iii) permeability barriers to the influx of antibiotics, (iv) efflux pumps that actively extrude antibiotics from the bacterial cell(112), and (v) combinations of the above mechanisms, which can extend the spectrum of resistance phenotypes. Beta-lactam resistance in Gram-negative bacilli is typically associated with enzymatic mechanisms, including narrow-spectrum penicillinases or cephalosporinases, or broad-spectrum mechanisms such as extended-spectrum beta-lactamases or carbapenemases, with some of the latter being capable of hydrolysing all classes of beta-lactams. Additional mechanisms which can enhance the enzymatic activity of beta-lactam-hydrolysing enzymes include: (i) alterations in porin genes(113), (ii) modification of drug efflux systems(114), (iii) promoter/attenuator mutations which can increase the expression of enzymes(115), and (iv) gene duplications or variation in copy number(116). Underlying genotypes can consist of a number of different mechanisms encoding for similar or overlapping phenotypes(117), highlighting the limitations of our knowledge on fitness costs associated with resistance, and the extent of its reversibility in the absence of continued exposure(118, 119).

Hydrolytic enzymes that degrade beta-lactams are encoded chromosomally in certain species of *Enterobacteriaceae*, such as *bla*$_{\text{SHV}}$ in *K. pneumoniae*, *bla*$_{\text{OXY}}$ in *Klebsiella oxytoca*, or *bla*$_{\text{KLUG}}$ in *Kluyvera georgiana*. A number of these mechanisms have also become mobilised on transmissible genetic elements, for example plasmids, such as *bla*$_{\text{SHV}}$ (the same gene as can be chromosomally integrated as above), *bla*$_{\text{CMY-2}}$ and *bla*$_{\text{CTX-M}}$, facilitating both intra- and inter-species spread. This horizontal transfer of resistance genes is then typically thought to be associated with wider dissemination and increasing prevalence(120).

Beta-lactam antibiotics were first widely introduced into clinical practice in the 1940s, and resistance was observed in *Enterobacteriaceae* soon after. The beta-lactamase TEM-1 ($bla_{\text{TEM-1}}$; TEMoneira-1) is a plasmid-mediated, narrow spectrum beta-lactamase first observed in a non-enteropathogenic *E. coli* isolate in Greece in 1965. Since then, following the introduction of ampicillin in the 1960s and amoxicillin in the 1970s, it has become hugely prevalent amongst ampicillin/amoxicillin-resistant non-enteropathogenic *E. coli* and has given rise to over 200 variants, some of which have an extended spectrum of hydrolysis encompassing beta-lactam/beta-lactamase inhibitor combinations(121, 122). In the 1980s, following the introduction of cephalosporins, the first plasmid-mediated SHV, extended-spectrum TEM and CTX-M beta-lactamase enzymes were observed. To date, a large number of diverse genetic families conferring beta-lactam resistance have been identified(120, 123, 124), with transmissible carbapenemases(125-127) emerging as the most recent healthcare threat. With these enzymes, which are often associated with resistance mechanisms to other antimicrobial classes, current antibiotics are in some cases rendered completely ineffective(128).

Beta-lactamases are conventionally classified into four categories on the basis of structure and amino acid homology (Ambler classification) (Figure 2.7.), or on the basis of substrate specificity and function (Bush-Jacoby-Medeiros classification) (Table 2.8.)(129, 130). Essentially two main mechanisms of action exist – either the use of zinc ions to disrupt the beta-lactam ring (the metallo-beta-lactamases, group B), or through the formation and subsequent hydrolysis of covalent bonds with the antibiotic through a free hydroxyl group on a serine residue, which leads to antibiotic inactivation (the serine beta-lactamases, groups A, C and D). In Gram-negative species, these enzymes are typically located in the periplasmic space(131), and their

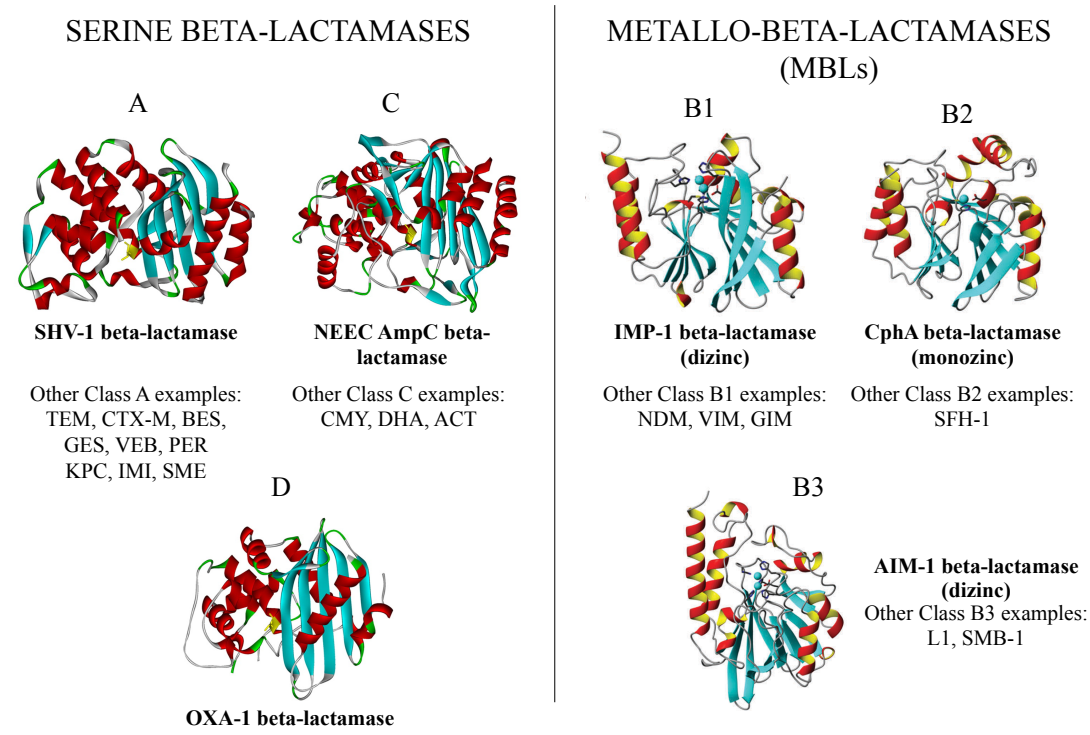


Figure 2.7. Major structural sub-classes of beta-lactamase. Adapted from (132, 133).

Summary of classification schemes for bacterial beta-lactamases

Bush-Jacoby group (2009)	Bush-Jacoby-Medeiros group (1995)	Molecular class (subclass)	Distinctive substrate(s)	Inhibited by		Defining characteristic(s)	Representative enzyme(s)
				CA or TZB[a]	EDTA		
1	1	C	Cephalosporins	No	No	Greater hydrolysis of cephalosporins than benzylpenicillin; hydrolyzes cephamycins	*E. coli* AmpC, P99, ACT-1, CMY-2, FOX-1, MIR-1
1e	NI[b]	C	Cephalosporins	No	No	Increased hydrolysis of ceftazidime and often other oxyimino-β-lactams	GC1, CMY-37
2a	2a	A	Penicillins	Yes	No	Greater hydrolysis of benzylpenicillin than cephalosporins	PC1
2b	2b	A	Penicillins, early cephalosporins	Yes	No	Similar hydrolysis of benzylpenicillin and cephalosporins	TEM-1, TEM-2, SHV-1
2be	2be	A	Extended-spectrum cephalosporins, monobactams	Yes	No	Increased hydrolysis of oxyimino-β-lactams (cefotaxime, ceftazidime, ceftriaxone, cefepime, aztreonam)	TEM-3, SHV-2, CTX-M-15, PER-1, VEB-1
2br	2br	A	Penicillins	No	No	Resistance to clavulanic acid, sulbactam, and tazobactam	TEM-30, SHV-10
2ber	NI	A	Extended-spectrum cephalosporins, monobactams	No	No	Increased hydrolysis of oxyimino-β-lactams combined with resistance to clavulanic acid, sulbactam, and tazobactam	TEM-50
2c	2c	A	Carbenicillin	Yes	No	Increased hydrolysis of carbenicillin	PSE-1, CARB-3
2ce	NI	A	Carbenicillin, cefepime	Yes	No	Increased hydrolysis of carbenicillin, cefepime, and cefpirome	RTG-4
2d	2d	D	Cloxacillin	Variable	No	Increased hydrolysis of cloxacillin or oxacillin	OXA-1, OXA-10
2de	NI	D	Extended-spectrum cephalosporins	Variable	No	Hydrolyzes cloxacillin or oxacillin and oxyimino-β-lactams	OXA-11, OXA-15
2df	NI	D	Carbapenems	Variable	No	Hydrolyzes cloxacillin or oxacillin and carbapenems	OXA-23, OXA-48
2e	2e	A	Extended-spectrum cephalosporins	Yes	No	Hydrolyzes cephalosporins. Inhibited by clavulanic acid but not aztreonam	CepA
2f	2f	A	Carbapenems	Variable	No	Increased hydrolysis of carbapenems, oxyimino-β-lactams, cephamycins	KPC-2, IMI-1, SME-1
3a	3	B (B1)	Carbapenems	No	Yes	Broad-spectrum hydrolysis including carbapenems but not monobactams	IMP-1, VIM-1, CcrA, IND-1
		B (B3)					L1, CAU-1, GOB-1, FEZ-1
3b	3	B (B2)	Carbapenems	No	Yes	Preferential hydrolysis of carbapenems	CphA, Sfh-1
NI	4	Unknown					

Table 2.8. Functional classification of beta-lactamases. Taken from (129).

kinetics are dependent on a combination of affinity for the substrate, the maximum rate of hydrolysis they are capable of, and their concentration. The latter will depend on whether they are constitutively expressed at a relatively fixed rate and concentration, or whether they are inducible, in which case resistance will depend on how effectively a substrate induces enzyme expression, and how labile it is to enzymatic action(131). Enzymes can also be influenced by physiochemical conditions, which may be important in the context of phenotyping (hence the use of standard susceptibility testing media in routine practice), and by the interplay of other factors. For *Enterobacteriaceae*, the rates of diffusion/efflux of beta-lactam into the periplasmic space may affect the enzyme kinetics of the beta-lactam itself, and would explain the rationale for giving combination therapy of, say colistin to disrupt the outer membrane, with a beta-lactam antibiotic, even in the presence of beta-lactam resistance(131).

2.7. THE GENETIC HIERARCHY OF ANTIMICROBIAL RESISTANCE IN NON-ENTEROPATHOGENIC *E. COLI* AND *K. PNEUMONIAE*: FROM GENES TO HOST STRAINS

2.7.1. RESISTANCE GENES

The simplest genetic mechanism of resistance is represented by the presence of a single gene or coding sequence giving rise to an alteration in susceptibility phenotype. The number of resistance gene families encoding for resistance to beta-lactams in non-enteropathogenic *E. coli* and *K. pneumoniae* is substantial, with over 100 gene families identifiable from public resources/publications. The nomenclature of these gene families can be particularly confusing, with similar names being given to genetically divergent enzymes, with different phenotypic properties.

The studies in this thesis are restricted to an investigation of the epidemiology of non-enteropathogenic *E. coli* and *K. pneumoniae* containing a number of increasingly common and important clinical resistance threats, including New Delhi Metallo-beta-lactamase (bla_{NDM}) and *Klebsiella Pneumoniae* Carbapenemase enzymes (bla_{KPC}), in this case in *K. pneumoniae* (Chapters 6 and 7), and the CefoTaxiMase family of ESBL enzymes (bla_{CTX-M}) in non-enteropathogenic *E. coli* (Chapter 8). These enzymes are therefore discussed in more detail below.

2.7.2. BETA-LACTAMASE ENZYMES OF PARTICULAR CLINICAL IMPORTANCE

2.7.2.1. NDM CARBAPENEMASES

The NDM class B carbapenemase is encoded for by the bla_{NDM} gene, a nucleotide sequence of 813 base-pairs (bp). It confers resistance to all broad-spectrum beta-lactamases except aztreonam, although it is frequently found with other enzymes, such as ESBLs, that will also preclude the successful clinical use of aztreonam. To date 11 enzymatic variants have been catalogued, with 14 variant sites distinguishing them from each other(121). It was first identified in 2008 in *K. pneumoniae* and non-enteropathogenic *E. coli* isolates cultured from a Swedish patient with a history of travel to and hospitalisation in India(134), since then it has been identified from a wide range of global locations, with geographic "hotspots" of acquisition considered to be India, Pakistan, Bangladesh, the Balkans and the Middle East (Figure 2.9(135)).

Figure 2.9. Countries in which NDM-positive isolates have been reported. Triangles indicate cases which have an epidemiological link to the Indian sub-continent. Taken from (135).

Although it was first identified in 2008, a retrospective analysis of isolates collected from three Indian hospitals has revealed its presence in *Enterobacteriaceae* as early as 2006(136). In the earlier case reports and series, isolation of NDM-1 positive bacteria was frequently associated with travel to the Indian sub-continent, but in recent years it has been isolated from a number of individuals with no history of foreign travel, suggesting local acquisition and dissemination(135). This would be consistent with its observed presence in a seepage/water samples in handful of environmental studies(137, 138), and with the fact that asymptomatic faecal carriage with NDM-positive bacteria can be prolonged(135).

2.7.2.2. KPC CARBAPENEMASES

The *bla*$_{KPC}$ gene is a 918bp sequence, encoding for the class A KPC carbapenemase, which hydrolyses all broad-spectrum beta-lactams including aztreonam. To date 17 sequences have been catalogued(121), with 18 variable sites between them (excluding sequences which have missing information and *bla*$_{KPC-14}$, which contains a deletion).

KPC-2 and KPC-3 are the most common variants, and are endemic in a number of regions, including the US, Greece and Israel (Figure 2.10.). Here, the presence of carbapenem resistance is typically attributed to the widespread presence of KPC genes, typically thought to be associated with a single clinically "successful" *K. pneumoniae* lineage, namely sequence type (ST) 258, as mentioned previously, and members of the same clonal complex.

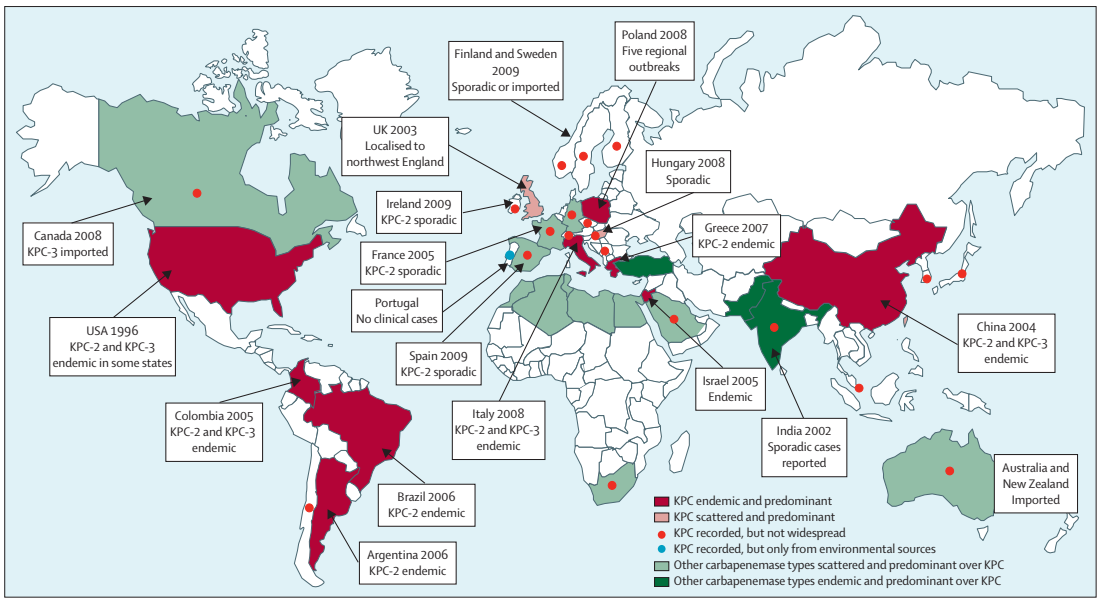


Figure 2.10. Global distribution of KPC variants and other carbapenemases. Taken from (139).

Although the number of infections is relatively low in the US, in Greece, up to 40% of all hospital-associated bloodstream *K. pneumoniae* isolates are KPC positive, and the organisms are found widely throughout all healthcare locations. In Israel, long-term care facilities have been recognised as a major potential reservoir(139).

2.7.2.3. CTX-M EXTENDED SPECTRUM BETA-LACTAMASES

The CTX-M enzymes were originally reported in the second half of the 1980s and are the most widely distributed ESBLs in non-enteropathogenic *E. coli* and *K.*

pneumoniae. They are class A enzymes, which hydrolyse oxyiminocephalosporins, but are susceptible to carbapenems and beta-lactam/beta-lactamase inhibitor combinations in vitro. The number of variants has increased dramatically over the last 25 years; currently there are approximately 150 variants catalogued(121). These variants are sub-clustered into families, which share less then 90% sequence similarity between them, and a couple of recombinant hybrids which fall between phylogenetic groups. They are thought to have been introduced into *Enterobacteriaceae* from several *Kluyvera* species (Figure 2.11.(140)).

The most common global variants are CTX-M-15 (family 1 in Figure 2.11.) and CTX-M-14 (family 9 in Figure 2.11.), or single nucleotide variants thereof, such as CTX-M-55/57 (actually representing the same nucleotide sequence) or CTX-M-27, respectively (141). Regional differences in distribution of common variants exists, with CTX-M-14-like variants predominating in South-East Asia, China and Taiwan, and CTX-M-15 in the Indian subcontinent, for example (Figure 2.12.). The genetic context of CTX-M genes is discussed in greater detail in the subsequent section.

Figure 2.11. Neighbour-joining tree of the 876 bp CTX-M gene. Drawn from variants catalogued up to June 2012. Putative source *Kluyvera* species are labelled on the tree. Scale bar represents 26 nucleotide variants.

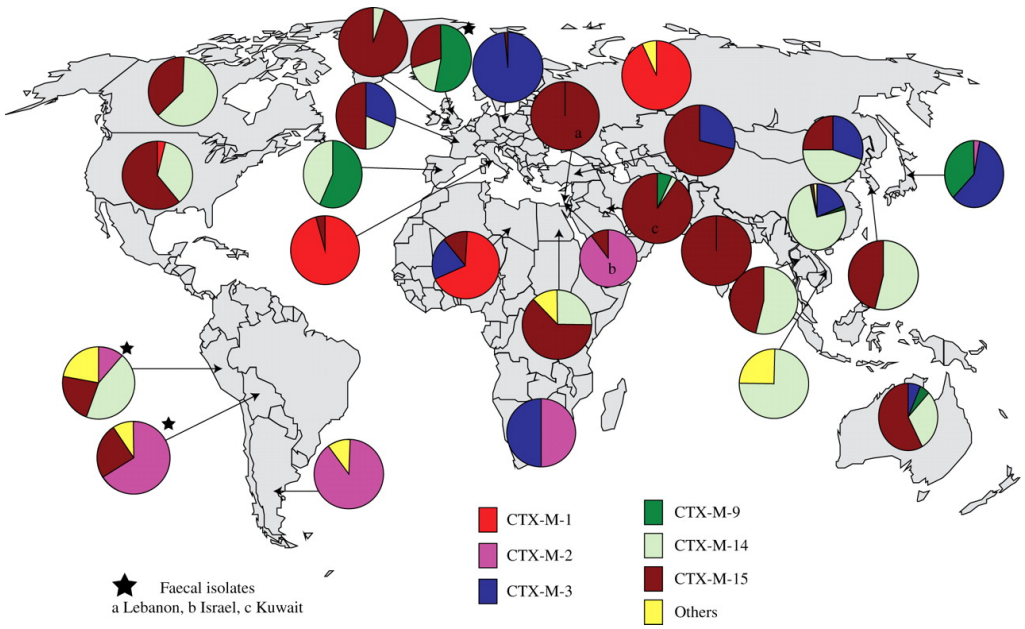


Figure 2.12. Global distribution of major CTX-M variants, 2009. Taken from (141).

2.7.3. INSERTION SEQUENCES (IS) AND TRANSPOSONS

Insertion sequences (IS) are the simplest transposable elements; they are types of DNA transposons (Class II). These can be loosely defined as segments of DNA <2.5Kb in size which are capable of inserting at multiple sites in a DNA molecule, be this the bacterial chromosome or an episomal element such as a plasmid. Again, as with resistance genes, the nomenclature can be confusing, with either a simple numeric assignation to the IS (e.g. IS*1*), or a numeric assignation plus some sort of text denoting the species in which it was first isolated (e.g. IS*Rm1*). Sometimes genetically similar ISs are designated as isoforms of a parent IS, and in some cases they have been given a separate number. Finally, some ISs have been assigned names that do not fit into this system of nomenclature(142).

Most ISs essentially represent a pair of inverted repeat sequences (IRL – inverted left repeat; IRR – inverted right repeat) of 10-40 base-pair (bp) flanking an open reading frame encoding a transposase enzyme. These inverted repeats include two functional domains: Domain A includes the 2-3 terminal bp, and is involved in cleavage and transfer reactions leading to transposition of the element; domain B is involved in transposase binding. On integration, most typically generate a series of short, directly repeated sequences (direct target repeats, 2-14bp) characteristic of each individual IS; these can be used as signatures of IS integration.

ISs are involved in the mobilisation and transfer of a number of different antimicrobial resistance genes, virulence genes and genes encoding metabolic functions, as well as being responsible for genetic rearrangements within chromosomes and plasmids, and in the integration of plasmids or parts of plasmids into the chromosome(142). They can activate or inactivate the expression of

neighbouring genes – the latter mostly occurring if they are inserted into a coding sequence itself(142). A large number of ISs have been catalogued in the ISFinder dataset(143) (https://www-is.biotoul.fr//), which has identified 26 major families.

Several associations of IS types with the broad-spectrum beta-lactamases being investigated in this thesis have been described. For example, it has been hypothesised that *bla*$_{NDM}$ was acquired horizontally from *Acinetobacter* spp., where it is mostly integrated in chromosomal locations downstream of a particular insertion sequence IS*Aba125*. A bleomycin resistance gene, *ble*$_{MBL}$, which can be found in the environment, is commonly co-located with *bla*$_{NDM}$. One hypothesis put forward is that the *bla*$_{NDM}$-*ble*$_{MBL}$ genes became integrated into the chromosome of an environmental *Acinetobacter* spp., and were then transposed onto a plasmid capable of replication and conjugative transfer amongst *Enterobacteriaceae*(144). Although the association with IS*Aba125* and *ble*$_{MBL}$ remains in many cases, NDM genes are now found on a wide range of plasmid vectors and have been reported in at least 10 different species(135). The current evidence mostly supports high mobilisation and transfer rates of the gene in a diverse range of genetic contexts(135).

IS*Ecp1* elements have been commonly found in conjunction with CTX-M genes from the group 1, 2 and 9 families. The successful spread of common variants of these families (CTX-M-15 and CTX-M-14) may in part be attributable to their association with this IS, as it is able to recognise a variety of DNA sequences as IRRs and mobilise DNA via a one-ended transposition process. It also significantly enhances CTX-M gene expression on mobilisation(145). A second common IS identified in association with CTX-M is IS*CR1*, which unlike most other ISs lacks the inverted repeats and mobilises instead via a system known as rolling circle transposition. This

mechanism can also lead to the one-ended mobilisation of adjacent genetic sequence. Like IS*Ecp1*, it also enhances gene expression.

Composite transposons consist of two ISs flanking an intervening stretch of DNA sequence – this may encode one or more genes, such as those encoding for antibiotic resistance. The flanking ISs may be identical or different, and can either mobilise the complete composite transposon structure, or be independently involved in transposition events as single ISs. These have been particularly associated with the transmission of several carbapenemases of interest, namely OXA-48 and KPC variants. OXA-48, for example, was first identified in 2004 as part of Tn*1999*(146), a composite transposon made up of two copies of IS*1999*. Since then, a number of Tn*1999* derivatives have been identified, one of which also contains the CTX-M-15 gene(147). KPC variants have been consistently isolated in the context of Tn*4401*, a large composite transposon (~10kB), made up of two *TnpA* genes (encoding for Tn*3* family transposons), a *TnpR* resolvase, and two insertion sequences, IS*Kpn6* and IS*Kpn7*. A number of different Tn*4401* "isoforms" have been described, on the basis of different deletions upstream of the KPC gene, with different effects on gene expression and carbapenem minimum inhibitory concentrations (MICs)(148, 149).

2.7.4. PLASMIDS

A plasmid is defined as a double-stranded, circular DNA molecule capable of autonomous replication, and contributes to the dissemination of resistance genes within and between species through a process known as conjugation. Conjugation involves the genetic exchange of material between two bacterial cells, and is a process undertaken by genes of plasmid conjugation modules, such as *tra* and *trb* genes.

Classification of plasmids has traditionally been based on a concept known as plasmid incompatibility, put forward in 1971(150). This was defined as the inability of two related plasmids, with common replication controls, to be maintained stably in the same cell line. Incompatibility group (Inc) typing is therefore based on the underlying nucleotide sequence of the genes (*rep*) involved in the initiation of replication. The first proper typing scheme was based on a Southern blot hybridisation method(151); this has now largely been superseded by more specific and straightforward PCR-based methods(152). One of these is a set of multiplex PCRs that defines 18 of the major families, namely: FIA, FIB, FIC, HI1, HI2, I1-Iγ, L/M, N, P, W, T, A/C, K, B/O, X, Y, F, and FIIA. There are however at least 27 known incompatibility groups, and a number of plasmids remain untypable in epidemiological surveys, suggesting that most studies represent only a subset of the real diversity(153). Plasmid incompatibility or *rep* controls are also associated with plasmid host range, namely the number of divergent species that a plasmid is able to transfer to given its specific conjugation and transfer modules. IncF plasmids, for example have been traditionally thought to have a narrow host range; whereas IncN plasmids have a broad host range(154).

NDM genes have been found on a range of plasmid vectors in both non-enteropathogenic *E. coli* and *K. pneumoniae*, including those of the A/C, FII, and L/M incompatibility groups(135). Interestingly, a relatively large proportion of NDM plasmids appear to be of the untypable variety, which may either reflect the emergence of new families of plasmids as significant clinical entities, or may just be a feature of increased interest and sampling. NDM-carrying plasmids vary significantly in size (from 50kB to ~300kb), and are commonly associated with a number of other

antimicrobial resistance genes, including CTX-M genes, aminoglycoside resistance genes, and quinolone resistance genes(135).

KPC genes are also associated with a number of different plasmids, including the IncFII-like plasmids pKpQIL in Israel (KPC-3), and pKpQIL-IT (KPC-3) in Italy, both associated with the ST258 lineages, and both having similarity to plasmids found in a previously sequenced antibiotic-resistant (although KPC negative) reference strain from the 1990s, MGH78578(155, 156). WGS-based analyses of ST258 isolates in the US have demonstrated additional complexity, with a within-lineage segregation of IncFIIk-based (and Tn*4401*a/KPC-2) KPC carriage in a sublineage of ST258 – ST258a; and IncI-based (Tn*4401*b/KPC-3) carriage in sublineage ST258b(109). Three distinct KPC-positive FIIk plasmids were identified, in two cases associated with marked plasmid backbone differences compared to pKpQIL(109).

CTX-M gene variants have been found in association with almost the complete range of plasmid Inc types, although CTX-M-15 is most often found in plasmids belonging to the IncF group(153). This plasmid family is typically present in low copy number (that is, regulated to maintain 1-5 plasmid copies per bacterial cell), and its members also frequently carry more than one replicon. Plasmids with multiple replicons are likely co-integrates of two separate descendant plasmids and may have an evolutionary advantage, with one replicon strongly conserved to maintain replication integrity, whilst the other replicon is free to diverge(157). A plasmid with a novel *rep* variant would be better able to invade local bacterial populations already occupied with other plasmids of ancestral incompatibility types(157). IncF plasmids are also associated with Tn*3*-associated TEM-1 genes, and the Tn*3* transposase is thought to

act as a site for integration of the IS*Ecp1*/CTX-M element(153), which may have facilitated its acquisition and dispersal by this specific group of plasmids.

2.7.5. CLONAL EPIDEMIC LINEAGES

The host bacterial lineage may also contribute to the spread of resistance, as some strains may be fitter and/or more clinically significant, either because they are associated with significant antimicrobial resistance, or because they are more virulent, or both. "Successful" global lineages are effective transmission vectors for resistance genes because these are typically transmitted vertically to progeny, and because there are more opportunities to transfer resistance genes horizontally intra- and inter-species(158).

Amongst non-enteropathogenic *E. coli*, several drug-resistant clones are recognised, most notably ST131, which is the subject of one of this thesis' sub-studies (chapter 8), and has emerged in association with CTX-M-15, and also CTX-M-14-like variants in Eastern and South-eastern Asia(159, 160). ST405, ST10, ST38 and ST648 are further examples (Figure 2.13.(161, 162)).

For *K. pneumoniae*, the predominant lineage described in association with significant clinical resistance is ST258, in conjunction with bla_{KPC}(158). Recent evidence suggests that ST258 has emerged as a hybrid strain from a large recombination event between ST11 and ST442(163). ST11 itself is also a drug-resistant, clinically common clone, and is has been identified in a number of epidemiological surveys with a variety of carbapenemases, including outbreaks in Brazil, China, Greece and Spain(164-167). Other important emerging *K. pneumoniae* lineages include ST15 and ST147(168-170).

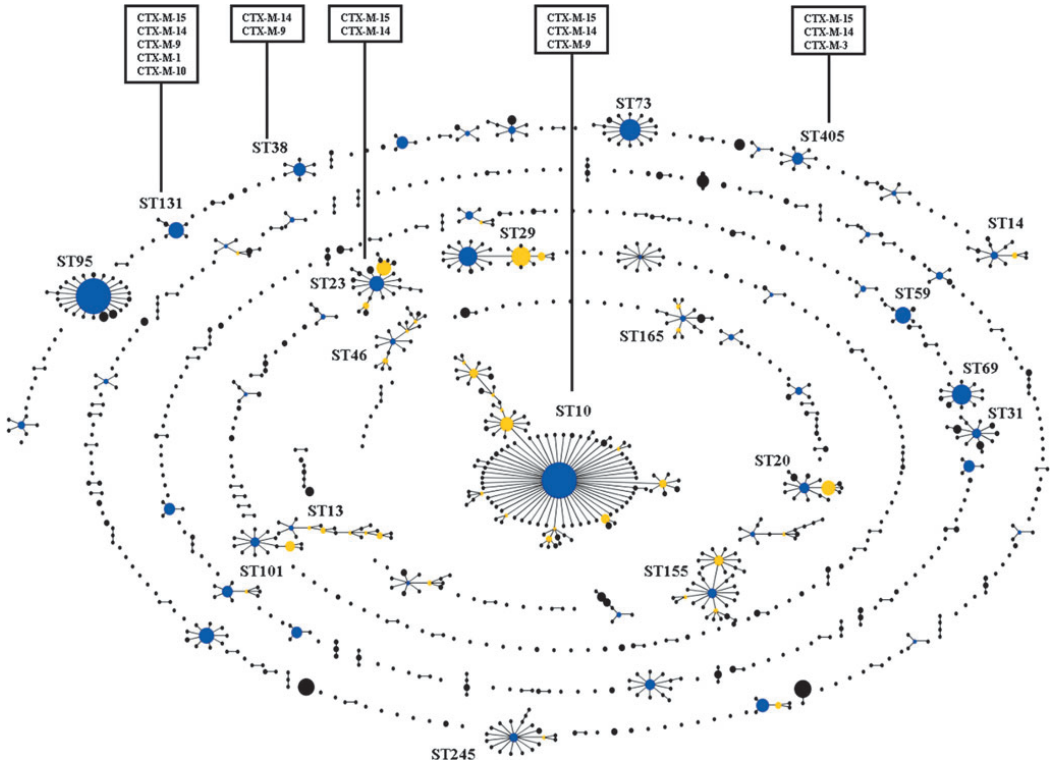


Figure 2.13. Eburst population snapshot of non-enteropathogenic *E. coli*; major clones and associations with CTX-M genes are highlighted. Taken from (161).

2.7.6. CHROMOSOMAL INTEGRATION OF RESISTANCE GENES

The genetic location of resistance genes is an important consideration in the analysis of their evolution and transmission, as mobile-element associated genes can be horizontally transferred between isolates, but may also be lost in bacterial fission and therefore be absent in offspring. Chromosomally integrated genes however will typically be passed on to descendants, and are potentially more likely to become part of the core genetic repertoire of that lineage. The clinical implications of the chromosomal integration of broad-spectrum beta-lactmases in non-enteropathogenic *E. coli* and *K. pneumoniae* are therefore significant.

One of the earliest descriptions of the chromosomal integration of CTX-M-15 in non-enteropathogenic *E. coli* was in a small collection of isolates sampled from 2000-

2006(160). Here, of 43 strains investigated, 8 were shown to be chromosomally located, and in a number of different host backgrounds (2 ST131 strains, 2 ST405, and 3 to other members of the D and A phylogroups). Two strains had evidence for both a plasmid and chromosomal location for the CTX-M gene. This phenomenon was also observed in a number of non-enteropathogenic *E. coli* strains in association with CTX-M-14 in isolates from Korea in 2005(171). Here, of 35 isolates, 8 showed evidence of a chromosomal location for the CTX-M-14 gene and 5 showed evidence for the concomitant location of CTX-M-14 in both the chromosome and on IncF plasmids. Again this was found in a wide range of host backgrounds, represented in 8 STs, most of which are recognised as clinically very important lineages, such as ST131 and ST405.

A recent study went further in specifically investigating the transposition units involved in, and the locations of, chromosomally integrated CTX-M-15 in ST131 in Japan. Although all of the chromosomally integrated CTX-M-15 were flanked by the same IS*Ecp1*–ORF477 element, they were identified in three different chromosomal regions with small, additional differences in genetic rearrangements observed at these sites for individual cases(172). The relative frequency, restrictions on location of integration, and stability and duration of the chromosomal location, of these genes remain to be elucidated.

For *K. pneumoniae*, the first observation of a chromosomal location for CTX-M was in a set of isolates sampled in Barcelona, Spain, between 2005-2008(173). In this case, CTX-M-15 was found located chromosomally, but only in association with a single ST, ST1.

Chromosomal integration of CTX-M is also not restricted to human isolates. Again recent data show that CTX-M-2 has been found chromosomally integrated in 19 isolates downstream of an IS*CR*-element in healthy chickens from two broiler farms in Brazil(174); these isolates represented three different STs from three different phylogroups, again suggesting that this is not a lineage-restricted phenomenon.

The chromosomal integration of CTX-M-15 has also been described in a number of other *Enterobacteriaceae*. In a study of *Salmonella* Concord isolates this was also seen as early as 2001(175), with the authors hypothesising that this came about as a result of IS*26*-mediated transfer from CTX-M-15 containing plasmid structures to the chromosome. CTX-M variants associated with CTX-M families less commonly found in non-enteropathogenic *E. coli* and *K. pneumoniae*, namely the -2 and -25, families, have been observed in a couple of studies in *Proteus mirabilis*, all in association with IS*Ecp1* elements(176-178). The implications are that the evolution of and selection pressures on these resistance genes may be acting in a number of different species and environments, and that transmission networks are likely to be vast.

2.8. THESIS OUTLINE

The overarching goal of this thesis is to harness the additional genetic resolution that WGS can provide over traditional typing formats to investigate resistance genes, flanking mobile elements, plasmid vectors and host bacterial strains using a unified, "one-stop" method. As I am a clinician in infectious diseases training, I chose to focus on beta-lactam resistance mechanisms since these present a major, current clinical threat to the management of infections; and on pathogens that contribute significantly to the burden of global disease, namely non-enteropathogenic *E. coli* and *K. pneumoniae*.

The first analysis chapter (chapter 5) discusses the initial experiment to ensure that I was able to identify relevant resistance genes of interest in clinical isolates, namely a pilot study to predict antimicrobial susceptibility phenotype from genotype. At the time the work was undertaken, it was considered that this would be too challenging for complex resistance genotypes in ram-negative organisms; this view is now mostly changing.

Chapters 6 and 7 focus on using WGS to characterise clinical outbreaks of *K. pneumoniae* in two different settings with two different carbapenemase resistance mechanisms, namely $bla_{\text{NDM-1}}$ and bla_{KPC}. These both present evidence that challenge the views of the epidemiology of these resistance mechanisms that prevailed at the time the work was being undertaken. For NDM, the historical view has been that the gene is extremely mobile, without much association with plasmid structures or host strains; for KPC, that this was typically considered strongly associated with a single global lineage of *K. pneumoniae*. My work, undertaken in collaboration with colleagues (see Attributions section in Chapter 1 – Preface), demonstrates that NDM can be maintained as part of a single plasmid population within a single lineage over relatively long period of time, and that KPC can be transmitted across multiple genetically divergent lineages of *K. pneumoniae* within in a single institution in a relatively short period of time.

Finally, I used WGS to investigate the population structure of ST131 non-enteropathogenic *E. coli* in a global sampling frame (chapter 8). Although there were two other research groups who used WGS to investigate aspects of the same research question, my work complements and expands on theirs(179, 180). I have the largest collection of sequenced ST131 strains in my analysis, have defined evolutionary rates,

have included non-CTX-M-15 associated ST131s in my analyses, and have a more substantially detailed assessment of the role played by plasmids – although I have by no means answered all the questions that remain.

CHAPTER 2 REFERENCES

1. **Pham HN, Ohkusu K, Mishima N, Noda M, Monir Shah M, Sun X, Hayashi M, Ezaki T.** 2007. Phylogeny and species identification of the family Enterobacteriaceae based on dnaJ sequences. Diagn Microbiol Infect Dis **58:**153-161.
2. **Didelot X, Bowden R, Wilson DJ, Peto TE, Crook DW.** 2012. Transforming clinical microbiology with bacterial genome sequencing. Nature reviews. Genetics **13:**601-612.
3. **Escherich T.** 1989. The intestinal bacteria of the neonate and breast-fed infant. 1885. Reviews of infectious diseases **11:**352-356.
4. **von Baumgarten P.** 1889. Jahresbericht ueber die Fortschritte in der Lehre von den pathogenen Mikroorganismen umfassend Bakterien, Pilze und Protozooen. **5**.
5. **Uhlenhuth.** Beitrag zur Pathogenität des Bacterium coli commune. Zeitschrift für Hygiene und Infektionskrankheiten.
6. **Croxen MA, Law RJ, Scholz R, Keeney KM, Wlodarska M, Finlay BB.** 2013. Recent advances in understanding enteric pathogenic Escherichia coli. Clinical microbiology reviews **26:**822-880.
7. **Kaper JB, Nataro JP, Mobley HL.** 2004. Pathogenic Escherichia coli. Nature reviews. Microbiology **2:**123-140.
8. **Russo TA, Johnson JR.** 2000. Proposal for a new inclusive designation for extraintestinal pathogenic isolates of Escherichia coli: ExPEC. The Journal of infectious diseases **181:**1753-1754.
9. **Kohler CD, Dobrindt U.** 2011. What defines extraintestinal pathogenic Escherichia coli? International journal of medical microbiology : IJMM **301:**642-647.

10. **Leimbach A, Hacker J, Dobrindt U.** 2013. E. coli as an all-rounder: the thin line between commensalism and pathogenicity. Current topics in microbiology and immunology **358:**3-32.

11. **Brisse SG, F.; Grimont, P.A.D.** 2006. The Genus Klebsiella. Prokaryotes **6:**159-196.

12. **Brisse S, Fevre C, Passet V, Issenhuth-Jeanjean S, Tournebize R, Diancourt L, Grimont P.** 2009. Virulent clones of Klebsiella pneumoniae: identification and evolutionary scenario based on genomic and phenotypic characterization. PloS one **4:**e4982.

13. **Glode MP, Sutton A, Robbins JB, McCracken GH, Gotschlich EC, Kaijser B, Hanson LA.** 1977. Neonatal meningitis due of Escherichia coli K1. The Journal of infectious diseases **136 Suppl:**S93-97.

14. **Tsay RW, Siu LK, Fung CP, Chang FY.** 2002. Characteristics of bacteremia between community-acquired and nosocomial Klebsiella pneumoniae infection: risk factor for mortality and the impact of capsular serotypes as a herald for community-acquired infection. Archives of internal medicine **162:**1021-1027.

15. **Fang CT, Lai SY, Yi WC, Hsueh PR, Liu KL, Chang SC.** 2007. Klebsiella pneumoniae genotype K1: an emerging pathogen that causes septic ocular or central nervous system complications from pyogenic liver abscess. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America **45:**284-293.

16. **Orskov F, Orskov I.** 1992. Escherichia coli serotyping and disease in man and animals. Canadian journal of microbiology **38:**699-704.

17. **Milkman R.** 1973. Electrophoretic variation in Escherichia coli from natural sources. Science **182:**1024-1026.

18. **Stanley TG, Wilson I.** 2003. Multilocus enzyme electrophoresis: a practical guide. Molecular biotechnology **24:**203-220.

19. **Clermont O, Bonacorsi S, Bingen E.** 2000. Rapid and simple determination of the Escherichia coli phylogenetic group. Applied and environmental microbiology **66:**4555-4558.

20. **Clermont O, Christenson JK, Denamur E, Gordon DM.** 2013. The Clermont Escherichia coli phylo-typing method revisited: improvement of specificity and detection of new phylo-groups. Environmental microbiology reports **5:**58-65.

21. **Foley SL, Lynne AM, Nayak R.** 2009. Molecular typing methodologies for microbial source tracking and epidemiological investigations of Gram-negative bacterial foodborne pathogens. Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases **9:**430-440.

22. **Blackwood RA, Rode CK, Pierson CL, Bloch CA.** 1997. Pulsed-field gel electrophoresis genomic fingerprinting of hospital Escherichia coli bacteraemia isolates. Journal of medical microbiology **46:**506-510.

23. **Meunier JR, Grimont PA.** 1993. Factors affecting reproducibility of random amplified polymorphic DNA fingerprinting. Research in microbiology **144:**373-379.

24. **Welsh J, McClelland M.** 1990. Fingerprinting genomes using PCR with arbitrary primers. Nucleic acids research **18:**7213-7218.

25. **Wirth T, Falush D, Lan R, Colles F, Mensa P, Wieler LH, Karch H, Reeves PR, Maiden MC, Ochman H, Achtman M.** 2006. Sex and virulence in Escherichia coli: an evolutionary perspective. Molecular microbiology **60:**1136-1151.

26. **Diancourt L, Passet V, Verhoef J, Grimont PA, Brisse S.** 2005. Multilocus sequence typing of Klebsiella pneumoniae nosocomial isolates. Journal of clinical microbiology **43:**4178-4182.

27. **Russo TA, Johnson JR.** 2003. Medical and economic impact of extraintestinal infections due to Escherichia coli: focus on an increasingly important endemic problem. Microbes and infection / Institut Pasteur **5:**449-456.

28. **Stamm WE, Norrby SR.** 2001. Urinary tract infections: disease panorama and challenges. The Journal of infectious diseases **183 Suppl 1:**S1-4.

29. **Foxman B.** 2003. Epidemiology of urinary tract infections: incidence, morbidity, and economic costs. Disease-a-month : DM **49:**53-70.

30. **Matthews SJ, Lancaster JW.** 2011. Urinary tract infections in the elderly population. The American journal of geriatric pharmacotherapy **9:**286-309.

31. **Wilson J, Guy R, Elgohari S, Sheridan E, Davies J, Lamagni T, Pearson A.** 2011. Trends in sources of meticillin-resistant Staphylococcus aureus (MRSA) bacteraemia: data from the national mandatory surveillance of MRSA bacteraemia in England, 2006-2009. The Journal of hospital infection **79:**211-217.

32. **Schlackow I, Stoesser N, Walker AS, Crook DW, Peto TE, Wyllie DH, Infections in Oxfordshire Research Database T.** 2012. Increasing incidence of Escherichia coli bacteraemia is driven by an increase in antibiotic-resistant isolates: electronic database study in Oxfordshire 1999-2011. The Journal of antimicrobial chemotherapy **67:**1514-1524.

33. **de Kraker ME, Jarlier V, Monen JC, Heuer OE, van de Sande N, Grundmann H.** 2013. The changing epidemiology of bacteraemias in Europe: trends from the European Antimicrobial Resistance Surveillance System.

Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases **19:**860-868.

34. **Control ECfDPa.** 2013. Annual Epidemiological Report 2013.

35. **Reddy EA, Shaw AV, Crump JA.** 2010. Community-acquired bloodstream infections in Africa: a systematic review and meta-analysis. The Lancet infectious diseases **10:**417-432.

36. **Deen J, von Seidlein L, Andersen F, Elle N, White NJ, Lubell Y.** 2012. Community-acquired bacterial bloodstream infections in developing countries in south and southeast Asia: a systematic review. The Lancet infectious diseases **12:**480-487.

37. **Liu L, Johnson HL, Cousens S, Perin J, Scott S, Lawn JE, Rudan I, Campbell H, Cibulskis R, Li M, Mathers C, Black RE, Child Health Epidemiology Reference Group of WHO, Unicef.** 2012. Global, regional, and national causes of child mortality: an updated systematic analysis for 2010 with time trends since 2000. Lancet **379:**2151-2161.

38. **Simonsen KA, Anderson-Berry AL, Delair SF, Davies HD.** 2014. Early-onset neonatal sepsis. Clinical microbiology reviews **27:**21-47.

39. **Podschun R, Ullmann U.** 1998. Klebsiella spp. as nosocomial pathogens: epidemiology, taxonomy, typing methods, and pathogenicity factors. Clinical microbiology reviews **11:**589-603.

40. **England PH.** 2014. Voluntary surveillance of Klebsiella spp. bacteraemia in England, Wales and

Northern Ireland: 2009-2013.

41. **Control ECfDPa.** 2013. Point prevalence survey of healthcare-associated infections and antimicrobial use in European long-term care facilities.

42. **Control ECfDPa** 2014, posting date. Microorganisms and antimicrobial resistance in HAIs. [Online.]

43. **Doyle D, Peirano G, Lascols C, Lloyd T, Church DL, Pitout JD.** 2012. Laboratory detection of Enterobacteriaceae that produce carbapenemases. Journal of clinical microbiology **50:**3877-3880.

44. **Moore R, O'Shea D, Geoghegan T, Mallon PW, Sheehan G.** 2013. Community-acquired Klebsiella pneumoniae liver abscess: an emerging infection in Ireland and Europe. Infection **41:**681-686.

45. **Palmer C, Bik EM, DiGiulio DB, Relman DA, Brown PO.** 2007. Development of the human infant intestinal microbiota. PLoS biology **5:**e177.

46. **Ringel-Kulka T, Cheng J, Ringel Y, Salojarvi J, Carroll I, Palva A, de Vos WM, Satokari R.** 2013. Intestinal microbiota in healthy U.S. young children and adults--a high throughput microarray analysis. PloS one **8:**e64315.

47. **Berg RD.** 1996. The indigenous gastrointestinal microflora. Trends in microbiology **4:**430-435.

48. **Tenaillon O, Skurnik D, Picard B, Denamur E.** 2010. The population genetics of commensal Escherichia coli. Nature reviews. Microbiology **8:**207-217.

49. **Adlerberth I, Jalil F, Carlsson B, Mellander L, Hanson LA, Larsson P, Khalil K, Wold AE.** 1998. High turnover rate of Escherichia coli strains in the intestinal flora of infants in Pakistan. Epidemiology and infection **121:**587-598.

50. **Freter R, Brickner H, Fekete J, Vickerman MM, Carey KE.** 1983. Survival and implantation of Escherichia coli in the intestinal tract. Infection and immunity **39:**686-703.

51. **Fanaro S, Chierici R, Guerrini P, Vigi V.** 2003. Intestinal microflora in early infancy: composition and development. Acta paediatrica **91:**48-55.

52. **Sharma N, Chaudhry R, Panigrahi P.** 2012. Quantitative and qualitative study of intestinal flora in neonates. Journal of global infectious diseases **4:**188-192.

53. **Degener JE, Smit AC, Michel MF, Valkenburg HA, Muller L.** 1983. Faecal carriage of aerobic Gram-negative bacilli and drug resistance of Escherichia coli in different age-groups in Dutch urban communities. Journal of medical microbiology **16:**139-145.

54. **Chung DR, Lee H, Park MH, Jung SI, Chang HH, Kim YS, Son JS, Moon C, Kwon KT, Ryu SY, Shin SY, Ko KS, Kang CI, Peck KR, Song JH.** 2012. Fecal carriage of serotype K1 Klebsiella pneumoniae ST23 strains closely related to liver abscess isolates in Koreans living in Korea. European journal of clinical microbiology & infectious diseases : official publication of the European Society of Clinical Microbiology **31:**481-486.

55. **Franiczek R, Sobieszczanska B, Grabowski M, Mowszet K, Pytrus T.** 2003. Occurrence of extended-spectrum beta-lactamases among Escherichia coli isolates from hospitalized and healthy children. Folia microbiologica **48:**243-247.

56. **Mirelis B, Navarro F, Miro E, Mesa RJ, Coll P, Prats G.** 2003. Community transmission of extended-spectrum beta-lactamase. Emerging infectious diseases **9:**1024-1025.

57. **Woerther PL, Burdet C, Chachaty E, Andremont A.** 2013. Trends in human fecal carriage of extended-spectrum beta-lactamases in the community: toward the globalization of CTX-M. Clinical microbiology reviews **26:**744-758.

58. **Skurnik D, Bonnet D, Bernede-Bauduin C, Michel R, Guette C, Becker JM, Balaire C, Chau F, Mohler J, Jarlier V, Boutin JP, Moreau B, Guillemot D, Denamur E, Andremont A, Ruimy R.** 2008. Characteristics of human intestinal Escherichia coli with changing environments. Environmental microbiology **10:**2132-2137.

59. **Escobar-Paramo P, Grenet K, Le Menac'h A, Rode L, Salgado E, Amorin C, Gouriou S, Picard B, Rahimy MC, Andremont A, Denamur E, Ruimy R.** 2004. Large-scale population structure of human commensal Escherichia coli isolates. Applied and environmental microbiology **70:**5698-5700.

60. **Andriatahina T, Randrianirina F, Hariniana ER, Talarmin A, Raobijaona H, Buisson Y, Richard V.** 2010. High prevalence of fecal carriage of extended-spectrum beta-lactamase-producing Escherichia coli and Klebsiella pneumoniae in a pediatric unit in Madagascar. BMC infectious diseases **10:**204.

61. **Woerther PL, Angebault C, Jacquier H, Clermont O, El Mniai A, Moreau B, Djossou F, Peroz G, Catzeflis F, Denamur E, Andremont A.** 2013. Characterization of fecal extended-spectrum-beta-lactamase-producing Escherichia coli in a remote community during a long time period. Antimicrobial agents and chemotherapy **57:**5060-5066.

62. **Adler A, Baraniak A, Izdebski R, Fiett J, Salvia A, Samso JV, Lawrence C, Solomon J, Paul M, Lerman Y, Schwartzberg Y, Mordechai E, Rossini A, Fierro J, Lammens C, Malhotra-Kumar S, Goossens H, Hryniewicz W, Brun-Buisson C, Gniadkowski M, Carmeli Y, team M.** 2014. A multinational study of colonization with extended spectrum beta-lactamase-producing Enterobacteriaceae in healthcare personnel and family members of carrier patients hospitalized in rehabilitation centres. Clinical microbiology

and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases.

63. **Escobar-Paramo P, Le Menac'h A, Le Gall T, Amorin C, Gouriou S, Picard B, Skurnik D, Denamur E.** 2006. Identification of forces shaping the commensal Escherichia coli genetic structure by comparing animal and human isolates. Environmental microbiology **8:**1975-1984.

64. **Gordon DM, Cowling A.** 2003. The distribution and genetic structure of Escherichia coli in Australian vertebrates: host and geographic effects. Microbiology **149:**3575-3586.

65. **Costa D, Poeta P, Saenz Y, Vinue L, Rojo-Bezares B, Jouini A, Zarazaga M, Rodrigues J, Torres C.** 2006. Detection of Escherichia coli harbouring extended-spectrum beta-lactamases of the CTX-M, TEM and SHV classes in faecal samples of wild animals in Portugal. The Journal of antimicrobial chemotherapy **58:**1311-1312.

66. **Guenther S, Aschenbrenner K, Stamm I, Bethe A, Semmler T, Stubbe A, Stubbe M, Batsajkhan N, Glupczynski Y, Wieler LH, Ewers C.** 2012. Comparable high rates of extended-spectrum-beta-lactamase-producing Escherichia coli in birds of prey from Germany and Mongolia. PloS one **7:**e53039.

67. **Veldman K, van Tulden P, Kant A, Testerink J, Mevius D.** 2013. Characteristics of cefotaxime-resistant Escherichia coli from wild birds in the Netherlands. Applied and environmental microbiology **79:**7556-7561.

68. **Radhouani H, Igrejas G, Goncalves A, Estepa V, Sargo R, Torres C, Poeta P.** 2013. Molecular characterization of extended-spectrum-beta-lactamase-producing Escherichia coli isolates from red foxes in Portugal. Archives of microbiology **195:**141-144.

69. **Stephan R, Hachler H.** 2012. Discovery of extended-spectrum beta-lactamase producing Escherichia coli among hunted deer, chamois and ibex. Schweizer Archiv fur Tierheilkunde **154:**475-478.

70. **Tamang MD, Nam HM, Jang GC, Kim SR, Chae MH, Jung SC, Byun JW, Park YH, Lim SK.** 2012. Molecular characterization of extended-spectrum-beta-lactamase-producing and plasmid-mediated AmpC beta-lactamase-producing Escherichia coli isolated from stray dogs in South Korea. Antimicrobial agents and chemotherapy **56:**2705-2712.

71. **Guenther S, Grobbel M, Beutlich J, Guerra B, Ulrich RG, Wieler LH, Ewers C.** 2010. Detection of pandemic B2-O25-ST131 Escherichia coli harbouring the CTX-M-9 extended-spectrum beta-lactamase type in a feral urban brown rat (Rattus norvegicus). The Journal of antimicrobial chemotherapy **65:**582-584.

72. **Literak I, Dolejska M, Radimersky T, Klimes J, Friedman M, Aarestrup FM, Hasman H, Cizek A.** 2010. Antimicrobial-resistant faecal Escherichia coli in wild mammals in central Europe: multiresistant Escherichia coli producing extended-spectrum beta-lactamases in wild boars. Journal of applied microbiology **108:**1702-1711.

73. **Janatova M, Albrechtova K, Petrzelkova KJ, Dolejska M, Papousek I, Masarikova M, Cizek A, Todd A, Shutt K, Kalousova B, Profousova-Psenkova I, Modry D, Literak I.** 2014. Antimicrobial-resistant Enterobacteriaceae from humans and wildlife in Dzanga-Sangha Protected Area, Central African Republic. Veterinary microbiology **171:**422-431.

74. **Johnson JR, Stell AL, Delavari P.** 2001. Canine feces as a reservoir of extraintestinal pathogenic Escherichia coli. Infection and immunity **69:**1306-1314.

75. **Johnson JR, Stell AL, Delavari P, Murray AC, Kuskowski M, Gaastra W.** 2001. Phylogenetic and pathotypic similarities between Escherichia coli isolates from urinary tract infections in dogs and extraintestinal infections in humans. The Journal of infectious diseases **183:**897-906.

76. **Osugui L, de Castro AF, Iovine R, Irino K, Carvalho VM.** 2014. Virulence genotypes, antibiotic resistance and the phylogenetic background of extraintestinal pathogenic Escherichia coli isolated from urinary tract infections of dogs and cats in Brazil. Veterinary microbiology **171:**242-247.

77. **Hoyle DV, Knight HI, Shaw DJ, Hillman K, Pearce MC, Low JC, Gunn GJ, Woolhouse ME.** 2004. Acquisition and epidemiology of antibiotic-resistant Escherichia coli in a cohort of newborn calves. The Journal of antimicrobial chemotherapy **53:**867-871.

78. **Kang SG, Lee DY, Shin SJ, Ahn JM, Yoo HS.** 2005. Changes in patterns of antimicrobial susceptibility and class 1 integron carriage among Escherichia coli isolates. Journal of veterinary science **6:**201-205.

79. **Smith JL, Drum DJ, Dai Y, Kim JM, Sanchez S, Maurer JJ, Hofacre CL, Lee MD.** 2007. Impact of antimicrobial usage on antimicrobial resistance in commensal Escherichia coli strains colonizing broiler chickens. Applied and environmental microbiology **73:**1404-1414.

80. **Geser N, Stephan R, Kuhnert P, Zbinden R, Kaeppeli U, Cernela N, Haechler H.** 2011. Fecal carriage of extended-spectrum beta-lactamase-producing Enterobacteriaceae in swine and cattle at slaughter in Switzerland. Journal of food protection **74:**446-449.

81. **Ho PL, Chow KH, Lai EL, Lo WU, Yeung MK, Chan J, Chan PY, Yuen KY.** 2011. Extensive dissemination of CTX-M-producing Escherichia coli with multidrug resistance to 'critically important' antibiotics among food

animals in Hong Kong, 2008-10. The Journal of antimicrobial chemotherapy **66:**765-768.

82. **Schoster A, Staempfli HR, Arroyo LG, Reid-Smith RJ, Janecko N, Shewen PE, Weese JS.** 2012. Longitudinal study of Clostridium difficile and antimicrobial susceptibility of Escherichia coli in healthy horses in a community setting. Veterinary microbiology **159:**364-370.

83. **Duan RS, Sit TH, Wong SS, Wong RC, Chow KH, Mak GC, Yam WC, Ng LT, Yuen KY, Ho PL.** 2006. Escherichia coli producing CTX-M beta-lactamases in food animals in Hong Kong. Microbial drug resistance **12:**145-148.

84. **Hansen KH, Damborg P, Andreasen M, Nielsen SS, Guardabassi L.** 2013. Carriage and fecal counts of cefotaxime M-producing Escherichia coli in pigs: a longitudinal study. Applied and environmental microbiology **79:**794-798.

85. **Leverstein-van Hall MA, Dierikx CM, Cohen Stuart J, Voets GM, van den Munckhof MP, van Essen-Zandbergen A, Platteel T, Fluit AC, van de Sande-Bruinsma N, Scharinga J, Bonten MJ, Mevius DJ, National Esg.** 2011. Dutch patients, retail chicken meat and poultry share the same ESBL genes, plasmids and strains. Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases **17:**873-880.

86. **Johnson JR, McCabe JS, White DG, Johnston B, Kuskowski MA, McDermott P.** 2009. Molecular Analysis of Escherichia coli from retail meats (2002-2004) from the United States National Antimicrobial Resistance Monitoring System. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America **49:**195-201.

87. **Whitehouse CA, Keirstead N, Taylor J, Reinhardt JL, Beierschmitt A.** 2010. Prevalence of hypermucoid Klebsiella pneumoniae among wild-caught and captive vervet monkeys (Chlorocebus aethiops sabaeus) on the island of St. Kitts. Journal of wildlife diseases **46:**971-976.

88. **Elgderi RM, Ghenghesh KS, Berbash N.** 2006. Carriage by the German cockroach (Blattella germanica) of multiple-antibiotic-resistant bacteria that are potentially pathogenic to humans, in hospitals and households in Tripoli, Libya. Annals of tropical medicine and parasitology **100:**55-62.

89. **Rahuma N, Ghenghesh KS, Ben Aissa R, Elamaari A.** 2005. Carriage by the housefly (Musca domestica) of multiple-antibiotic-resistant bacteria that are potentially pathogenic to humans, in hospital and other urban environments in Misurata, Libya. Annals of tropical medicine and parasitology **99:**795-802.

90. **Guerra B, Fischer J, Helmuth R.** 2014. An emerging public health problem: acquired carbapenemase-producing microorganisms are present in food-producing animals, their environment, companion animals and wild birds. Veterinary microbiology **171:**290-297.

91. **Schmiedel J, Falgenhauer L, Domann E, Bauerfeind R, Prenger-Berninghoff E, Imirzalioglu C, Chakraborty T.** 2014. Multiresistant extended-spectrum beta-lactamase-producing Enterobacteriaceae from humans, companion animals and horses in central Hesse, Germany. BMC microbiology **14:**187.

92. **Orskov I, Orskov F, Jann B, Jann K.** 1977. Serology, chemistry, and genetics of O and K antigens of Escherichia coli. Bacteriological reviews **41:**667-710.

93. **Brenner DJ, Fanning GR, Skerman FJ, Falkow S.** 1972. Polynucleotide sequence divergence among strains of Escherichia coli and closely related organisms. Journal of bacteriology **109:**953-965.

94. **Ochman H, Whittam TS, Caugant DA, Selander RK.** 1983. Enzyme polymorphism and genetic population structure in Escherichia coli and Shigella. Journal of general microbiology **129:**2715-2726.

95. **Whittam TS, Ochman H, Selander RK.** 1983. Multilocus genetic structure in natural populations of Escherichia coli. Proceedings of the National Academy of Sciences of the United States of America **80:**1751-1755.

96. **Ochman H, Selander RK.** 1984. Standard reference strains of Escherichia coli from natural populations. Journal of bacteriology **157:**690-693.

97. **Herzer PJ, Inouye S, Inouye M, Whittam TS.** 1990. Phylogenetic distribution of branched RNA-linked multicopy single-stranded DNA among natural isolates of Escherichia coli. Journal of bacteriology **172:**6175-6181.

98. **Chaudhuri RR, Henderson IR.** 2012. The evolution of the Escherichia coli phylogeny. Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases **12:**214-226.

99. **Desjardins P, Picard B, Kaltenbock B, Elion J, Denamur E.** 1995. Sex in Escherichia coli does not disrupt the clonal structure of the population: evidence from random amplified polymorphic DNA and restriction-fragment-length polymorphism. Journal of molecular evolution **41:**440-448.

100. **Reid SD, Herbelin CJ, Bumbaugh AC, Selander RK, Whittam TS.** 2000. Parallel evolution of virulence in pathogenic Escherichia coli. Nature **406:**64-67.

101. **Escobar-Paramo P, Clermont O, Blanc-Potard AB, Bui H, Le Bouguenec C, Denamur E.** 2004. A specific genetic background is required for

acquisition and expression of virulence factors in Escherichia coli. Molecular biology and evolution **21:**1085-1094.

102. **Touchon M, Hoede C, Tenaillon O, Barbe V, Baeriswyl S, Bidet P, Bingen E, Bonacorsi S, Bouchier C, Bouvet O, Calteau A, Chiapello H, Clermont O, Cruveiller S, Danchin A, Diard M, Dossat C, Karoui ME, Frapy E, Garry L, Ghigo JM, Gilles AM, Johnson J, Le Bouguenec C, Lescat M, Mangenot S, Martinez-Jehanne V, Matic I, Nassif X, Oztas S, Petit MA, Pichon C, Rouy Z, Ruf CS, Schneider D, Tourret J, Vacherie B, Vallenet D, Medigue C, Rocha EP, Denamur E.** 2009. Organised genome dynamics in the Escherichia coli species results in highly diverse adaptive paths. PLoS genetics **5:**e1000344.

103. **Didelot X, Meric G, Falush D, Darling AE.** 2012. Impact of homologous and non-homologous recombination in the genomic evolution of Escherichia coli. BMC genomics **13:**256.

104. **Darling AE, Mau B, Perna NT.** 2010. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PloS one **5:**e11147.

105. **Didelot X, Falush D.** 2007. Inference of bacterial microevolution using multilocus sequence data. Genetics **175:**1251-1266.

106. **Brisse S, Verhoef J.** 2001. Phylogenetic diversity of Klebsiella pneumoniae and Klebsiella oxytoca clinical isolates revealed by randomly amplified polymorphic DNA, gyrA and parC genes sequencing and automated ribotyping. International journal of systematic and evolutionary microbiology **51:**915-924.

107. **Brisse S, Duijkeren E.** 2005. Identification and antimicrobial susceptibility of 100 Klebsiella animal clinical isolates. Veterinary microbiology **105:**307-312.

108. **Deleo FR, Chen L, Porcella SF, Martens CA, Kobayashi SD, Porter AR, Chavda KD, Jacobs MR, Mathema B, Olsen RJ, Bonomo RA, Musser JM, Kreiswirth BN.** 2014. Molecular dissection of the evolution of carbapenem-resistant multilocus sequence type 258 Klebsiella pneumoniae. Proceedings of the National Academy of Sciences of the United States of America **111:**4988-4993.

109. **Wright MS, Perez F, Brinkac L, Jacobs MR, Kaye K, Cober E, van Duin D, Marshall SH, Hujer AM, Rudin SD, Hujer KM, Bonomo RA, Adams MD.** 2014. Population Structure of KPC-Producing Klebsiella pneumoniae Isolates from Midwestern U.S. Hospitals. Antimicrobial agents and chemotherapy **58:**4961-4965.

110. **Wright GD.** 2007. The antibiotic resistome: the nexus of chemical and genetic diversity. Nature reviews. Microbiology **5:**175-186.

111. **Boucher HW, Talbot GH, Benjamin DK, Jr., Bradley J, Guidos RJ, Jones RN, Murray BE, Bonomo RA, Gilbert D, Infectious Diseases Society of A.** 2013. 10 x '20 Progress--development of new drugs active against Gram-negative bacilli: an update from the Infectious Diseases Society of America. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America **56:**1685-1694.

112. **Torres JA, Villegas MV, Quinn JP.** 2007. Current concepts in antibiotic-resistant Gram-negative bacteria. Expert review of anti-infective therapy **5:**833-843.

113. **Delcour AH.** 2009. Outer membrane permeability and antibiotic resistance. Biochimica et biophysica acta **1794:**808-816.

114. **Schweizer HP.** 2012. Understanding efflux in Gram-negative bacteria: opportunities for drug discovery. Expert opinion on drug discovery **7:**633-642.

115. **Caroff N, Espaze E, Berard I, Richet H, Reynaud A.** 1999. Mutations in the ampC promoter of Escherichia coli isolates resistant to oxyiminocephalosporins without extended spectrum beta-lactamase production. FEMS microbiology letters **173:**459-465.

116. **Seetulsingh PS, Hall LM, Livermore DM.** 1991. Activity of clavulanate combinations against TEM-1 beta-lactamase-producing Escherichia coli isolates obtained in 1982 and 1989. The Journal of antimicrobial chemotherapy **27:**749-759.

117. **Moland ES, Hong SG, Thomson KS, Larone DH, Hanson ND.** 2007. Klebsiella pneumoniae isolate producing at least eight different beta-lactamases, including AmpC and KPC beta-lactamases. Antimicrobial agents and chemotherapy **51:**800-801.

118. **Andersson DI, Hughes D.** 2010. Antibiotic resistance and its cost: is it possible to reverse resistance? Nature reviews. Microbiology **8:**260-271.

119. **Canton R, Ruiz-Garbajosa P.** 2011. Co-resistance: an opportunity for the bacteria and resistance genes. Current opinion in pharmacology **11:**477-485.

120. **Paterson DL.** 2006. Resistance in Gram-negative bacteria: Enterobacteriaceae. American journal of infection control **34:**S20-28; discussion S64-73.

121. **Jacoby GB, K.** 2012, posting date. β-Lactamase Classification and Amino Acid Sequences for TEM, SHV and OXA Extended-spectrum and Inhibitor-resistant Enzymes. [Online.]

122. **Canton R, Morosini MI, de la Maza OM, de la Pedrosa EG.** 2008. IRT and CMT beta-lactamases and inhibitor resistance. Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases **14 Suppl 1:**53-62.

123. **Jacoby GA.** 2009. AmpC beta-lactamases. Clinical microbiology reviews **22:**161-182, Table of Contents.

124. **Naas T, Poirel L, Nordmann P.** 2008. Minor extended-spectrum beta-lactamases. Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases **14 Suppl 1:**42-52.

125. **Canton R, Akova M, Carmeli Y, Giske CG, Glupczynski Y, Gniadkowski M, Livermore DM, Miriagou V, Naas T, Rossolini GM, Samuelsen O, Seifert H, Woodford N, Nordmann P, European Network on C.** 2012. Rapid evolution and spread of carbapenemases among Enterobacteriaceae in Europe. Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases **18:**413-431.

126. **Nordmann P, Poirel L, Walsh TR, Livermore DM.** 2011. The emerging NDM carbapenemases. Trends in microbiology **19:**588-595.

127. **Poirel L, Pitout JD, Nordmann P.** 2007. Carbapenemases: molecular diversity and clinical consequences. Future microbiology **2:**501-512.

128. **Maltezou HC.** 2009. Metallo-beta-lactamases in Gram-negative bacteria: introducing the era of pan-resistance? International journal of antimicrobial agents **33:**405 e401-407.

129. **Bush K, Jacoby GA.** 2010. Updated functional classification of beta-lactamases. Antimicrobial agents and chemotherapy **54:**969-976.

130. **Bush K, Jacoby GA, Medeiros AA.** 1995. A functional classification scheme for beta-lactamases and its correlation with molecular structure. Antimicrobial agents and chemotherapy **39:**1211-1233.

131. **Livermore DM.** 1995. beta-Lactamases in laboratory and clinical resistance. Clinical microbiology reviews **8:**557-584.

132. **Drawz SM, Bonomo RA.** 2010. Three decades of beta-lactamase inhibitors. Clinical microbiology reviews **23:**160-201.

133. **Karsisiotis AI, Damblon CF, Roberts GC.** 2014. A variety of roles for versatile zinc in metallo-beta-lactamases. Metallomics : integrated biometal science **6:**1181-1197.

134. **Yong D, Toleman MA, Giske CG, Cho HS, Sundman K, Lee K, Walsh TR.** 2009. Characterization of a new metallo-beta-lactamase gene, bla(NDM-1), and a novel erythromycin esterase gene carried on a unique genetic structure in Klebsiella pneumoniae sequence type 14 from India. Antimicrobial agents and chemotherapy **53:**5046-5054.

135. **Johnson AP, Woodford N.** 2013. Global spread of antibiotic resistance: the example of New Delhi metallo-beta-lactamase (NDM)-mediated carbapenem resistance. Journal of medical microbiology **62:**499-513.

136. **Castanheira M, Deshpande LM, Mathai D, Bell JM, Jones RN, Mendes RE.** 2011. Early dissemination of NDM-1- and OXA-181-producing Enterobacteriaceae in Indian hospitals: report from the SENTRY Antimicrobial Surveillance Program, 2006-2007. Antimicrobial agents and chemotherapy **55:**1274-1278.

137. **Walsh TR, Weeks J, Livermore DM, Toleman MA.** 2011. Dissemination of NDM-1 positive bacteria in the New Delhi environment and its implications for human health: an environmental point prevalence study. The Lancet infectious diseases **11:**355-362.

138. **Isozumi R, Yoshimatsu K, Yamashiro T, Hasebe F, Nguyen BM, Ngo TC, Yasuda SP, Koma T, Shimizu K, Arikawa J.** 2012. bla(NDM-1)-positive

Klebsiella pneumoniae from environment, Vietnam. Emerging infectious diseases **18:**1383-1385.

139. **Munoz-Price LS, Poirel L, Bonomo RA, Schwaber MJ, Daikos GL, Cormican M, Cornaglia G, Garau J, Gniadkowski M, Hayden MK, Kumarasamy K, Livermore DM, Maya JJ, Nordmann P, Patel JB, Paterson DL, Pitout J, Villegas MV, Wang H, Woodford N, Quinn JP.** 2013. Clinical epidemiology of the global expansion of Klebsiella pneumoniae carbapenemases. The Lancet infectious diseases **13:**785-796.

140. **Bonnet R.** 2004. Growing group of extended-spectrum beta-lactamases: the CTX-M enzymes. Antimicrobial agents and chemotherapy **48:**1-14.

141. **Hawkey PM, Jones AM.** 2009. The changing epidemiology of resistance. The Journal of antimicrobial chemotherapy **64 Suppl 1:**i3-10.

142. **Mahillon J, Chandler M.** 1998. Insertion sequences. Microbiology and molecular biology reviews : MMBR **62:**725-774.

143. **Siguier P, Perochon J, Lestrade L, Mahillon J, Chandler M.** 2006. ISfinder: the reference centre for bacterial insertion sequences. Nucleic acids research **34:**D32-36.

144. **Nordmann P, Dortet L, Poirel L.** 2012. Carbapenem resistance in Enterobacteriaceae: here is the storm! Trends in molecular medicine **18:**263-272.

145. **Poirel L, Naas T, Nordmann P.** 2008. Genetic support of extended-spectrum beta-lactamases. Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases **14 Suppl 1:**75-81.

146. **Poirel L, Heritier C, Tolun V, Nordmann P.** 2004. Emergence of oxacillinase-mediated resistance to imipenem in Klebsiella pneumoniae. Antimicrobial agents and chemotherapy **48:**15-22.

147. **Potron A, Nordmann P, Rondinaud E, Jaureguy F, Poirel L.** 2013. A mosaic transposon encoding OXA-48 and CTX-M-15: towards pan-resistance. The Journal of antimicrobial chemotherapy **68:**476-477.

148. **Kitchel B, Rasheed JK, Endimiani A, Hujer AM, Anderson KF, Bonomo RA, Patel JB.** 2010. Genetic factors associated with elevated carbapenem resistance in KPC-producing Klebsiella pneumoniae. Antimicrobial agents and chemotherapy **54:**4201-4207.

149. **Naas T, Cuzon G, Villegas MV, Lartigue MF, Quinn JP, Nordmann P.** 2008. Genetic structures at the origin of acquisition of the beta-lactamase bla KPC gene. Antimicrobial agents and chemotherapy **52:**1257-1263.

150. **Datta N, Hedges RW.** 1971. Compatibility groups among fi - R factors. Nature **234:**222-223.

151. **Couturier M, Bex F, Bergquist PL, Maas WK.** 1988. Identification and classification of bacterial plasmids. Microbiological reviews **52:**375-395.

152. **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.

153. **Carattoli A.** 2009. Resistance plasmid families in Enterobacteriaceae. Antimicrobial agents and chemotherapy **53:**2227-2238.

154. **Johnson TJ, Nolan LK.** 2009. Pathogenomics of the virulence plasmids of Escherichia coli. Microbiology and molecular biology reviews : MMBR **73:**750-774.

155. **Garcia-Fernandez A, Villa L, Carta C, Venditti C, Giordano A, Venditti M, Mancini C, Carattoli A.** 2012. Klebsiella pneumoniae ST258 producing KPC-3 identified in italy carries novel plasmids and OmpK36/OmpK35 porin variants. Antimicrobial agents and chemotherapy **56:**2143-2145.

156. **Leavitt A, Carmeli Y, Chmelnitsky I, Goren MG, Ofek I, Navon-Venezia S.** 2010. Molecular epidemiology, sequence types, and plasmid analyses of KPC-producing Klebsiella pneumoniae strains in Israel. Antimicrobial agents and chemotherapy **54:**3002-3006.

157. **Sykora P.** 1992. Macroevolution of plasmids: a model for plasmid speciation. Journal of theoretical biology **159:**53-65.

158. **Woodford N, Turton JF, Livermore DM.** 2011. Multiresistant Gram-negative bacteria: the role of high-risk clones in the dissemination of antibiotic resistance. FEMS microbiology reviews **35:**736-755.

159. **Nicolas-Chanoine MH, Blanco J, Leflon-Guibout V, Demarty R, Alonso MP, Canica MM, Park YJ, Lavigne JP, Pitout J, Johnson JR.** 2008. Intercontinental emergence of Escherichia coli clone O25:H4-ST131 producing CTX-M-15. The Journal of antimicrobial chemotherapy **61:**273-281.

160. **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 beta-lactamase CTX-M-15. Emerging infectious diseases **14:**195-200.

161. **Naseer U, Sundsfjord A.** 2011. The CTX-M conundrum: dissemination of plasmids and Escherichia coli clones. Microbial drug resistance **17:**83-97.

162. **Ewers C, Bethe A, Stamm I, Grobbel M, Kopp PA, Guerra B, Stubbe M, Doi Y, Zong Z, Kola A, Schaufler K, Semmler T, Fruth A, Wieler LH,**

Guenther S. 2014. CTX-M-15-D-ST648 Escherichia coli from companion animals and horses: another pandemic clone combining multiresistance and extraintestinal virulence? The Journal of antimicrobial chemotherapy **69:**1224-1230.

163. **Chen L, Mathema B, Pitout JD, DeLeo FR, Kreiswirth BN.** 2014. Epidemic Klebsiella pneumoniae ST258 is a hybrid strain. mBio **5:**e01355-01314.

164. **Cao X, Xu X, Zhang Z, Shen H, Chen J, Zhang K.** 2014. Molecular characterization of clinical multidrug-resistant Klebsiella pneumoniae isolates. Annals of clinical microbiology and antimicrobials **13:**16.

165. **Andrade LN, Vitali L, Gaspar GG, Bellissimo-Rodrigues F, Martinez R, Darini AL.** 2014. Expansion and evolution of a virulent, extensively drug-resistant (polymyxin B-resistant), QnrS1-, CTX-M-2-, and KPC-2-producing Klebsiella pneumoniae ST11 international high-risk clone. Journal of clinical microbiology **52:**2530-2535.

166. **Pena I, Picazo JJ, Rodriguez-Avial C, Rodriguez-Avial I.** 2014. Carbapenemase-producing Enterobacteriaceae in a tertiary hospital in Madrid, Spain: high percentage of colistin resistance among VIM-1-producing Klebsiella pneumoniae ST11 isolates. International journal of antimicrobial agents **43:**460-464.

167. **Voulgari E, Gartzonika C, Vrioni G, Politi L, Priavali E, Levidiotou-Stefanou S, Tsakris A.** 2014. The Balkan region: NDM-1-producing Klebsiella pneumoniae ST11 clonal strain causing outbreaks in Greece. The Journal of antimicrobial chemotherapy **69:**2091-2097.

168. **Novais A, Rodrigues C, Branquinho R, Antunes P, Grosso F, Boaventura L, Ribeiro G, Peixe L.** 2012. Spread of an OmpK36-modified ST15

Klebsiella pneumoniae variant during an outbreak involving multiple carbapenem-resistant Enterobacteriaceae species and clones. European journal of clinical microbiology & infectious diseases : official publication of the European Society of Clinical Microbiology **31:**3057-3063.

169. **Rodrigues C, Machado E, Ramos H, Peixe L, Novais A.** 2014. Expansion of ESBL-producing Klebsiella pneumoniae in hospitalized patients: A successful story of international clones (ST15, ST147, ST336) and epidemic plasmids (IncR, IncFII). International journal of medical microbiology : IJMM.

170. **Rodrigues C, Novais A, Machado E, Peixe L.** 2014. Detection of VIM-34, a novel VIM-1 variant identified in the intercontinental ST15 Klebsiella pneumoniae clone. The Journal of antimicrobial chemotherapy **69:**274-275.

171. **Kim J, Bae IK, Jeong SH, Chang CL, Lee CH, Lee K.** 2011. Characterization of IncF plasmids carrying the blaCTX-M-14 gene in clinical isolates of Escherichia coli from Korea. The Journal of antimicrobial chemotherapy **66:**1263-1268.

172. **Hirai I, Fukui N, Taguchi M, Yamauchi K, Nakamura T, Okano S, Yamamoto Y.** 2013. Detection of chromosomal blaCTX-M-15 in Escherichia coli O25b-B2-ST131 isolates from the Kinki region of Japan. International journal of antimicrobial agents **42:**500-506.

173. **Coelho A, Gonzalez-Lopez JJ, Miro E, Alonso-Tarres C, Mirelis B, Larrosa MN, Bartolome RM, Andreu A, Navarro F, Johnson JR, Prats G.** 2010. Characterisation of the CTX-M-15-encoding gene in Klebsiella pneumoniae strains from the Barcelona metropolitan area: plasmid diversity and chromosomal integration. International journal of antimicrobial agents **36:**73-78.

174. **Ferreira JC, Penha Filho RA, Andrade LN, Berchieri AJ, Darini AL.** 2014. Detection of chromosomal bla in diverse Escherichia coli isolates from healthy broiler chickens. Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases.

175. **Fabre L, Delaune A, Espie E, Nygard K, Pardos M, Polomack L, Guesnier F, Galimand M, Lassen J, Weill FX.** 2009. Chromosomal integration of the extended-spectrum beta-lactamase gene blaCTX-M-15 in Salmonella enterica serotype Concord isolates from internationally adopted children. Antimicrobial agents and chemotherapy **53:**1808-1816.

176. **Harada S, Ishii Y, Saga T, Kouyama Y, Tateda K, Yamaguchi K.** 2012. Chromosomal integration and location on IncT plasmids of the blaCTX-M-2 gene in Proteus mirabilis clinical isolates. Antimicrobial agents and chemotherapy **56:**1093-1096.

177. **Navon-Venezia S, Chmelnitsky I, Leavitt A, Carmeli Y.** 2008. Dissemination of the CTX-M-25 family beta-lactamases among Klebsiella pneumoniae, Escherichia coli and Enterobacter cloacae and identification of the novel enzyme CTX-M-41 in Proteus mirabilis in Israel. The Journal of antimicrobial chemotherapy **62:**289-295.

178. **Song W, Kim J, Bae IK, Jeong SH, Seo YH, Shin JH, Jang SJ, Uh Y, Shin JH, Lee MK, Lee K.** 2011. Chromosome-encoded AmpC and CTX-M extended-spectrum beta-lactamases in clinical isolates of Proteus mirabilis from Korea. Antimicrobial agents and chemotherapy **55:**1414-1419.

179. **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.

180. 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.

CHAPTER 3: THESIS METHODS

3.1. INTRODUCTION

The work in this thesis required the development of a range of skills, including laboratory work, classical epidemiology and statistics, the acquisition of a basic understanding of linux, simple scripting in python, and the use of a number of software packages designed for the processing of sequencing data and for population genetic analyses.

3.2. LABORATORY METHODS

3.2.1. SAMPLE COLLECTION AND SAMPLING FRAMES

The sampling frame for this thesis was made possible as a result of close collaboration with an number of different research groups, including the South-East Asian network operating out of the Mahidol-Oxford Tropical Research Unit (MORU), based in Bangkok, Thailand; and collaborative efforts set up with Prof James Johnson, of the University of Minnesota; the Paediatric Study Team operating out of the University of Oxford and Patan Hospital in Kathmandu; Drs Amy Mathers and Costi Sifri working at the School of Medicine, University of Virginia; Dr Amee Manges of the University of British Columbia, Vancouver, and Drs Veronica Kos and Bob McLaughlin at AstraZeneca. Most of the samples collected from South-East Asian units were retrieved from frozen stocks by either myself or colleagues during a period of time preceding the thesis in which I was undertaking epidemiological research and clinical work in Thailand and Cambodia.

3.2.2. *ESCHERCHIA COLI* AND *KLEBSIELLA PNEUMONIAE* CULTURE AND IDENTIFICATION

Culture for non-enteropathogenic *E. coli* and *K. pneumoniae* from frozen stocks was undertaken on a variety of different agar plates depending on local availability. Selective agars included MacConkey, chromogenic agar in the form of Orientation UTI agar, and Luria-Bertani agar infused with antibiotics that I prepared in the laboratory in cases where this was required for the selection of antimicrobial-resistant strains.

Typical culture conditions involved streaking out the inoculum on the agar, and incubating in air at 37°C for 18-24 hours.

3.2.3. DNA EXTRACTION FOR BACTERIAL STRAINS AND PLASMIDS

DNA extraction for bacterial strains and electroporated transformants for all sequencing methods was carried out using a commercial, kit-based method manufactured by Quickgene (DNA Tissue Kit S; Fujifilm/Kurabo Biomedical, Tokyo, Japan), with an additional mechanical lysis step using the FastPrep homogeniser (MP Biomedicals, Santa Ana, USA) and its "lysing matrix B", containing 0.1mm silica beads adapted for the lysis of bacteria.

Briefly, a 10microl loopful of freshly sub-cultured bacterial growth was suspended in a mixture of sterile saline, proprietary lysis buffer (LDT) and proteinase K (EDT), to chemically lyse the bacterial cell walls, and digest any contaminating proteins, particularly nucleases, present in the lysate. This mixture was then processed in the FastPrep benchtop homogeniser in the lysing matrix B tubes for 40 seconds at a speed

of 6m/second as a mechanical cell lysis step – this was done twice. Subsequently all tubes were incubated in a heat block at 70°C for 10 minutes to denature proteins/enzymes, and then centrifuged for a further 10 minutes at 13,200 rpm in a microcentrifuge to pellet any cellular debris and the matrix beads.

The supernatant (approximately 450µl) was then added to 240µl of 99% ethanol and the mixture pipetted up and down briefly to mix and precipitate the DNA, before being transferred to the QuickGene extraction columns. These were then subjected to high pressure in one of the QuickGene extraction units (QuickGene mini-80), initially to trap DNA within the cartridge matrix, then to wash away impurities with proprietary wash buffer (this step was undertaken three times), and then elute the DNA with the kit elution buffer. This consistently provided high-yield, good-quality DNA which proved appropriate for sequencing.

In preparation for Illumina HiSeq runs, DNA was diluted to a normalised concentration of 20ng/µl following quantification with PicoGreen, a proprietary fluorochrome that selectively binds dsDNA(1). Batches of 96 samples, including a DNA extract of a reference strain, were included in each microtitre plate set up for sequencing. Plates were then transferred to the Wellcome Trust Centre for Human Genetics Sequencing hub, where all the Illumina HiSeq sequencing for this project was undertaken.

Plasmid DNA was extracted from freshly prepared sub-cultures of frozen stock grown overnight on blood agar using the Qiagen plasmid mini-kit (Qiagen, Venlo, Netherlands), in accordance with the manufacturer's instructions. Essentially this is a

modified alkaline lysis procedure(2), in which detergent (sodium dodecyl sulphate SDS) is used to disrupt cellular membranes and sodium hydroxide is used to selectively denature sheared, linearized chromosomal DNA whilst the plasmid DNA remains in solution. Potassium acetate is then added to acidify the solution, causing precipitation of: (i) renatured chromosomal DNA; (ii) SDS as potassium dodecyl sulphate (PDS); (iii) contaminating protein in the form of protein-PDS complexes; and (iv) high molecular weight RNA. These are then removed in a centrifugation step, and plasmid DNA is precipitated with ethanol and bound to an anion-exchange matrix in spin columns under appropriate low-salt and pH conditions. Any residual RNA, protein and other contaminants are subsequently removed by a medium-salt wash. Plasmid DNA is eluted from the column matrix in a high-salt buffer, and then concentrated and desalted by isopropanol precipitation.

The plasmid extraction protocol was modified by the addition of Glycoblue^{TM} co-precipitate (Life Technologies, Carlsbad, USA) to the DNA eluates prior to isopropanol precipitation to enable better visualisation of the DNA pellet, thus making sure that it was not inadvertently lost in processing. Selective starter cultures were used, depending on the main resistance mechanism of interest (Luria-Bertani broth and ceftriaxone for broad-spectrum beta-lactamase containing plasmids, or chloramphenicol, or ampicillin at 1, 8 or 50 μg/mL respectively). Plasmid DNA was re-dissolved in distilled water and then typically electroporated on the same day, or stored in the fridge prior to electroporation within 24 hours.

3.2.4. PLASMID ELECTROPORATION

In order to enable the specific isolation of antimicrobial resistance plasmids on a subset of strains, transformation by electroporation of extracted plasmids was undertaken. Electroporation is an experimental method of introducing foreign DNA into cells by means of the application of a high-voltage electrical force of short duration, reversibly permeabilising the cell membrane and allowing the uptake of external genetic material(3). Electroporation was selected as the preferred method of transformation, because it rarely results in the transfer of more than a single plasmid, unlike the other main alternative method, conjugation.

Commercially prepared DH10B *E. coli* (ElectroMAX™ DH10B™ Cells; Invitrogen/Life Technologies, Carlsbad, California, USA) were used as the recipient cell strain for plasmid electroporation, because of their high transformation efficiencies and the fact that the strain has been fully sequenced (NCBI RefSeq: NC_010473.1)(4). Sequenced DNA belonging to this strain could therefore be identified and bioinformatically filtered, leaving plasmid-associated sequences available for further analysis. pUC19 DNA was provided with the purchased cells and was used as a positive control for the success of transformation. pUC19 is a small (2868bp) high-copy plasmid that is commonly used as a cloning vector and has an ampicillin resistance gene, enabling its selection on ampicillin-containing agar.

Briefly, 2µl of plasmid DNA (extracted as above) were mixed with 20µl of electrocompetent cells in a small Eppendorf tube on ice. The mix was then pipetted into a pre-chilled 0.2cm electroporation cuvette (Bio-Rad, Hercules, California, USA), placed in the MicroPulser™ electroporator, and an electric shock was applied

(Ec2 settings; typically 2.5 kV applied for less than 5msec). The shocked cells were immediately suspended in 0.9mls of pre-warmed SOC medium (Super Optimal Broth with Catabolite repression) in a clean Eppendorf tube, and incubated at 37°C, whilst being shaken at 220rpm, for one hour. 100µl of the transformant cell suspension was then plated onto pre-warmed locally-made Luria-Bertani agar (Becton Dickinson, Franklin Lakes, New Jersey, USA), infused with ceftriaxone (1µg/mL) for the selection of broad-spectrum beta-lactamase containing plasmids, chloramphenicol (8 µg/mL) for pOX38 as a control plasmid for sequencing, or ampicillin (50µg/mL) for the selection of pUC19 transformants as the positive controls as both a control for plasmid sequencing, and as a generic experimental control. Selective agars were incubated with known positive and negative control strains with each set of transformations. Typically five or more transformed colonies were picked for long-term storage.

3.2.5. ANTIMICROBIAL PHENOTYPING

For the work in this PhD, susceptibility phenotyping was performed using both agar and broth-based methods. In all cases, EUCAST susceptibility breakpoints were used for the interpretation of susceptible (S), intermediate (I), and resistant (R) categories(5).

3.2.5.1 AGAR-BASED METHODS – THE ETEST® METHOD

Standardised inocula (using the 0.5 MacFarlane standard; and inoculating sterile peptone water) of fresh sub-cultures grown on Columbia Blood Agar (Oxoid, ThermoFisher Scientific, Basingstoke, Hampshire, UK) were prepared, and streaked evenly over an Iso-Sensitest Agar (ISA) plate (Oxoid, ThermoFisher Scientific,

Basingstoke, Hampshire, UK), with a sterile cotton swab soaked in the inoculated peptone water. ISA is developed and standardised commercially, meeting standards required for susceptibility testing in routine diagnostic microbiology.

Etest® (bioMérieux; Marcy l'Etoile, France) strips for the relevant antimicrobial being tested were then applied to the surface of the inoculated agar using sterile tweezers. Plates were incubated at 37°C in air overnight, and read the following morning. The minimum inhibitory concentration (MIC) of the antibiotic was measured off the strip at the point at which bacterial growth was inhibited.

3.2.5.2 BROTH MICRODILUTION METHOD

Broth microdilution susceptibility testing assays were done using the BD Phoenix Automated Microbiology System (Becton Dickinson, Franklin Lakes, New Jersey, USA). The Phoenix has an incubator-reader that enables the processing of large numbers of susceptibility testing panels that contain doubling dilutions of test antimicrobials. The Phoenix monitors each panel every 20 minutes using both turbidometric and colorimetric (oxidation-reduction indicator) growth detection(6).

I used the NMIC-84 test panel, enabling testing against the following antimicrobials: amikacin, amoxicillin-clavulanate (co-amoxiclav), ampicillin, aztreonam, cafazolin, cefepime, cefoxitin, ceftazidime, ceftriaxone, cefuroxime, ciprofloxacin, colistin, ertapenem, gentamicin, levofloxacin, meropenem, nitrofurantoin, piperacillin, piperacillin-tazobactam, temocillin, tobramycin, trimethoprim and trimethoprim-sulfamethoxazole. Briefly, bacterial inocula were prepared in proprietary susceptibility testing broths at a MacFarland standard of 0.5-0.6. They were then

inoculated with a dye, poured into the test cartridge, and placed in the reader-incubator overnight. Results were obtained the following day.

3.3. SEQUENCE DATA GENERATION

3.3.1. SEQUENCING METHODS

Frederick Sanger and colleagues developed the first sequencing method in 1977(7), using the incorporation of labelled dideoxynucleotides within separate sequencing reactions (one for each of the four nucleotides) to both terminate subsequent primer elongation along a DNA template, and act as a marker for the type of nucleotide that was represented at a given position with respect to the fragment being sequenced. Fragments for each of the four sequencing reactions were then separated using gel electrophoresis, and the nucleotide found at each position could be determined from the consecutive pattern of bands on the gel. This enabled the sequencing of fragments of the order of 400-900 bp.

From 1986 onwards, the labelling of these dideoxynucleotides with different fluorophores increased turnaround times by enabling a unification of the process in a single sequencing reaction, and allowed for optical, faster, and more automated reading of sequences on the basis of chromatograms generated by the incorporation of these fluorescently-labelled dideoxynucleotides during the sequencing reaction(8). Using this method, the first bacterial genome, that of *Haemophilus influenzae*, was sequenced in 1995, by randomly fragmenting *H. influenzae* DNA, inserting and amplifying these sequences in cloning vectors, sequencing the clones, and then using software to align overlapping fragments from the query sequence into a single contiguous sequence in a method known as “shotgun sequencing”(9, 10).

The evolution of next generation, massively parallel, sequencing methods in the 1990s and 2000s led to an explosion of sequencing data by decreasing turnaround times and costs. A number of methods were part of this revolution, including the sequencing-by-synthesis approach epitomised by the Illumina (formerly Solexa) platform and used for much of the work in this thesis, as well as the single-molecule real-time sequencing method represented by the Pacific Biosciences (PacBio) platform(11) (Figure 3.1.).

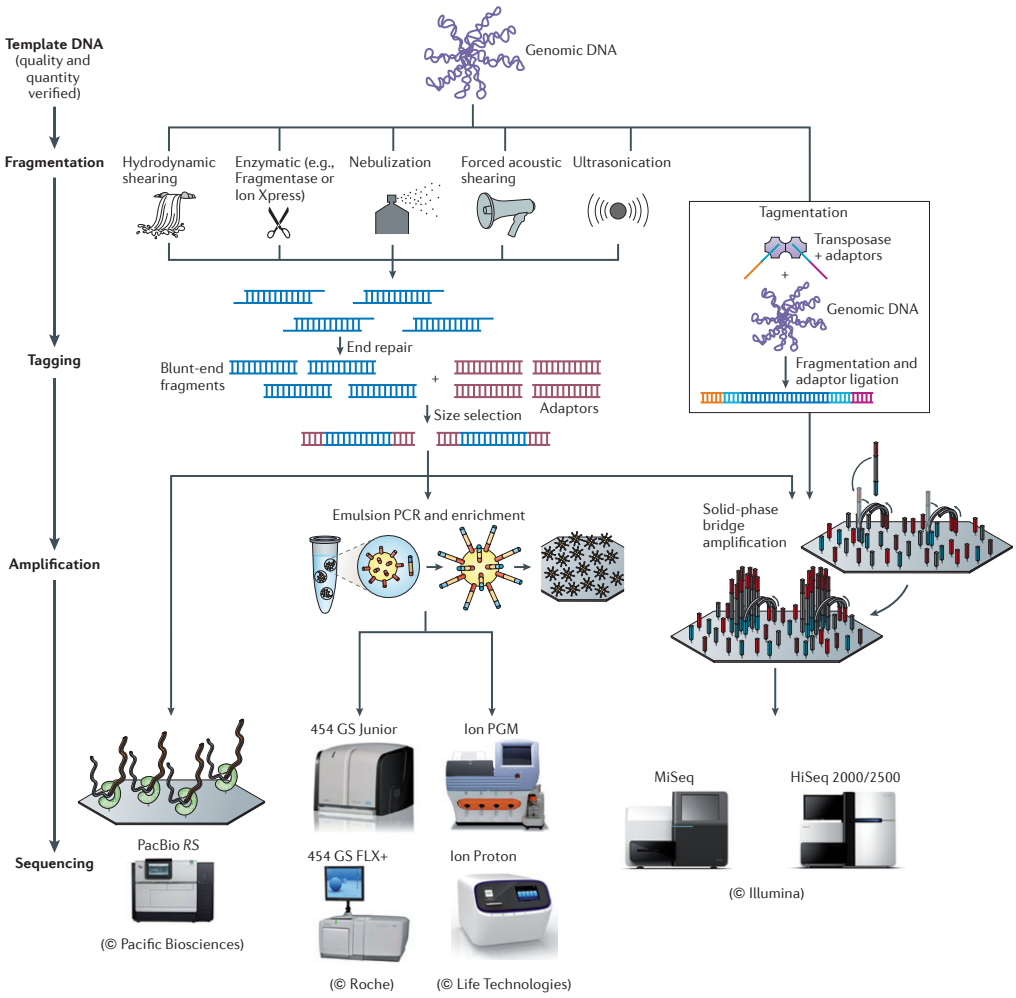


Figure 3.1. Common high-throughput sequencing platforms and their respective workflows. Taken from (12).

Both of these methods require the preparation of sequencing libraries. In the case of Illumina, early sequencing data in this thesis were generated using fragmentation by sonication, followed by use of the Illumina TruSeq library preparation kits; more

recently, data have been generated using the Nextera transposase-based fragmentation kits, which use transposase enzymes to randomly fragment the DNA and attach adaptors in a single step known as tagmentation. The adaptors: (i) are indexed to allow the pooling of samples; (ii) immobilise the template on a solid substrate (the flow cell); and (iii) enable a universal sequencing primer to bind allowing for a pre-sequencing amplification step and the generation of dense clusters of the same template.

For paired-end sequencing, DNA templates are prepared with a set of paired adaptors, each one of a pair attached to opposing ends of the fragmented DNA molecule. DNA fragments are typically relatively short, at 200-600bp long. Reads are generated by first sequencing one side of the molecule with the other end immobilised on the flow cell, and then sequencing the other end of the DNA template. Sequencing reads therefore only represent the nucleotides present at the end of the template sequenced, with the middle portion – the "insert" – remaining uncharacterised for any particular template. Each sequencing read in a paired-end set is described as the "mate" of the other read. The advantage of paired-end sequences is that they impart some positional information by relating two reads to each other with respect to their approximate genetic distance in the original DNA template – this information can be used to improve the accuracy of both read mapping and *de novo* assembly. Confusingly, a method known as "mate-pair" sequencing also exists, in which libraries of long DNA fragments with very large insert sizes of several kilobases are created. This method was not however used in this thesis and is not discussed further.

Illumina data are generated using a process of cyclic reversible termination: streams of fluorescently labelled nucleotides are passed over the clusters of template immobilised on the sequencing flow cell, and are either incorporated by DNA polymerases bound to the templates, or washed away. Incorporation of a nucleotide simultaneously terminates polymerisation, at which point the attached nucleotide is imaged. Each base call is a consensus of the nucleotide added across each template in a sequencing cluster. Fluorescent tags on incorporated nucleotides are then cleaved, the 3'-OH group is regenerated using the reducing agent tris(2-carboxyethyl)phosphine (TCEP) enabling further polymerisation, and the cycle begins again.

Error rates for Illumina data are some of the lowest reported for next generation sequencing technologies, with overall error rates of 0.3% for the Illumina HiSeq 2000/2500, and 0.8% for data generated on the Illumina Genome Analyser (GAIIx) and the MiSeq benchtop platform(13). All three of these platforms also vary in terms of their data output, run-times, read lengths and costs: For the HiSeq 2500 for example, 150-180 Gb of data using 150bp paired-end reads can be generated for 192 samples in 40 hours, giving rise to approximately 100-200x coverage for a 5Mb genome, at a cost of around £40/sample; for the MiSeq, the cost is approximately 2-fold higher, but data can be generated in a "benchtop" fashion for smaller batches within 24 hours.

For PacBio, used in some of the later work in this thesis, the sequencing approach is a single-molecule approach, in which DNA templates are extended by DNA polymerases immobilised in tiny wells, known as zero-mode waveguides (ZMWs).

Around 150,000 of these ZMWs are located on a sequencing cell, and the incorporation of labelled nucleotides by the polymerase is imaged simultaneously in real-time. For PacBio the major advantage is the capacity to generate very long reads (~10Kb), enabling closure of genomes. However, the error rates are of the order of 10-13%(13), requiring downstream read correction either with low-error data, such as those generated by Illumina, or by using high sequencing coverage, to enable bioinformatic auto-correction of errors in reads based on consensus calls made across a position in a deep read pile-up(14). The cost, at approximately £200-2000/sample depending on whether collaborative agreements are in place, is also generally prohibitive for sequencing large datasets; nevertheless, it can be extremely useful for generating dataset-specific references.

3.4. SEQUENCE READ PROCESSING

Raw sequencing reads generated by next generation sequencing platforms typically require downstream processing before they can be used for analyses. Two approaches are commonly used: (i) mapping/alignment of reads to a reference sequence, or (ii) *de novo* assembly, using a variety of algorithms to re-construct contiguous sequences (contigs) by calculating the extent of overlap of reads, without necessarily incorporating any external reference data.

Sequencing read information is typically summarised in a file format known as FASTQ (.fq), a text-based format storing read identifiers, their nucleotide sequences, and corresponding quality scores(15). The sequencing read identifiers are unique, and incorporate information about the instrument, the run, and the flowcell used, and for paired-end sequences include a “1” or “2” tag which can be used in linking paired

sequences together. Nucleotide data are represented as “A”, “C”, “G”, or “T”, with quality scores for each base on a separate line, again as a series of standardised text-based characters. The quality score defines the probability of error associated with the sequencing base call. This is now mostly based on a metric known as the Phred score, defined as $Q_{phred} = -10\log_{10}e$, where e is the estimated probability of a base being wrong. A Phred score of 10 therefore represents a 1 in 10 chance of the base having been called incorrectly (base call accuracy of 90%), a score of 20 a base call accuracy of 99%, a score of 30 a base call accuracy of 99.9%, and so on. Acceptable quality is commonly defined as any base carrying a Phred score of 20 or higher(16).

3.4.1. MAPPING-BASED APPROACHES AND VARIANT CALLING

A large number of methods have been developed for read mapping, and offer different trade-offs in terms of the number of mismatches allowed, speed, memory requirements, and the capacity to cope with gapped alignments allowing for the interpretation of indels(17). For this thesis, most reference-based mapping was undertaken as part of the standard group pipeline, using a software called Stampy(18). This was written for Illumina data, has high specificity, incorporates the use of base quality scores and paired-end information, and is sensitive to the identification of short indels. Its major downside is its speed, and the inability to parallelise its functions, resulting in a processing time of approximately six hours for a non-enteropathogenic *E. coli* genome. For additional ex-pipeline mapping, a second software called BWA was also used(19). This is a much faster approach, which can also identify indels, but does not include base quality scores in its algorithm.

Following read mapping, a summary of the number and quality of the reads mapping at each site is produced, in this case with SAMTools(20). In the same way that each base within a read is assigned a quality score (the Phred score) during the sequencing process, which defines the probability of error associated with the sequencing base call, reads are given a mapping quality score which determines the probability that they have been inappropriately aligned to the reference at that position. The combination of coverage, base and mapping quality scores can subsequently be used to filter output data, flag low confidence base calls, and determine variation with respect to the reference with great sensitivity and specificity, and this information is summarised in variant call format (.vcf) file. This approach does not however enable an assessment of variation within regions that are not present in the reference sequence, and can fail to accurately represent highly divergent areas, or repetitive regions. For the group pipeline, repetitive regions within the reference are determined with a self-self nucleotide BLAST and masked out prior to mapping.

During the thesis, mapping was carried out to a number of different references depending on the requirements of the dataset, but standard references were also used to enable comparisons across all sequenced isolates, and to estimate the error rate for the group's mapping algorithm with respect to non-enteropathogenic *E. coli* and *K. pneumoniae*.

3.4.2. *DE NOVO* ASSEMBLY METHODS

De novo assembly methods attempt to reconstruct genomes without reference to other sequences, typically to assess non-core components of the genome such as mobile genetic elements. Most assemblers rely on one of three assembly paradigms: (i)

Greedy, in which the assembler always makes the choice that is associated with the most immediate benefit, but does not take into account any global associations between reads (such as incorporating the use of paired read information); (ii) Overlap-layout-consensus (OLC), in which reads that overlap sufficiently well are represented as nodes in a graph, with overlaps represented as edges; (iii) de Bruijn graph algorithms, where nodes in the graph represent substrings of length k extracted from the input reads (k-mers), and edges represent overlaps of k-1. Greedy methods have been limited by their inherently local approach; OLC methods are computationally intensive, a limitation which has been partly overcome by the implementation of string graph approaches, where redundant information is iteratively removed from the graph; and de Bruijn graph algorithms rely on exact matches, and can therefore only be used for data with low sequencing errors(21).

The *de novo* assembly method originally selected for the group pipeline consists of a de Bruijn graph-based algorithm encompassed in a software package known as Velvet(22). Automated optimisation of parameter values for the construction of the assemblies is achieved using the VelvetOptimizer program(23). As a comparison to the use of this approach, additional assembly methods were explored for the analysis of sequenced non-enteropathogenic *E. coli* and *K. pneumoniae* (this is described in greater detail in chapter 4: Thesis Methods – evaluation).

3.5. PHYLOGENETIC ANALYSIS AND COMPARISONS OF SEQUENCE DATA

Phylogenetics relates to the study of the evolutionary history of genetic units such as species or genes, by classifying their relatedness through trees or phylogenies.

Phylogenies enable the determination of shared properties of members of each genetically related group or clade, thus enabling us to infer ancestral properties on the basis of sampled datasets.

3.5.1. SEQUENCE ALIGNMENT

Sequence alignment seeks to identify the homologous nucleotides in two or more genomes, and in this way identify nucleotides that have descended from a single site in an ancestral organism. Sequence alignment is notoriously difficult, and universal, effective approaches for benchmarking the appropriateness of alignment strategies have not been fully defined(24-26). Computational approaches to sequence alignment generally fall into two groups: (i) global alignment, which forces the alignment to span the entire length of all query sequences, and (ii) local alignment, which identifies regions of similarity, but is not trying to infer similarity on a larger scale. Semi-global alignment is a hybrid of the two methods, and can be useful when comparing sequences of hugely different lengths, such as a chromosome and a gene. Most alignment methods depend either on slow and accurate dynamic programming algorithms, or on faster but less accurate heuristic or probabilistic methods, which are designed for large-scale problems.

Sequence alignment can be restricted to two sequences, known as pairwise alignments, or can be used to align multiple sequences, known as multiple sequence alignment. Pairwise alignment methods include: (i) dot-matrix methods, in which the two sequences are set up as two-dimensional matrix and dots are placed at any point where characters match; (ii) dynamic programming methods, which typically assign positive scores to sequence matches, negative scores to mismatches and negative

penalties to gaps – the "correct" alignment is then the one with the highest score; and (iii) word methods, which are an example of fast, heuristic-based approach in which relative positions of short, non-overlapping subsequences of the query sequences (words) are compared and more detailed sensitive alignment is only applied to regions identified as being similar. Multiple sequence alignment typically relies on: (i) progressive or hierarchical methods in which the most similar sequences are aligned first, and other sequences are then incorporated in a stepwise fashion; and/or, (ii) iterative methods, in which a global alignment of sequences is defined first, and used to realign sequence subsets, which are then aligned to produce the next iteration's global alignment.

In the context of whole bacterial genomes, as part of evolution, both local and large-scale genetic changes can occur, with the former typically affecting a smaller number of nucleotides and incorporating substitution and small insertion-deletion (indel) events, and the latter including loss, gain or duplication of large segments of DNA, generated by unequal recombination. A further complication is that sets of genomes may contain pairs of sequence positions whose evolutionary history can be defined by any of the major sub-classes of homology, namely orthology, paralogy and xenology (Figure 3.2.). Traditionally, orthologous segments are considered of most importance for generating core genomic phylogenies, although this may not be true in every case(27).

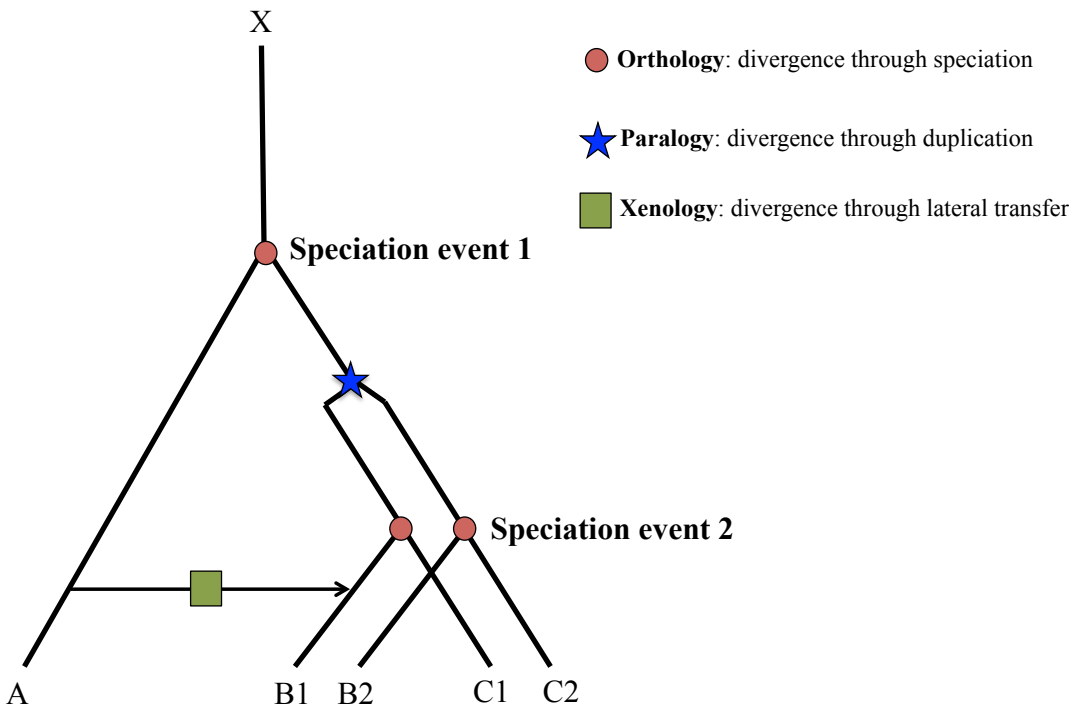


Figure 3.2. Representation of the sub-classes of homology, where "X" is the common ancestor, and orthology, paralogy and xenology represent the major mechanisms of diversification.

There are currently a limited number of approaches to whole-genome alignment, and these tend to specifically predict relationships between orthologous sequences. Some of the main methods have been used in this thesis, namely a hierarchical/iterative approach, using a software called ProgressiveMauve, and the local approach, using NUCmer as part of the MUMmer package(28). For ProgressiveMauve, an initial genome alignment is split into a subset of smaller "global" alignment problems through the identification of collinear (i.e. unaffected by genomic rearrangement since their most recent common ancestor), homologous segments and their representation in a homology map; these segments are then individually subjected to nucleotide level alignments. ProgressiveMauve is able to identify and represent rearrangements in genetic segments, and a degree of sequence variability within collinear blocks. For MUMmer, a large set of nucleotide level alignments are produced first, which are then merged and filtered to produce multiple sets of pairwise alignments of

homologous sequences. MUMmer only handles comparisons between pairs of genomes, whereas ProgressiveMauve can be used for multiple sequence alignment.

Challenges in association with whole genome alignment include the capacity to scale comparisons to large numbers of genomes, and to enable an assessment of the degree of uncertainty associated with a particular alignment. In addition, validation of the accuracy of these methods is difficult given the lack of knowledge surrounding the "true" evolutionary history of a set of sequences. Validation of the whole genome alignment methods used was beyond the scope of this thesis, but it is necessary to be aware of some of their limitations.

3.5.2. TREE-BUILDING APPROACHES AND SOFTWARE PACKAGES

There are several tree-building methods available – these can be broadly classified as distance or character-based methods(29). Distance-based methods include the unweighted pair group method using arithmetic averages (UPGMA) and the neighbour-joining method; these use a model of evolution to generate a distance matrix between sequences (based on nucleotide differences in this thesis) from which a tree is calculated by means of progressive clustering. The main advantages of these methods is that they are fast, and can easily handle large numbers of sequences. The main disadvantage of UPGMA trees is that they assume the same evolutionary speed for all lineages, which is unlikely, particularly for divergent lineages; for this reason, neighbour-joining methods are typically the preferred distance-based method.

Character-based methods operate directly on the sequence alignment data, and can be used to infer ancestral sequences and evolutionary parameters. The three main

methods here are: (i) maximum-parsimony (MP), in which the minimum number of substitutions over all sites (known as the tree length) is computed for each possible tree topology, and the MP tree is the tree with the minimum tree length; (ii) Maximum-likelihood (ML), in which the likelihood of observing a given set of sequence data for a specific model of evolution is maximised for each topology, and the tree with the highest maximum-likelihood is then chosen; and (iii) Bayesian inference, which differs from ML in that model parameters chosen are not unknown fixed constants, but are random variables with statistical distributions. In Bayesian methods, parameters are assigned a prior distribution before analysis, which is then combined and updated with the available data during the analysis to generate a posterior distribution.

ML and Bayesian methods are considered the most robust of the tree-building methods; they are also the most computationally intensive. The major components in model selection typically include: (i) what proportion of sites are described as invariant; (ii) the choice of nucleotide substitution model specifying the matrix defining the rates of change between each of the nucleotide states, and (iii) whether nucleotide substitution rates can vary across sites. There is scope for modelling a number of other parameters using some Bayesian methods, such as an estimate of the rate of evolution within the dataset – the so-called "molecular clock".

For this thesis a combination of ML software packages was used – initially PhyML(30), and subsequently RaxML(31), which is faster and can be parallelised, but outputs topologies that are thought to be equally robust. Bayesian phylogenetic analysis was undertaken using BEAST(32) mostly due to the expertise available

within the team, although there are other software packages available, such as MrBayes(33) and ClonalFrame(34). For maximum-likelihood approaches, the standard nucleotide substitution model used was the generalised time-reversible model (GTR), which allows for different rates of change between all four nucleotide character states. In addition, the models were set-up to include the possibility of four relative rates of change across sites, but allowed all sites to be subject to an evolutionary rate (i.e. the proportion of invariant sites was fixed at 0%). For ML methods trees were bootstrapped (sequences resampled with replacement) to generate a measure of statistical support for the nodes represented in the final consensus topology, with nodes closer to 1 (or 100%) being better supported; for Bayesian methods, posterior support values for tree nodes are generated as part of the assessment of the tree topology.

3.5.3. ANNOTATION

Genome annotation – essentially the attribution of a biological function to raw nucleotide sequence - requires a number of steps(35). In the initial stage, several bioinformatic methods are linked up, typically in a pipeline structure, to predict the likely location of genes and to describe the likely function of gene products. This is done by making reference to a series of databases such as GenBank/EMBL, an annotated collections of all publicly available DNA sequences (http://www.ncbi.nlm.nih.gov/Genbank/; http://www.ebi.ac.uk/embl/), HAMAP, a database of annotated proteomes and protein families (http://ca.expasy.org/sprot/hamap/), or EcoCyc (http://ecocyc.org/), describing annotations and metabolic pathways specific to *E. coli* K-12. For nucleotide level annotation, a number of prediction software modules can then be applied, such as

GLIMMER [http://www.genomics.jhu.edu/Glimmer/] or GeneMark [http://exon.gatech.edu/genemark/], which predict coding sequences for proteins, and Aragorn and RNAmmer [http://lowelab.ucsc.edu/tRNAscan-SE/, http://www.cbs.dtu.dk/services/RNAmmer/], which are used to identify tRNA and rRNA motifs respectively. For this thesis a number of methods were used – initially web-based annotation platforms, namely BASys and RAST [http://wishart.biology.ualberta.ca/basys], which have turnaround times of the order of days per sequence, and then PROKKA, a command-line based prokaryotic annotation pipeline that can be downloaded and installed on the server locally and can annotate a 5Mb genome in less than 20 minutes(36). PROKKA uses a combination of tools to predict the coordinates of protein coding sequences (Prodigal), define tRNA and rRNA sequences (RNAmmer, Aragorn), signal leader peptides (SignalP), and non-coding RNA (Infernal). Putative gene products are assigned to coding sequences in a hierarchical manner by comparisons to a number of databases using BLAST+, initially UniProt, then proteins from finished bacterial genomes in RefSeq, and then a series of hidden Markov model profile databases, including Pfam and TIGRFAMS (using hmmscan). If no matches can be found, coding sequences are labelled as "hypothetical proteins".

Annotated genomes enable an alternative approach to comparative genomics circumventing some of the difficulties associated with multiple sequence alignment, enabling the clustering of nucleotide or amino acid sequences defined as being of likely coding relevance and their classification into core (present in all isolates) or accessory components (present in at least one isolate) of the pangenome (representing all genetic content) for a specific dataset. An example of a clustering software used

for this purpose is CD-HIT(37). In this thesis, this approach was used as one of the method for determining the similarity of plasmid transformants generated in the population structure analysis of an important global non-enteropathogenic *E. coli* lineage, namely ST131 (Chapter 8).

3.6. SUMMARY

A number of different methods were used at all stages of the thesis, from laboratory methods, sequencing preparation and technologies, to raw data processing and genetic inference. The rationale for different approaches was partly driven by the rapid evolution of sequencing technologies and analysis approaches during the project, as well as logistical issues including the availability of hardware resource for computing and access to PacBio sequences. Every approach has benefits and disadvantages, and an awareness of these is important in the interpretation of the data. Although it was not possible to carry out a comparative evaluation of all the methods presented here, a limited evaluation of the robustness of some of my more widely used methods are presented in the next chapter.

CHAPTER 3 REFERENCES

1. **Ahn SJ, Costa J, Emanuel JR.** 1996. PicoGreen quantitation of DNA: effective evaluation of samples pre- or post-PCR. Nucleic acids research **24:**2623-2625.
2. **Birnboim HC, Doly J.** 1979. A rapid alkaline extraction procedure for screening recombinant plasmid DNA. Nucleic acids research **7:**1513-1523.
3. **Miller JF, Dower WJ, Tompkins LS.** 1988. High-voltage electroporation of bacteria: genetic transformation of Campylobacter jejuni with plasmid DNA. Proceedings of the National Academy of Sciences of the United States of America **85:**856-860.
4. **Durfee T, Nelson R, Baldwin S, Plunkett G, 3rd, Burland V, Mau B, Petrosino JF, Qin X, Muzny DM, Ayele M, Gibbs RA, Csorgo B, Posfai G, Weinstock GM, Blattner FR.** 2008. The complete genome sequence of Escherichia coli DH10B: insights into the biology of a laboratory workhorse. Journal of bacteriology **190:**2597-2606.
5. **Testing ECoAS.** 2014. Breakpoint tables for interpretation of MICs and zone diameters.
6. **Jorgensen JH, Ferraro MJ.** 2009. Antimicrobial susceptibility testing: a review of general principles and contemporary practices. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America **49:**1749-1755.
7. **Sanger F, Nicklen S, Coulson AR.** 1977. DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences of the United States of America **74:**5463-5467.

8. **Smith LM, Sanders JZ, Kaiser RJ, Hughes P, Dodd C, Connell CR, Heiner C, Kent SB, Hood LE.** 1986. Fluorescence detection in automated DNA sequence analysis. Nature **321:**674-679.

9. **Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM, et al.** 1995. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science **269:**496-512.

10. **Staden R.** 1979. A strategy of DNA sequencing employing computer programs. Nucleic acids research **6:**2601-2610.

11. **MacLean D, Jones JD, Studholme DJ.** 2009. Application of 'next-generation' sequencing technologies to microbial genetics. Nature reviews. Microbiology **7:**287-296.

12. **Loman NJ, Constantinidou C, Chan JZ, Halachev M, Sergeant M, Penn CW, Robinson ER, Pallen MJ.** 2012. High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity. Nature reviews. Microbiology **10:**599-606.

13. **Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, Bertoni A, Swerdlow HP, Gu Y.** 2012. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC genomics **13:**341.

14. **Chin CS, 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. Nature methods **10:**563-569.

15. **Cock PJ, Fields CJ, Goto N, Heuer ML, Rice PM.** 2010. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic acids research **38:**1767-1771.

16. **Richterich P.** 1998. Estimation of errors in "raw" DNA sequences: a validation study. Genome research **8:**251-259.

17. **Wikipedia** 2014, posting date. List of sequence alignment software. [Online.]

18. **Lunter G, Goodson M.** 2011. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. Genome research **21:**936-939.

19. **Li H, Durbin R.** 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics **26:**589-595.

20. **Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S.** 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics **25:**2078-2079.

21. **Nagarajan N, Pop M.** 2013. Sequence assembly demystified. Nature reviews. Genetics **14:**157-167.

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

23. **Gladman SS, T.**, posting date. VelvetOptimiser. [Online.]

24. **Iantorno S, Gori K, Goldman N, Gil M, Dessimoz C.** 2014. Who watches the watchmen? An appraisal of benchmarks for multiple sequence alignment. Methods in molecular biology **1079:**59-73.

25. **Dewey CN.** 2012. Whole-genome alignment. Methods in molecular biology **855:**237-257.

26. **Loytynoja A.** 2012. Alignment methods: strategies, challenges, benchmarking, and comparative overview. Methods in molecular biology **855:**203-235.

27. **Studer RA, Robinson-Rechavi M.** 2009. How confident can we be that orthologs are similar, but paralogs differ? Trends in genetics : TIG **25:**210-216.

28. **Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL.** 2004. Versatile and open software for comparing large genomes. Genome biology **5:**R12.

29. **Sleator RD.** 2013. A beginner's guide to phylogenetics. Microbial ecology **66:**1-4.

30. **Guindon S, Delsuc F, Dufayard JF, Gascuel O.** 2009. Estimating maximum likelihood phylogenies with PhyML. Methods in molecular biology **537:**113-137.

31. **Stamatakis A, Ludwig T, Meier H.** 2005. RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. Bioinformatics **21:**456-463.

32. **Drummond AJ, Rambaut A.** 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC evolutionary biology **7:**214.

33. **Huelsenbeck JP, Ronquist F.** 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics **17:**754-755.

34. **Didelot X, Falush D.** 2007. Inference of bacterial microevolution using multilocus sequence data. Genetics **175:**1251-1266.

35. **Medigue C, Moszer I.** 2007. Annotation, comparison and databases for hundreds of bacterial genomes. Research in microbiology **158:**724-736.

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

37. **Li W, Godzik A.** 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics **22:**1658-1659.

CHAPTER 4: THESIS METHODS - EVALUATION

4.1. INTRODUCTION

In order to determine the robustness of the pipeline mapping, variant calling and *de novo* assembly approaches for the two species under investigation, a set of "technical replicates" of known, fully sequenced reference strains (obtained from central culture collections) was established. These "technical replicates" encompassed replication at the following levels: (i) sequencing of the same DNA extract, (ii) sequencing of DNA extracted from the same frozen stock of isolate but subcultured and extracted on several occasions, (iii) sequencing of DNA extracts from (i) and (ii) above in different sequencing runs and flowcells.

4.2. ROBUSTNESS OF MAPPING AND VARIANT CALLING ALGORITHMS

For standard pipeline mapping of non-enteropathogenic *E. coli*, I used the reference *E. coli* strain CFT073 (RefSeq accession: NC_004431), which has a chromosomal length of 5,231,428bp, and is devoid of episomal content. Repetitive regions identified by self-self comparisons using BLASTn represent 303,037bp or 5.8% of the genome. Eighteen replicates were sequenced, as shown in Table 4.1.a. A median of 2,454,289 reads mapped (range: 1,237,346-14,625,659), representing 99.6% of reads generated in sequencing (range: 98.3-99.7%), and 4,772,918 sites in the reference were called across all replicates (91.2% of the reference; 96.8% of the masked reference). No genetic differences were observed between any of these technical replicates (error rate of 0 single nucleotide variants [SNVs]/genome; 95% CI: 0 to 0.204 SNVs/genome).

For standard pipeline mapping of *K. pneumoniae*, I used the reference *K. pneumoniae* strain MGH78578 (RefSeq accession: NC_009653.1), which has a chromosomal length of 5,315,120bp. This strain additionally has five plasmids: pKPN3 (175,879bp), pKPN4 (107,576bp), pKPN5 (88,582bp), pKPN6 (4,259bp), and pKPN7 (3,478bp), although mapping was restricted to the chromosomal reference only. Repetitive regions identified by self-self comparisons using BLASTn were excluded from variant calling as previously, and represent 149,107bp or 2.8% of the genome. Seventeen replicates were sequenced, as shown in Table 4.1.b. A median of 1,858,016 reads mapped (range: 1,245,217-7,615,552), representing 89.3% of sequencing reads generated (range: 87-89.8%), and 5,019,242 sites in the reference were called across all replicates (94.4% of the reference; 97.1% of the masked reference). Again, no genetic differences were observed between any of these technical replicates (error rate of 0 SNVs/genome; 95% CI: 0 to 0.200 SNVs/genome).

This therefore demonstrates that the group's pipeline variant calling algorithm was is an accurate and reproducible method of core nucleotide variant detection across species, time, extraction and sequencing batches, in the context of mapping to these two references.

Sample Name	DNA extract	Sequencing batch
CFT073_I1_E2	A	1
CFT073_I1_E2	A	2
CFT073_I1_E3	B	3
CFT073_I1_E4	C	4
CFT073_I1_E4	C	5
CFT073_I1_E5	D	6
CFT073_I1_E5	D	7
CFT073_I1_E12	E	8
CFT073_I1_E12	E	8
CFT073_I1_E12	E	8
CFT073_I1_E11	F	8
CFT073_I1_E11	F	8
CFT073_I1_E11	F	8
CFT073_I1_E10	G	8
CFT073_I1_E10	G	8
CFT073_I1_E10	G	8
CFT073_I1_E10	G	8
CFT073_I1_E10	G	9

Table 4.1.a. Details of non-enteropathogenic *E. coli* replicates sequenced. Identical letters denote sequencing from identical DNA extracts; identical numbers denote identical sequencing runs and flowcells.

Sample Name	DNA extract	Sequencing batch
MGH78578_I1_E1	A	1
MGH78578_I1_E11	B	2
MGH78578_I1_E11	B	2
MGH78578_I1_E11	B	2
MGH78578_I1_E12	C	2
MGH78578_I1_E12	C	2
MGH78578_I1_E12	C	2
MGH78578_I1_E12	C	3
MGH78578_I1_E12	C	3
MGH78578_I1_E13	D	2
MGH78578_I1_E13	D	2
MGH78578_I1_E13	D	2
MGH78578_I1_E13	D	2
MGH78578_I1_E2	E	4
MGH78578_I1_E2	E	5
MGH78578_I1_E2	E	6
MGH78578_I1_E2	E	7

Table 4.1.b. Details of *K. pneumoniae* replicates sequenced. Identical letters denote sequencing from identical DNA extracts; identical numbers denote identical sequencing runs and flowcells.

4.3. ROBUSTNESS OF *DE NOVO* ASSEMBLIES

An assessment of the reliability and reproducibility of the group pipeline's *de novo* assembly approach compared with other *de novo* assemblers was undertaken for sequenced reference strains, again in order to characterise error rates and the nature of errors introduced, and to see whether other approaches might be more useful for subsets of data analysed for this work.

Standard, optimised Velvet assemblies undertaken using the group's pipeline were compared with three other assembly methods – two of which had achieved the highest ratings in a published comparison of methods for bacterial *de novo* assembly (B-GAGE)(1), namely SPAdes(2) and the Maryland Super-Read Celera Assembler (MaSuRCA)(3). The third published method used in the comparisons was Andrew And Aaron's Awesome Assembly (A5) pipeline(4).

SPAdes implements a multi-sized version of the de Bruijn graph assembly method, allowing for different sizes of k-mer (*k*) within the same assembly. Smaller values of *k* tend to collapse DNA repeats together, but enable assembly in lower coverage regions and reduce fragmentation in these contexts; larger values of *k* facilitate the assembly of repeat regions, but are not adapted to assembling through low coverage regions.

MaSuRCA contains an algorithm for the creation of "super-reads", which are extensions of sequenced reads. The basic concept of this is to extend each original read in both directions, as long as the next base added in the extension is unique. In the first instance, a *k*-mer count look-up table is created, containing the counts of all

possible *k*-mers in the read dataset. Given a *k*-mer found at the end of a read, there are four possibilities by which this could be extended, namely the string formed by appending A, C, G or T to the last *k*-1 bases in the read. This set of hypothetical *k*-mers is compared with the look-up table, and if only one of the four possible *k*-mers is found, this is considered a unique following *k*-mer, and that base is then appended to the end of the read. This process continues at both ends of the read until the read can no longer be extended uniquely, creating a smaller set of longer reads (super-reads). From this dataset, a smaller subset of maximal super-reads is created, including only those super-reads that are not exact substrings of another super-read. A modified version of the CABOG (Celera Assembler with Best Overlap Graph) assembler(5), an OLC-type algorithm, is then applied to this dataset of maximal super-reads and is used to create the final assembly.

The A5 pipeline emerged out of an evaluation of numerous assembly algorithms, and is an end-to-end pipeline which takes raw fastq files (quality-tagged raw read data), and outputs the final scaffolded assembly. The pipeline includes steps for data cleaning, error correction, contig assembly and assembly checks, and scaffolding, using a number of previously developed software tools, in conjunction with a novel algorithm for assessing assembly quality control (Figure 4.2). It takes around 2 hours to assemble a complete non-enteropathogenic *E. coli* genome.

All of these assemblers have the option of additional scaffolding steps, which will attempt to join contigs together using, for example, positional information derived from estimated insert sizes and read-pairing information. In our standard pipeline, the scaffolding option is turned on, and it is unclear whether this is prone to creating

further misassemblies. Comparisons were therefore also made with and without scaffolding for each of the assemblers.

The same parameters for each of the read datasets were used for each of the assemblers as follows:

1. For Velvet – this was performed as part of the group's pipeline using the VelvetOptimiser wrapper, run for the maximum range of *k*-values from the default lowest possible value of 19 to the maximum *k*-mer length in the dataset being processed.
2. For SPAdes (version 3.0.0):

   -python ./spades.py -t 1 -m 32 --careful -1 ./input_reads_forward_1.fq -2 ./input_reads_reverse_2.fq -o ./output_directory_name

   -t represents the number of cores to be used for processing

   -m specifies the memory limit in Gb

   --careful invokes a module which attempts to reduce the number of mismatches and short indels, and is recommended by the authors

   default parameters include *k*-mer values (default lengths: 21,33,55) for assemblies, and auto-detection of PHRED quality cut-offs for the input reads.
3. For MaSuRCA – the config file was set up for auto-optimisation of k-mer lengths (GRAPH_KMER_SIZE=auto), Illumina reads (USE_LINKING_MATES=1), the jellyfish hash size as advised (JF_SIZE=5000000), homopolymer trimming turned off (DO_HOMOPOLYMER_TRIM=0)
4. For A5 (Figure 4.2.) – the pipeline was invoked using the command:

./a5_pipeline.pl ./input_reads_forward.fq ./input_reads_backward.fq

./output_prefix

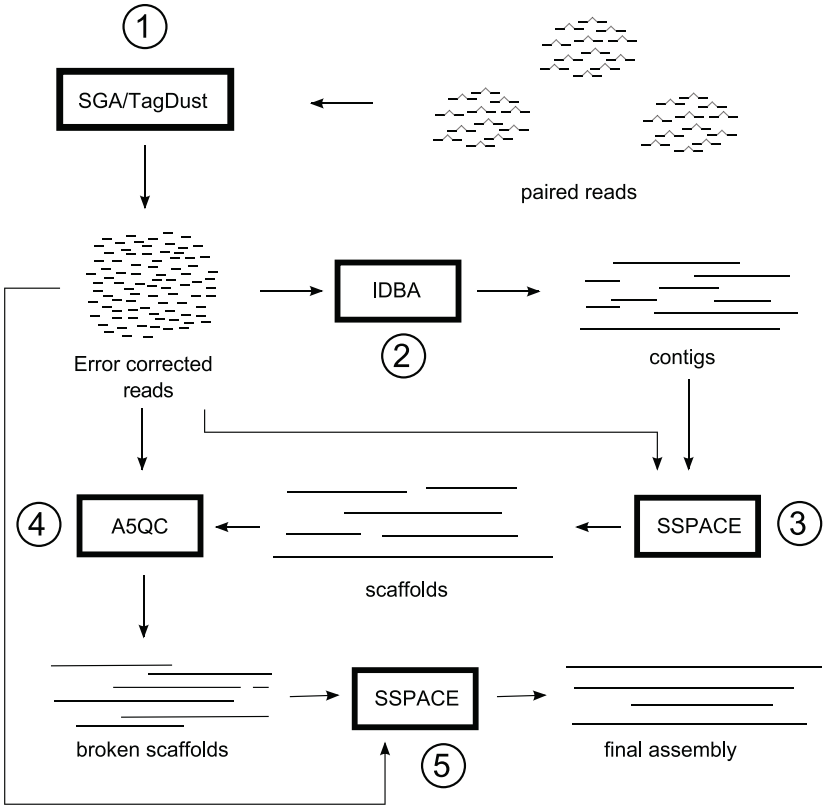


Figure 4.2. Overview of the stages in the A5 pipeline. The first stage cleans reads. These are then assembled with a de-Bruijn-based assembler, IDBA and scaffolded using the original read set with SSPACE. Scaffolds are then quality controlled and re-scaffolded using the original readset(4).

For both the non-enteropathogenic *E. coli* CFT073 reference and the *K. pneumoniae* MGH78578 reference, I took data from three extracts of each reference strain which had been sequenced a number of times: in both cases two extracts in triplicate, and one extract four times (these were all sequencing datasets which had been shown to be indistinguishable on reference-based mapping above). Each sequence dataset was assembled twice with each of the software programs as specified, with scaffolding features turned on and off, respectively. The assembly was then compared to the known reference sequence, and quality metrics were extracted and visualised using the QUAST software package(6).

The summary across the ten replicates of CFT073 is presented in Table 4.3.a., with a visual representation in Figure 4.4. panels a-d. Scaffolded Velvet data is sub-optimal if it is being used to assess genetic structure and gene synteny, but can be used to retrieve content for ~96% of genes. MaSuRCA "overassembled " the genome, and generated the largest number of contigs and the largest number of misassemblies with or without scaffolding. SPAdes produced the most consistent assemblies across the different sequenced replicates, but generates a number of misassemblies. Unscaffolded A5 resulted in the lowest number of misassemblies, and was therefore selected for use for *de novo* assemblies performed ex-pipeline for this thesis; any scaffolding was shown to introduce errors, and was therefore avoided. This comparison also demonstrated that any assessment of nucleotide-level variation using *de novo* assemblies would hugely over-estimate the number of SNVs present. Unscaffolded A5 was the least error-prone in this respect, but for a 5Mb genome would still be calling around 200 false-positive SNVs. *De novo* assemblies could therefore ideally only be used for gene-level or structural similarities, or as a further reference for re-mapping to define SNVs.

For *K. pneumoniae* MGH78578, the summary across the datasets is presented in Table 4.3.b. The graphical depictions are very similar to that for the non-enteropathogenic *E. coli* data, and are therefore omitted here.

Quality metric	Velvet scaffolded	Velvet unscaffolded	A5 scaffolded	A5 unscaffolded	SPAdes scaffolded	SPAdes unscaffolded	MaSuRCA scaffolded	MaSuRCA unscaffolded
Total length (>500bp)	5144375	5143564	5144375	5087103	5130510	5128978	5505425	5504255
Number of contigs (>500bp)	78	163	78	240	142	148	231	251
Largest contig	684449	342377	684449	293541	425198	425198	310433	304050
Percentage of reference assembled	98	99	98	98	99	99	105	105
N50	177360	101461	177360	73047	135213	111509	87400	78776
Number of contigs with misassemblies	10	1	10	0	3	2	24	16
Length of misassembled contigs	1722309	46641	1722309	0	192386	157664	886638	387963
Number of contigs unaligned to reference	0	1	0	0	6	6	1	1
Number of contigs partially unaligned to reference	0	1	0	0	1	1	10	11
Unaligned length	183	979	183	3	4738	4774	3520	4293
Percentage of bases in assembly aligned to reference	98	98	98	97	98	98	99	99
Error statistics (per 100kbp of assembly, compared to reference)								
Average number of uncalled bases	6	0	6	0	21	0	8	0
Average number of miscalled bases	7	6	7	4	13	11	36	36
Average number of indels	5	4	5	4	4	4	6	6
Gene-based metrics (compared to reference)								
Number of complete genes in assembly	5153	5207	5153	5100	5203	5201	5250	5250
Percentage of complete genes in assembly	96	97	96	95	97	97	98	98
Number of partial genes in assembly (>100bp present)	147	101	147	161	83	84	91	91
Number of complete + partial genes in assembly	5301	5308	5301	5261	5286	5285	5341	5341
Percentage of complete and partial genes in assembly	99	99	99	98	98	98	99	99

Table 4.3.a. Mean quality metrics for ten sequenced datasets of the non-enteropathogenic *E. coli* CFT073 strain, when compared to the published reference sequence. The N50 is defined as the cut-off length for which contigs of that length or longer contain at least half of the total length of the assembly.

Quality metric	Velvet scaffolded	Velvet unscaffolded	A5 scaffolded	A5 unscaffolded	SPAdes scaffolded	SPAdes unscaffolded	MaSuRCA scaffolded	MaSuRCA unscaffolded
Total length (>500bp)	5547389	5542462	5558607	5515563	5562328	5559498	5986770	5985774
Number of contigs (>500bp)	146	204	88	189	128	135	280	300
Largest contig	440855	248377	546453	200921	524692	435301	280405	247225
Percentage of reference assembled	97	97	98	97	98	98	105	105
N50	153361	81541	215316	85468	211844	179072	82096	70569
Number of contigs with misassemblies	3	2	13	1	3	2	27	20
Length of misassembled contigs	372146	130568	1663659	13306	643295	206185	1461632	755296
Number of contigs unaligned to reference	0	0	0	0	9	9	1	1
Number of contigs partially unaligned to reference	1	1	1	1	1	1	22	23
Unaligned length	20	327	806	31	6759	6705	6679	7789
Percentage of bases in assembly aligned to reference	97	97	97	97	97	97	98	98
Error statistics (per 100kbp of assembly, compared to reference)								
Average number of uncalled bases	70	0	7	0	26	0	8	0
Average number of miscalled bases	5	5	5	1	7	6	32	32
Average number of indels	2	1	2	1	1	1	3	3
Gene-based metrics (compared to reference)								
Number of complete genes in assembly	4992	4959	4992	4964	5039	5038	5048	5048
Percentage of complete genes in assembly	96	96	96	96	97	97	97	97
Number of partial genes in assembly (>100bp present)	84	117	106	96	46	48	87	88
Number of complete + partial genes in assembly	5076	5076	5097	5060	5085	5085	5135	5135
Percentage of complete and partial genes in assembly	98	98	98	98	98	98	99	99

Table 4.3.b. Mean quality metrics for ten sequenced datasets of the *K. pneumoniae* MGH78578 strain, when compared to the published reference sequence.

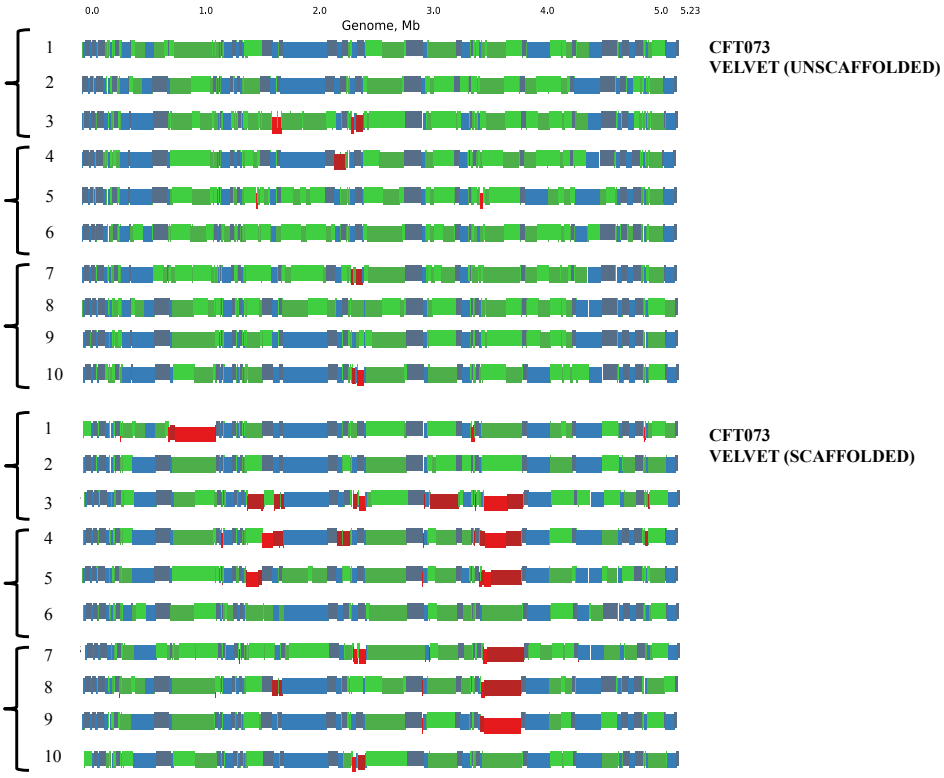
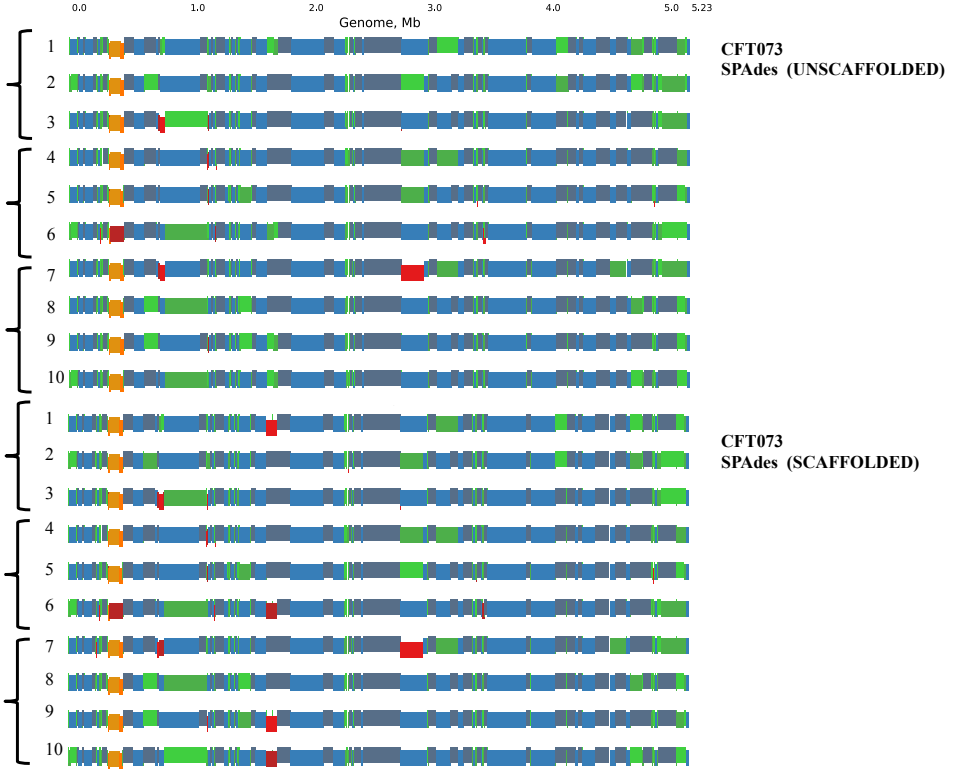


Figure 4.4. panels a and b. Comparisons of *de novo* assembled non-enteropathogenic *E. coli* CFT073 datasets using Velvet (top) and SPAdes (bottom). Each contig is represented by a block, staggered to represent contigs boundaries. Contigs consistent with the reference that align across the dataset are (i) coloured blue if the boundaries are consistent in at least half of the assemblies in the comparison, and (ii) green otherwise. Blocks of misassembled contigs are coloured orange if the boundaries are consistent in at least half of the assemblies in the comparison, and red otherwise. Bracketed datasets represent the same DNA extract.



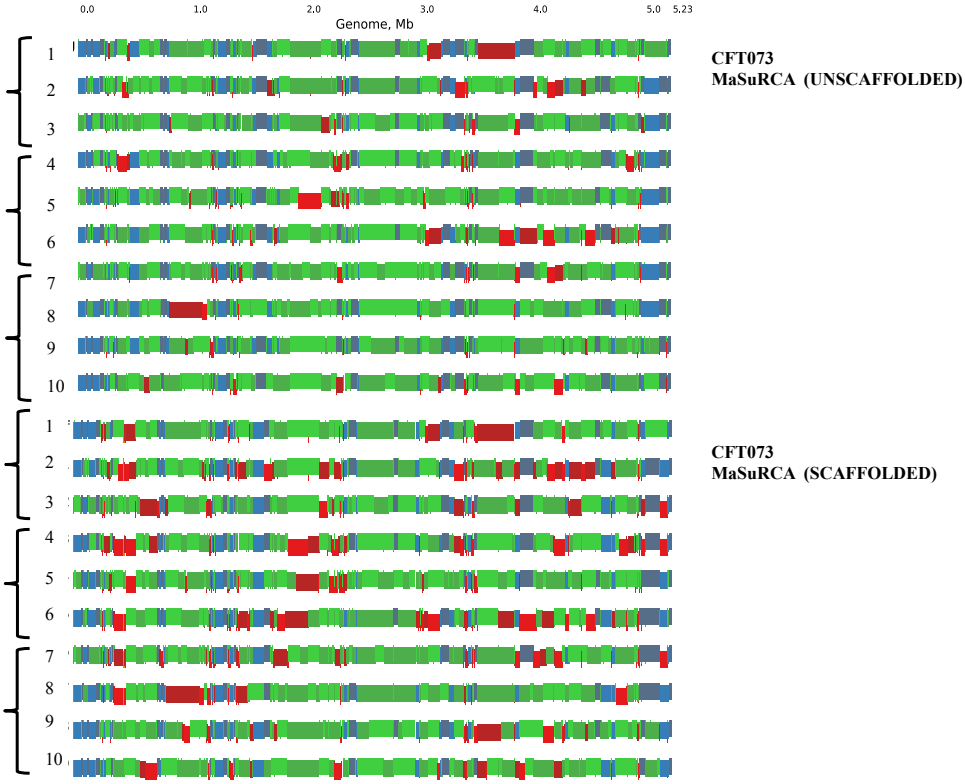
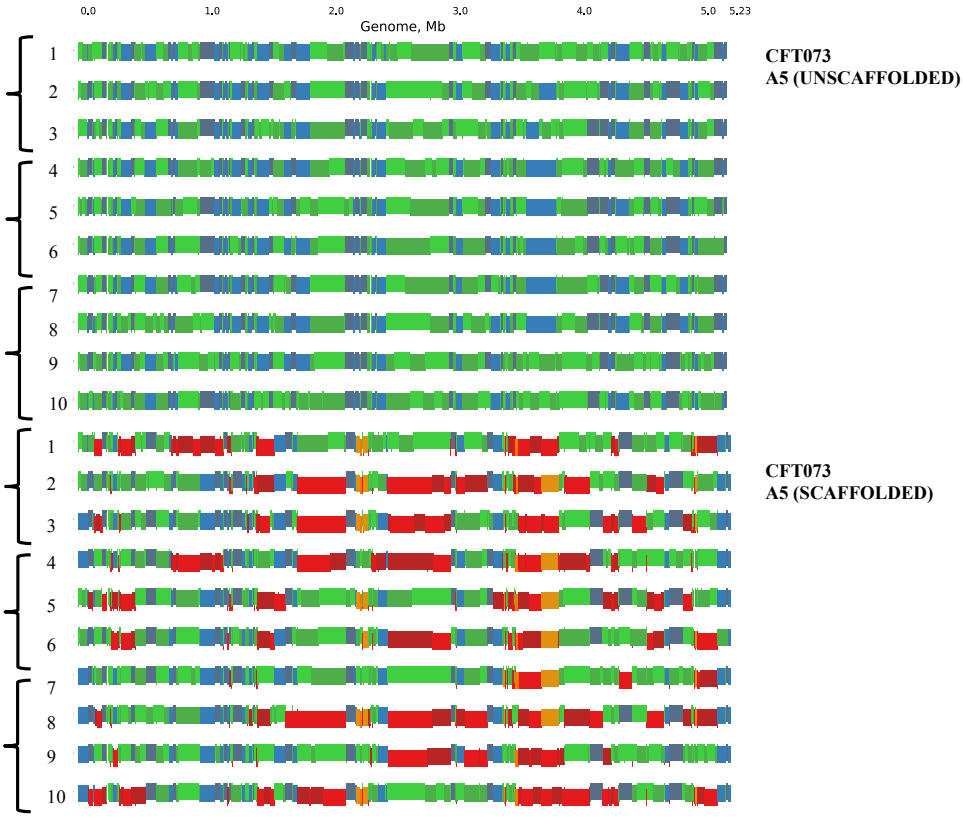


Figure 4.4. panels c and d. Comparisons of *de novo* assembled non-enteropathogenic *E. coli* CFT073 datasets using MaSuRCA (top) and A5 (bottom). Colouring as previously described.



4.4. ROBUSTNESS OF *DE NOVO* ASSEMBLIES OF TRANSFORMED PLASMIDS

In order to assess the quality of my method for sequencing plasmid transformants as a means to get accurate resistance plasmid assemblies, I undertook several control experiments. The first involved sequencing strains transformed with a number of known resistance plasmids of differing sizes, namely pUC19, a very small plasmid (2,686bp); pOX38-CL a moderately sized plasmid (57,129bp) represented by the largest *Hin*dIII digestion fraction of the F-plasmid(7), into which a chloramphenicol resistance cassette (~2210bp) had previously been inserted in our research laboratory; and pP46212, a CTX-M-15 plasmid extracted from one of the isolates analysed as part of a thesis dataset (on of the strains in the dataset described in chapter 8). pP46212 is a large (143,748bp), closed plasmid sequence I obtained from a pilot PacBio sequencing experiment. This represented a compromise set of reference plasmid strains, given the differing availability of appropriate plasmid reference strains and access to sequencing technology at various time-points during the thesis.

For these analyses, I opted to use the A5 assembler without scaffolding, as this had been shown to be the most reliable assembler when assembling *E. coli*/*K.pneumoniae* reference genomes, and these analyses were being carried out in relatively small numbers outwith the group's standard data processing pipeline. Following on from the across-reference *de novo* comparisons, the latest, upgraded version of the A5 pipeline was implemented, namely A5-MiSeq(8). This had been upgraded to cope with longer reads present in MiSeq data, used read mate-pair information during assembly, and included an improved read trimming algorithm. Once installed, the A5-MiSeq assembler was run as a default using:

./ a5_pipeline.pl input_ forward_fastq1.fq input_backward_fastq2.fq output_fasta.fa

Assembled data was annotated using PROKKA(9). Comparisons were made by aligning assembled test sequences with the respective reference sequences using progressiveMauve(10). Alignments were visualised and manually edited in Geneious(11).

For the sequenced DH10B-pUC19 transformant, reads were first mapped to the DH10B reference (RefSeq accession: NC_010473; 4,686,137bp) using the standard pipeline. Of the 2,210,686 reads available for mapping, 214,536 (9.7%; both read mates) remained unmapped and were extracted for assembly using SAMTools. *De novo* assembly with A5-MiSeq yielded a single contig of 2,543bp, which represented almost all of the reference pPUC19 strain (Figure 4.5. panels a and b). Part of the *lac* operon was missing in the *de novo* assembled plasmid – this is likely to be because this sequence is also observed in a number of bacterial chromosomes, including that of DH10B. These reads would therefore not have been in the extracted pool of unmapped reads, and would not have been available to include in the *de novo* plasmid assembly. The *de novo* assembly also had an additional short stretch of sequence not present in the reference, but the rest of the structure was identical.

pOX38-CL is a larger plasmid, albeit one mostly devoid of insertion sequences, which are known to cause problems for *de novo* assemblers as they are often repetitive, making ascertainment of the correct underlying structural linkages between sequenced reads difficult. Of 935,114 reads available for mapping from the sequenced

DH10B-pOX38-CL transformant, 76,248 (8.2%, both read mates) were unmapped to DH10B; these were assembled into seven contigs, totalling 57,733bp.
The alignments of *de novo* assembled sequence and reference (Figure 4.6. panels a and b) show good concordance, with the inserted resistance gene cassette clearly present in the *de novo* assembly, but absent from the reference, as expected. Small numbers of divergent sequence events were found, particularly in a region with relatively large numbers of short sequences annotated in the reference as repetitive regions (between positions: 33Kb and 40kb).

Plasmid pP46212 was the largest plasmid analysed, and likely to be most representative of resistance plasmids in my dataset, given that it had come from one of the datasets being studied. Here again, effectively almost all of the sequence was recovered (Figure 4.7. panel a and b) from my Illumina-processed transformants. Missing regions that had failed to assemble were annotated as transposases, or integrase core domain proteins associated with insertion sequences, features that Illumina sequencing would not be expected to recover.

Broadly speaking, the approach of transforming resistance plasmids of interest appeared a relatively robust way of assembling plasmid structures of a broad range of size and type, including the large plasmids involved in the transmission of antimicrobial resistance genes that are the main focus of interest in this study. Some loss of resolution around repetitive and mobile genetic elements such as transposons, insertion sequences and phage-related elements was observed, but this was not unexpected, and I demonstrated that the recovery of the majority of the plasmid sequence could be anticipated.

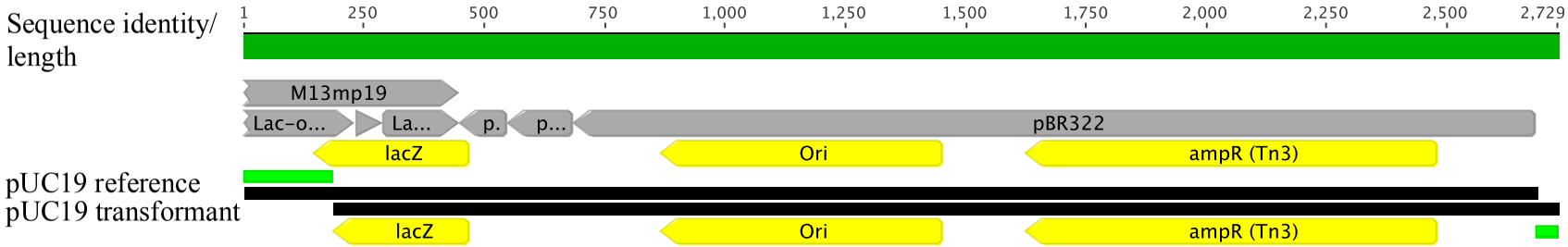


Figure 4.5. panel a. Annotated alignment of reference pUC19 sequence and assembled sequence (sequences in black; annotations in grey/ yellow) from the DH10B-pUC19 transformant. Combined sequence length and identity are represented by the top bar (dark green=100% identity). Additional portions of sequence found uniquely in either the reference (top sequence) or the transformant (bottom sequence) are highlighted in pale green.

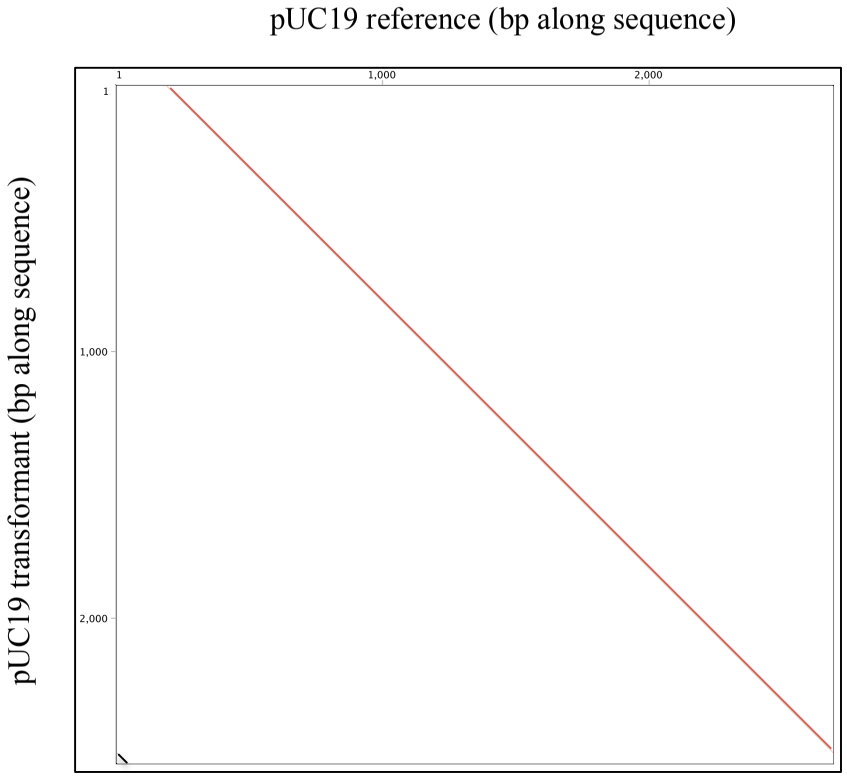


Figure 4.5. panel b. Dot-plot representation of pairwise comparison of the two aligned pUC19 sequences (reference and transformant) using a modified version of the EMBOSS dottup tool. The red line denotes sequence homology.

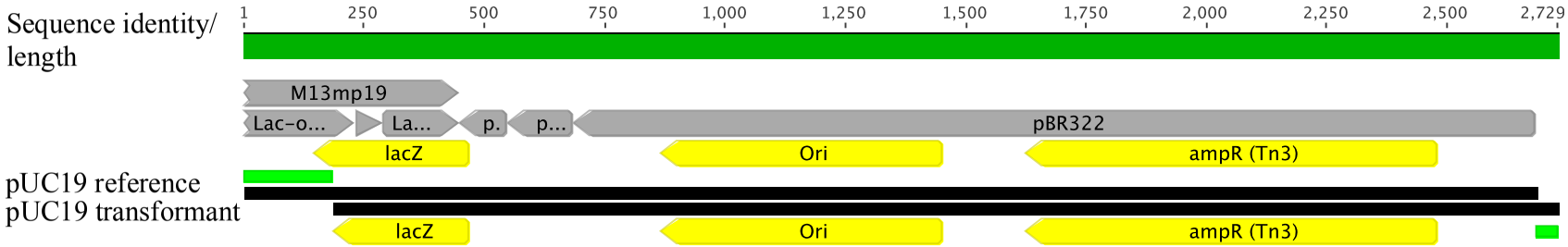


Figure 4.5. panel a. Annotated alignment of reference pUC19 sequence and assembled sequence (sequences in black; annotations in grey/ yellow) from the DH10B-pUC19 transformant. Combined sequence length and identity are represented by the top bar (dark green=100% identity). Additional portions of sequence found uniquely in either the reference (top sequence) or the transformant (bottom sequence) are highlighted in pale green.

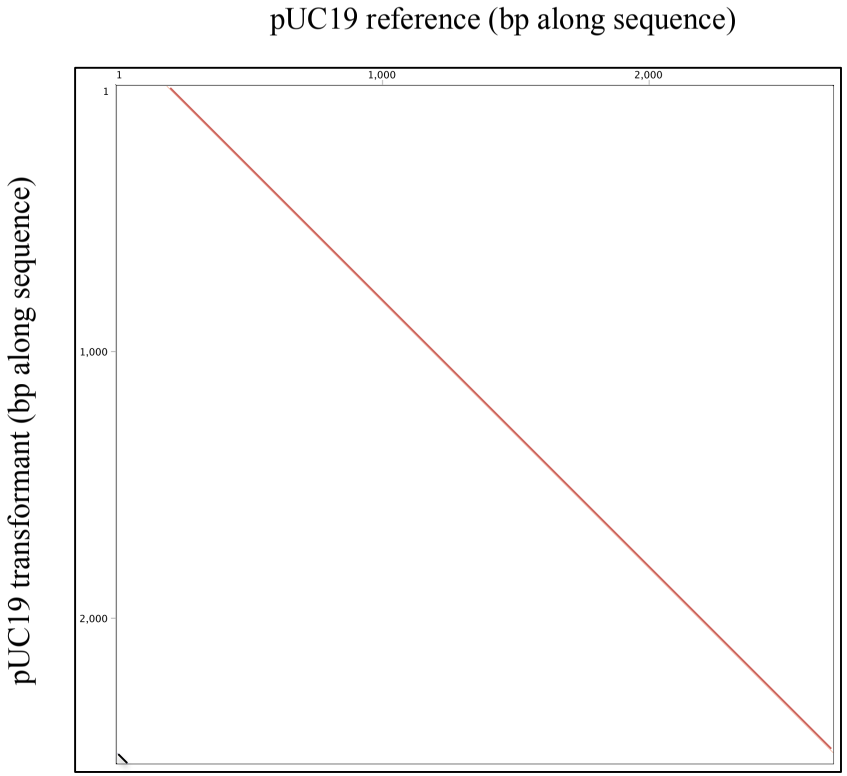


Figure 4.5. panel b. Dot-plot representation of pairwise comparison of the two aligned pUC19 sequences (reference and transformant) using a modified version of the EMBOSS dottup tool. The red line denotes sequence homology.

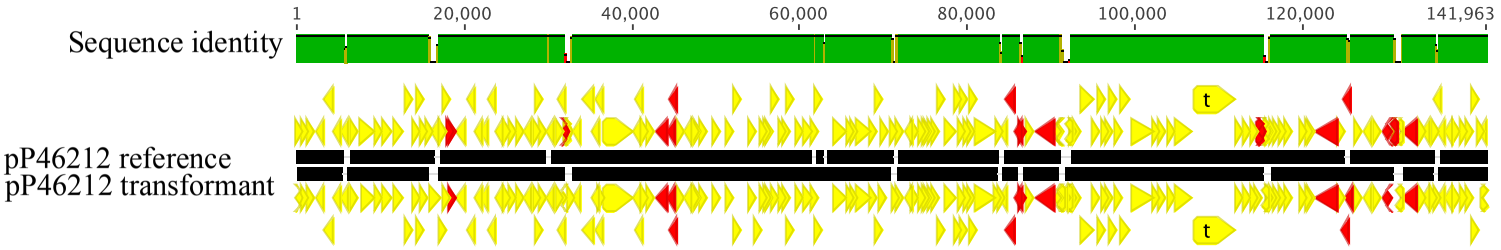


Figure 4.7. panel a. Annotated alignment of reference pP46212 PacBio sequence and assembled pP46212 sequence (sequences in black; annotations in yellow) from the DH10B-pP46212 transformant. Combined sequence length and identity are represented by the top bar (dark green=100%; gaps = 0%). Regions annotated as repetitive regions are shown as red arrows; these are associated with differences to the reference. The sequence is shorter as some short tracts are lost in the Mauve alignment output.

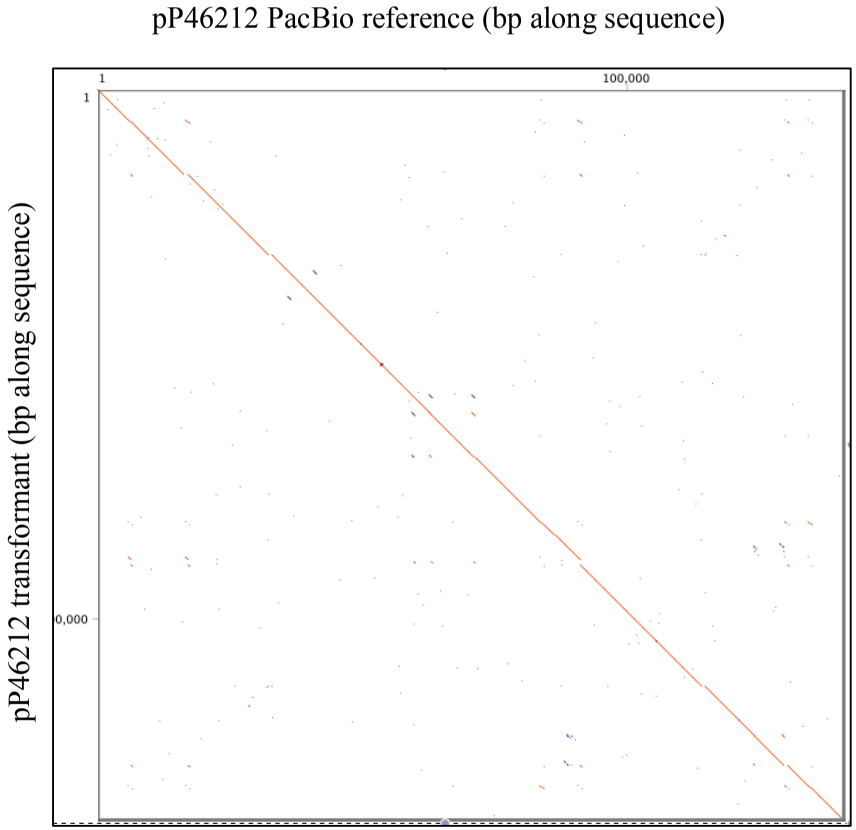


Figure 4.7. panel b. Dot-plot representation of pairwise comparison of the two aligned pP46212 sequences (reference and transformant) using a modified version of the EMBOSS dottup tool. The red line denotes sequence homology.

CHAPTER 4 REFERENCES

1. **Magoc T, Pabinger S, Canzar S, Liu X, Su Q, Puiu D, Tallon LJ, Salzberg SL.** 2013. GAGE-B: an evaluation of genome assemblers for bacterial organisms. Bioinformatics **29:**1718-1725.
2. **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.
3. **Zimin AV, Marcais G, Puiu D, Roberts M, Salzberg SL, Yorke JA.** 2013. The MaSuRCA genome assembler. Bioinformatics **29:**2669-2677.
4. **Tritt A, Eisen JA, Facciotti MT, Darling AE.** 2012. An integrated pipeline for de novo assembly of microbial genomes. PloS one **7:**e42304.
5. **Miller JR, Delcher AL, Koren S, Venter E, Walenz BP, Brownley A, Johnson J, Li K, Mobarry C, Sutton G.** 2008. Aggressive assembly of pyrosequencing reads with mates. Bioinformatics **24:**2818-2824.
6. **Gurevich A, Saveliev V, Vyahhi N, Tesler G.** 2013. QUAST: quality assessment tool for genome assemblies. Bioinformatics **29:**1072-1075.
7. **Guyer MS, Reed RR, Steitz JA, Low KB.** 1981. Identification of a sex-factor-affinity site in E. coli as gamma delta. Cold Spring Harbor symposia on quantitative biology **45 Pt 1:**135-140.
8. **Coil DJ, G.; Darling, A.** A5-miseq: an updated pipeline to assemble microbial genomes from Illumina MiSeq data. Quantitative Biology.

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

10. **Darling AE, Mau B, Perna NT.** 2010. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PloS one **5:**e11147.

11. **Biomatters.** Geneious, 7.1 ed.

CHAPTER 5: RESISTANCE GENE IDENTIFICATION IN *ESCHERICHIA COLI* AND *KLEBSIELLA PNEUMONIAE*, AND PREDICTION OF ANTIMICROBIAL SUSCEPTIBILITY PHENOTYPE FROM WHOLE GENOME SEQUENCING DATA

Text

5.1. INTRODUCTION

The lowest unit of antimicrobial resistance is the resistance gene itself, the presence and variant of which will be one of the major determinants of antimicrobial resistance phenotype. Accurate assessment of the catalogue of resistance genes within non-enteropathogenic *Escherichia coli*/*Klebsiella pneumoniae* is therefore a prerequisite to being able to investigate the molecular epidemiology of these organisms; historically this has been done at a limited level using a series of polymerase chain reaction (PCR) assays, either sequentially, or as a multiplex, with sequencing of PCR products to determine the specific variant present.

This chapter describes work done to generate a catalogue of resistance gene variants for beta-lactam resistance in non-enteropathogenic *E. coli*/*K. pneumoniae*, as well as for aminoglycoside and quinolone resistance, and to use this to assess the capacity for resistance determinants to be identified from Illumina sequence data. This was done in the context of a proof-of-principle experiment designed to compare the accuracy of susceptibility prediction by whole genome sequencing (WGS) with that of currently routinely used phenotyping methods. Conceivably, this could be part of the translational application of benchtop sequencers, such as the MiSeq (Illumina, San Diego, CA), to diagnostic microbiology, facilitating a one-stop approach to species

identification and susceptibility profiling of cultured bacterial isolates with turnaround times of less than a day(1).

At present, routine antimicrobial susceptibility testing is undertaken using a variety of approaches, including disc diffusion, gradient diffusion, and broth dilution methods (as described in the thesis "Methods" section), the latter being automated as part of commercial platforms such as the BD Phoenix (BD, Franklin Lakes, NJ) or the Vitek-2 (bioMérieux, Marcy L'Etoile, France)(2). Despite extensive efforts to standardise laboratory assays, problems with particular test methods for certain organism/antimicrobial combinations are well recognised and may relate to inherent properties of the organism or antimicrobial being tested. This has been noted as a particular problem with piperacillin-tazobactam(3, 4). Other errors can arise in inoculum preparation, culture conditions, or data entry.

Susceptibility phenotyping errors are typically classified as very major, resulting from a false-susceptible result, or major, resulting from a false-resistant result(2). The US Food and Drug Administration (FDA) stipulates rates must be <1.5% for very major errors and <3% for major errors prior to authorising marketing approval for new susceptibility testing devices; similar cut-offs have been proposed by others(2). In controlled research studies, overall error rates are 0-8%(5); in routine settings the error rates are assessed with quality control schemes, and vary between laboratories.

Routine genotypic prediction of bacterial antimicrobial susceptibility is currently used only in limited contexts, typically with single gene targets known to be highly associated with resistance, such as *mecA* assays to determine methicillin resistance in

Staphylococcus aureus. The prevailing view has been that genotypic assays would be too difficult to implement for complex patterns of antimicrobial resistance, for example those in major Gram-negative pathogens such as non-enteropathogenic *E. coli*/ *K. pneumoniae* (2). However, recent data investigating WGS approaches to identifying susceptibility phenotypes of porcine *Salmonella* Typhimurium, *E. coli*, *Enterococcus faecium*, and *Enterococcus faecalis* isolates for resistance surveillance purposes showed high concordance between phenotypic and predicted antimicrobial susceptibilities(6). Caveats to this acknowledged by the authors included the low complexity of the resistance genotypes in the bacterial populations studied (i.e., small numbers of resistance genes per isolate conferring resistance to the same antimicrobial class), and that no assessment of some important chromosomal markers of resistance, such as *gyrA* mutations for fluoroquinolones, was made.

Non-enteropathogenic *E. coli* and *K. pneumoniae* are the Gram-negative species most commonly identified in bacteraemic patients in the UK(7), with increases in incidence noted across Europe(8). As such, these organisms, in which multi-drug resistance is increasingly recognised(9, 10), represent species for which accurate and rapid antimicrobial susceptibility testing has the potential to deliver direct clinical benefit. Consequently, in this chapter I aimed to: (i) assess the potential of WGS to identify resistance genes of relevance to the study of the molecular epidemiology of antimicrobial resistance in non-enteropathogenic *E. coli* /*K. pneumoniae*; and (ii) to assess the feasibility of using WGS data from human blood culture isolates of non-enteropathogenic *E. coli* and *K. pneumoniae* representative of those seen in clinical practice to predict susceptibility phenotypes for antibiotics commonly used to manage infections caused by these organisms.

5.2. MATERIALS AND METHODS

5.2.1. Clinical isolate selection and *in vitro* antimicrobial susceptibility testing

I selected all retrievable extended-spectrum cephalosporin-resistant (commonly representative of multi-drug resistant phenotypes)(11) non-enteropathogenic *E. coli* and *K. pneumoniae* blood culture isolates obtained from patients at the Oxford University Hospitals NHS Trust, Oxford, UK, between January 2008 and November 2010 (non-enteropathogenic *E. coli*) or June 2011 (*K. pneumoniae*). Time-matched (by calendar year) susceptible control blood culture isolates were also selected at random and retrieved (Figure 5.1.).

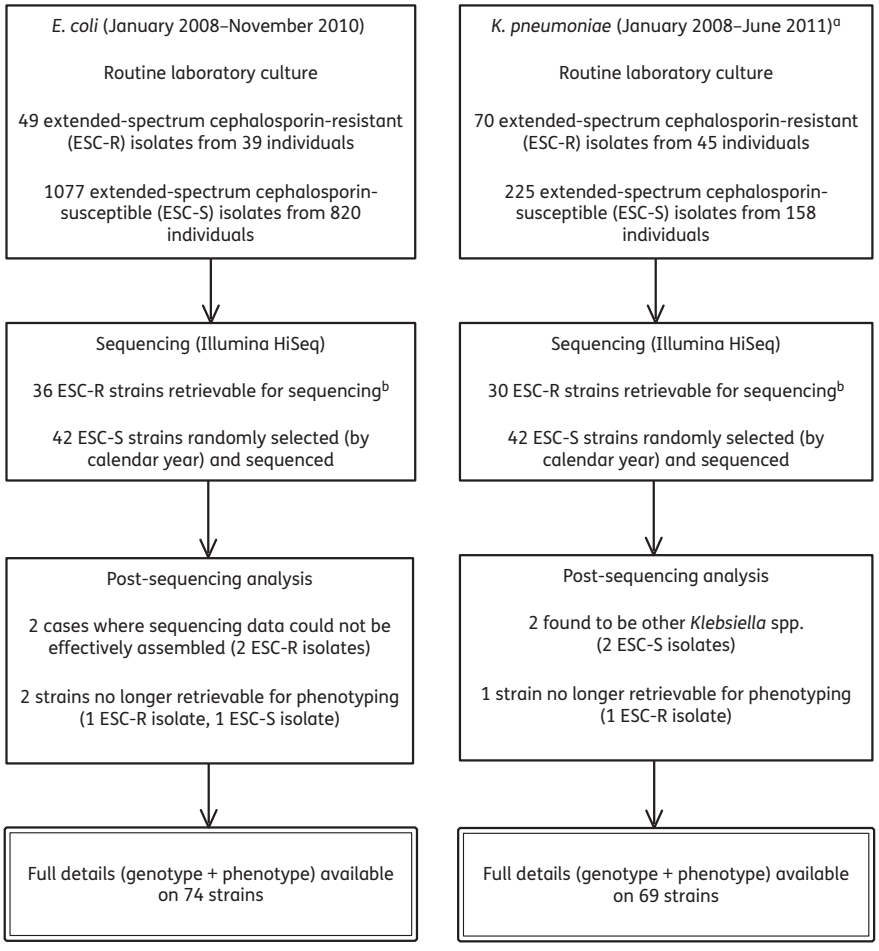


Figure 5.1. Sampling frame and processing of isolates. [a] The study time period for *K. pneumoniae* was extended to find similar numbers of organisms across both species groups. [b] Losses at the retrieval stage were mostly due to the fact that repeat isolates from individuals were not routinely stored; other missing isolates could not be found in the routine laboratory freezer.

Isolates were re-cultured from frozen stocks (-80°C) and underwent automated susceptibility testing in duplicate with the BD Phoenix system using EUCAST breakpoints(12) to allow comparisons with genotypic data. Intermediate BD Phoenix susceptibilities were considered as resistant (Tables 5.2.a. and 5.2.b.). In cases where duplicate BD Phoenix runs were concordant regarding an isolate's resistance category (susceptible/susceptible [S/S], resistant/resistant [R/R]), the consensus BD Phoenix phenotype was compared with the genotype. All discrepancies, whether discordance between BD Phoenix runs (S/R), or between predicted genotypic susceptibility and concordant BD Phoenix phenotype (S + R/R, or R + S/S respectively), were further investigated using gradient diffusion testing (Etest, bioMérieux, Basingstoke, UK; M.I.C Evaluator, Fisher Scientific UK, Loughborough, UK) on Iso-Sensitest agar in accordance with BSAC guidelines(13). In such cases, regardless of the nature of the discrepancy, the gradient diffusion result was adopted as the comparison standard phenotype; in all other cases the concordant BD Phoenix phenotype was the comparison standard.

5.2.2. Reference gene database

Genetic loci and sequence variants known to be associated with resistance to antimicrobial agents commonly used in our hospital to treat non-enteropathogenic *E. coli* and *K. pneumoniae* infections were identified from published reviews and web-based resources and were compiled as a reference gene database(14-22). Chromosomal and plasmid-mediated loci conferring resistance to amoxicillin, amoxicillin-clavulanate, ciprofloxacin, gentamicin, ceftriaxone, ceftazidime, and meropenem were included (full details of all mechanisms included are listed in Appendix 1). An additional search of complete coding sequences annotated as being

Table 5.2.a. Analysis of discordance in phenotypic and/or genotypic resistance predictions for 74 non-enteropathogenic *E. coli* bloodstream isolates.

Antibiotic	Discrepancies (n; % of total, 74 isolates)			Agreement of gradient diffusion with genotype in all discrepancies (n/total discrepancies; %)	Agreement of gradient diffusion with genotype in BD Phoenix-concordant discrepancies (n; %)
	S/R[a] discordant BD Phoenix	BD Phoenix S/S, genotype R	BD Phoenix R/R, genotype S		
Amoxicillin	0 (0)	2 (3)	0 (0)	1/2 (50)	1/2 (50)
Co-amoxiclav	5 (7)	0 (0)	15 (20)	20/20 (100)	15/15 (100)
Gentamicin	1 (1)	0 (0)	0 (0)	1/1 (100)	NA[b]
Ciprofloxacin	0 (0)	0 (0)	0 (0)	NA	NA
Ceftriaxone	0 (0)	1 (1)	1 (1)	0/2 (0)	0/2 (0)
Ceftazidime	1 (1)	11 (15)	1 (1)	1/13 (8)	1/12 (8)
Meropenem	0 (0)	0 (0)	0 (0)	NA	NA
Total	7/518 (1)	14/518 (3)	17/518 (3)	23/38 (61)	17/31 (55)

[a] S/R denotes susceptible/resistant category. Initial BD Phoenix intermediate results were counted as resistant – this occurred in one isolate with a S/R discrepancy for ceftazidime, and one isolate with a S/R discrepancy for gentamicin.
[b] NA = not applicable.

Table 5.2.b. Analysis of discordance in phenotypic and/or genotypic resistance predictions for 69 *K. pneumoniae* bloodstream isolates.

Antibiotic	Discrepancies (n; % of total, 69 isolates)			Agreement of gradient diffusion with genotype in all discrepancies (n/total discrepancies; %)	Agreement of gradient diffusion with genotype in BD Phoenix-concordant discrepancies (n; %)
	S/R[a] discordant BD Phoenix	Phoenix S/S, genotype R	Phoenix R/R, genotype S		
Amoxicillin	3 (4)	3[b] (4)	0 (0)	3/6 (50)	1/3 (33)
Co-amoxiclav	2 (3)	0 (0)	6 (9)	7/8 (88)	6/6 (100)
Gentamicin	1 (1)	0 (0)	1 (1)	1/2 (50)	0/1 (0)
Ciprofloxacin	1 (0)	2 (3)	7 (10)	4/10 (40)	3/9 (33)
Ceftriaxone	0 (0)	1 (1)	2 (3)	0/3 (0)	0/3 (0)
Ceftazidime	0 (0)	1 (1)	2 (3)	0/3 (0)	0/3 (0)
Meropenem	1 (1)	0 (0)	0 (0)	0/1 (0)	NA
Total	8/483 (2)	7/483 (1)	18/483 (4)	15/33 (45)	10/25 (40)

[a] SR denotes susceptible/resistant category. Initial BD Phoenix intermediate results were counted as resistant – this occurred in one isolate with a S/R discrepancy for ciprofloxacin, and one isolate with a S/R discrepancy for meropenem.
[b] This applies to a minimum inhibitory concentration (MIC)-based assessment of BD Phoenix results, disregarding interpretative guidelines (which would suggest that *K. pneumoniae* be universally reported as amoxicillin-resistant for clinical purposes, irrespective of MIC).
[c] NA = not applicable.

members of relevant bacterial (other than mycobacterial) resistance gene families deposited at the National Centre for Biotechnology Information (NCBI) was performed, using the following search terms: (a) “lactamase”, (b) “carbapenemase”, (c) “aminoglycoside” + “resistance”, and (d) “fluoroquinolone” + “resistance” (December 2012; references in Appendix 1).

5.2.3. DNA extraction and sequencing

DNA extraction, whole genome sequencing, and pipeline read-processing were carried out as per the thesis “Methods” section. *De novo* assembly quality was ensured by requiring >4 megabases (Mb) to be assembled into contigs, and contig N50 values of >30,000 bases.

5.2.4. *In silico* prediction of antimicrobial susceptibility phenotypes

BLASTn(23) was used to identify the presence of relevant resistance gene loci (from the compiled reference resistance gene database) in the *de novo* assembled contigs for each clinical isolate, with a word length of 11 and an expect value (E) cut-off of $1x10^{-4}$. All matches were visually inspected for confirmation. Matches with more than 80% identity at the nucleotide level and over 80% of the reference gene length were retained; this included partial matches with >80% sequence homology over 80% of the reference gene length, but distributed over several contigs. Overlapping fragments were then aligned in SeaView(24) and combined to give a single sequence. Chromosomal resistance gene sequences were analysed to identify mutations, including those known to be associated with resistance.

Each isolate's susceptibility phenotype was predicted from the genetic data on the basis of published associations with phenotypic resistance for each locus, and blinded to the BD Phoenix phenotype (details for susceptibility predictions for all profiles found shown in Tables I-III (non-enteropathogenic *E.* coli) and IV-VI (*K. pneumoniae*) at the end of this chapter. For any novel sequence variants identified, the genotypic susceptibility prediction mirrored that of the closest reference database variant. Discrepancies between BD phenotype and genotype were then investigated using gradient diffusion, as described above.

The sensitivity, specificity, and rates of major and very major errors for genotypic susceptibility predictions were calculated for each antibiotic and species against the comparison standard (determined as above). In the published manuscript, I defined major error as: the number of genotype R with phenotype S calls/total number of tests x100, and very major error as: the number of genotype S with phenotype R calls/total number of tests x 100. The FDA definitions by contrast use the total number of organisms susceptible by the reference method as the denominator for major errors (i.e. simply 1-specificity), and the total number of resistant organisms as the denominator for very major errors (i.e. simply 1-sensitivity). Statistical analyses were performed using Stata 11.2 (StataCorp, College Station, TX).

5.3. RESULTS

5.3.1. Quality of whole genome sequences

Two of the 76 candidate non-enteropathogenic *E. coli* study isolates were excluded because of poor sequence assembly (N50 <1,250 and < 0.3Mb assembled into contigs); two *K. pneumoniae* isolates were excluded because they were non-

pneumoniae Klebsiella spp. on the basis of mapping (see summary as part of Figure 5.1. above). Assemblies for the 74 remaining non-enteropathogenic *E. coli* isolates had a median of 394 contigs (range: 93-1,052) and N50 of 110,187 base-pairs (bp) (range: 32,391-189,171 bp). For the 69 *K. pneumoniae* study isolates, the corresponding medians were 255 contigs (range: 171-863) and n50 of 97,195 bp (range: 58,500-135,350 bp).

5.3.2. Investigation of phenotype-genotype discrepancies

Susceptibility phenotypes for seven antimicrobials were available for 143 study isolates (74 non-enteropathogenic *E. coli*, 69 *K. pneumoniae*), giving 1001 total susceptibility results (518 non-enteropathogenic *E. coli*, 483 *K. pneumoniae*), for comparison with the corresponding genotypic predictions. Gradient diffusion analysis was used to establish the phenotype for 71 antimicrobial-isolate combinations (involving 55 different isolates), which comprised seven (1%) non-enteropathogenic *E. coli* results and eight (2%) *K. pneumoniae* results with categorical (S/R) discordance in duplicate BD Phoenix testing, and 31 (6%) non-enteropathogenic *E. coli* and 25 (5%) *K. pneumoniae* results with discordance between the predicted genotypic susceptibility and the (concordant) BD Phoenix phenotype (Tables 5.2.a. [non-enteropathogenic *E. coli*] and 5.2.b. [*K. pneumoniae*]).

5.3.3. Genotypic prediction versus gold standard "reference" phenotype

Overall, the sensitivity of genotype for predicting resistance across all antibiotics for both species was 0.97 (95% confidence interval [95% CI]: 0.94-0.98), and specificity was 0.97 (95% CI: 0.95-0.98). Very major and major error rates, were, by the original definition used in the paper, at 1.2% and 2.1% respectively; for the FDA definitions at

3.4% and 3.2% respectively. For non-enteropathogenic *E. coli,* overall sensitivity was 0.99 (95% CI: 0.95-1.0) and specificity 0.96 (95% CI: 0.94-0.98) (Table 5.3.a), the major individual drug deficit being suboptimal specificity for ceftazidime (0.80; 95% CI: 66-89). Very major and major error rates were 0.3% and 3% by the original definitions in the paper; and 1.2% and 3.7% by the FDA definitions. For *K. pneumoniae*, overall sensitivity was 0.95 (95% CI: 0.90-0.97) and specificity 0.97 (95% CI: 0.95-0.99), with very major and major error rates of 2% by the original definitions in the paper, and 5.4% and 2.7% by the FDA definitions (Table 5.3.b.).

In non-enteropathogenic *E. coli,* in 23 (61%) of the 38 isolate-antimicrobial combinations with a phenotype-genotype discrepancy according to BD-Phoenix results, gradient diffusion analysis supported the genotypic prediction (Table 5.2.a. above). For the remaining 15 confirmed genotypic-phenotype discrepancies, results are summarised in Table 5.4. below. In 13 (87%) of these cases, a clear-cut genetic resistance mechanism was identified despite phenotypic susceptibility; for nine (69%) of these the gradient diffusion minimum inhibitory concentration (MIC) was at the susceptibility breakpoint. The remaining 2 (13%) of the 15 discrepant genotype-phenotype cases had no identifiable genetic resistance mechanism despite unequivocal phenotypic resistance.

In *K. pneumoniae,* in 15 (45%) of the 33 isolate-antimicrobial combinations with a phenotype-genotype discrepancy according to BD-Phoenix results, gradient diffusion analysis supported the genotypic prediction (Table 5.2.b. above). For the remaining 18 confirmed genotype-phenotype discrepancies, results are summarised in Table 5.4. below.

Table 5.3.a. Sensitivity and specificity of genotypic resistance predictions versus comparison with standard phenotype results for 74 non-enteropathogenic *E. coli* bloodstream isolates

	Susceptible by comparison standard phenotype		Resistant by comparison standard phenotype			
Antibiotic	susceptible by genotype (row %)	resistant by genotype (row %; major error)	susceptible by genotype (row %; very major error)	resistant by genotype (row %)	Sensitivity (95% CI)	Specificity (95% CI)
Amoxicillin	23 (31)	1 (1)	0 (0)	50 (68)	1.00 (0.91–1.00)	0.96 (0.77–1.00)
Co-amoxiclav	46 (62)	0 (0)	0 (0)	28 (38)	1.00 (0.85–1.00)	1.00 (0.90–1.00)
Gentamicin	60 (81)	0 (0)	0 (0)	14 (19)	1.00 (0.73–1.00)	1.00 (0.93–1.00)
Ciprofloxacin	48 (65)	0 (0)	0 (0)	26 (35)	1.00 (0.84–1.00)	1.00 (0.91–1.00)
Ceftriaxone	43 (58)	1 (1)	1 (1)	29 (40)	0.97 (0.81–1.00)	0.98 (0.87–1.00)
Ceftazidime	43 (58)	11 (15)	1 (1)	19 (26)	0.95 (0.73–1.00)	0.80 (0.66–0.89)
Meropenem	74 (100)	0 (0)	0 (0)	0 (0)	—	1.00 (0.94–1.00)
Total	337 (65)	13 (3)	2 (0.3)	166 (32)	0.99 (0.95–1.00)	0.96 (0.94–0.98)

Table 5.3.b. Sensitivity and specificity of genotypic resistance predictions versus comparison with standard phenotype results for 69 *K. pneumoniae* bloodstream isolates

	Susceptible by comparison standard phenotype		Resistant by comparison standard phenotype			
Antibiotic	susceptible by genotype (row %)	resistant by genotype (row %; major error)	susceptible by genotype (row %; very major error)	resistant by genotype (row %)	Sensitivity (95% CI)	Specificity (95% CI)
Amoxicillin	0 (0)	3 (4)	0 (0)	66 (96)	1.00 (0.93–1.00)	—
Co-amoxiclav	47 (68)	1 (1)	0 (0)	21 (30)	1.00 (0.81–1.00)	0.98 (0.88–1.00)
Gentamicin	45 (65)	0 (0)	1 (1)	23 (33)	0.96 (0.77–0.98)	1.00 (0.90–1.00)
Ciprofloxacin	45 (65)	2 (3)	4 (6)	18 (26)	0.90 (0.67–0.98)	0.92 (0.80–0.97)
Ceftriaxone	42 (61)	1 (1)	2 (3)	24 (35)	0.92 (0.73–0.99)	0.98 (0.86–1.00)
Ceftazidime	42 (61)	1 (1)	2 (3)	24 (35)	0.92 (0.73–0.99)	0.98 (0.86–1.00)
Meropenem	68 (99)	0 (0)	1 (1)	0 (0)	0 (0–0.95)	1.00 (0.93–1.00)
Total	289 (60)	8 (2)	10 (2)	176 (36)	0.95 (0.90–0.97)	0.97 (0.95–0.99)

In 6/18 (33%) instances, a recognised resistance mechanism was identified in phenotypically susceptible isolates; for two of these the MIC was at the susceptibility breakpoint. This group included two isolates predicted to be ciprofloxacin-resistant based on chromosomal mutations (double *gyrA* amino acid replacements) that were phenotypically ciprofloxacin-susceptible. In contrast, four isolates predicted to be ciprofloxacin-susceptible (based on a single mutation in *parC* [3 isolates with S80I], or combined mutations in *parC* [E84K] + *parE* [S458T]) were phenotypically resistant, suggesting the presence of unidentified resistance mechanisms. Similarly, six other isolates with unequivocal phenotypic resistance to one or more agents from several antibiotic classes (ceftriaxone, ceftazidime, meropenem, gentamicin; eight total agent-isolate combinations) had no identifiable resistance mechanism.

Table 5.4. List of relevant genotypic profiles for 13 non-enteropathogenic *E. coli* and 15 *K. pneumoniae* isolates with genotype-gradient diffusion susceptibility discrepancies for one or more antimicrobials.

Species	Number of isolates	Antibiotic discrepancy	Genotypic prediction	Genotypic mechanism for resistance prediction	Phenotypic result	MIC on gradient diffusion, μg/mL (EUCAST susceptibility breakpoint, μg/mL)
E. coli	1	Amoxicillin	R	P3 TEM-promoter and $bla_{\text{TEM-1}}$	S	6 (8)
E. coli	1	Ceftriaxone[a]	S	None	R	>32 (1)
		Ceftazidime[a]	S		R	4 (1)
E. coli	1	Ceftriaxone[b]	R	T-32A *ampC* promoter mutation	S	0.38 (1)
		Ceftazidime[b]	R		S	1 (1)
E. coli	1	Ceftazidime	R	$bla_{\text{CTX-M-15}}$	S	0.25 (1)
E. coli	7	Ceftazidime	R	$bla_{\text{CTX-M-15}}$	S	1 (1)
E. coli	1	Ceftazidime	R	$bla_{\text{CTX-M-14}}$	S	0.5 (1)
E. coli	1	Ceftazidime	R	$bla_{\text{CTX-M-1}}$	S	1 (1)
K. pneumoniae	2	Amoxicillin	R	bla_{LEN}	S	4, 8 (8)
K. pneumoniae	1	Amoxicillin	R	bla_{SHV}	S	6 (8)
K. pneumoniae	1	Co-amoxiclav	R	$bla_{\text{OXA-1}}$	S	8 (8)
K. pneumoniae	1	Ceftriaxone[c]	R	$bla_{\text{SHV-27}}$	S	0.064 (1)
		Ceftazidime[c]	R		S	0.25 (1)
K. pneumoniae	1	Ceftriaxone[d]	S	None	R	8 (1)
		Ceftazidime[d]	S		R	64 (1)
K. pneumoniae	1	Ceftriaxone[e]	S	None	R	>32 (1)
		Ceftazidime[e]	S		R	8 (1)
K. pneumoniae	1	Meropenem	S	None	R	>32 (2)
K. pneumoniae	1	Ciprofloxacin	R	2 *gyrA* mutations (S83F + D87A)	S	0.064 (0.5)
K. pneumoniae	1	Ciprofloxacin	R	2 *gyrA* mutations (S83I + D87N)	S	0.047 (0.5)
K. pneumoniae	2	Ciprofloxacin	S	1 *parC* mutation (S80I)	R	8, >32 (0.5)
K. pneumoniae	1	Ciprofloxacin	S	1 *parC* mutation (S80I) + *aac(6')-Ib-cr*	R	2 (0.5)
K. pneumoniae	1	Ciprofloxacin	S	1 *parC* mutation	R	>32 (0.5)

				(E84K) + 1 *parE* mutation (S458T)		
K. pneumoniae	1	Gentamicin	S	None	R	16 (2)

[a-e] Multiple genotype-phenotype discrepancies observed for several antibiotics for the same isolate

5.3.4.1. Resistance gene profiles – non-enteropathogenic *E. coli*

Genotypic resistance profiles in non-enteropathogenic *E. coli* are summarised in Tables I (beta-lactam resistance), II (fluoroquinolone resistance), and III (aminoglycoside resistance) at the end of this chapter. There were 15 distinct profiles for beta-lactam resistance mechanisms, 22 for ciprofloxacin associated resistance mechanisms, and 12 for aminoglycoside-associated resistance mechanisms.

5.3.4.1.1. Beta-lactam resistance

Twelve (16%) isolates had bla_{TEM}, $bla_{\text{OXA-1}}$, and $bla_{\text{CTX-M}}$ conferring beta-lactam resistance; 15 (20%) had two of these three mechanisms, 24 (32%) one, and 23 (31%) none. Most bla_{TEM}-containing isolates had $bla_{\text{TEM-1}}$ (35/36), with five distinct nucleotide sequences (including the reference sequence) observed. In addition to the P3 and Pa/Pb $bla_{\text{TEM-1}}$ promoters,[19] two novel promoter sequences were identified (single nucleotide polymorphisms compared with promoter P3, C➔T at position 75 [Sutcliffe numbering](25); G➔A at position 175). However, co-amoxiclav resistance was identified only in the presence of other explanatory mechanisms with these novel promoter sequences. All bla_{OXA} variants were $bla_{\text{OXA-1}}$ and most $bla_{\text{CTX-M}}$ variants were $bla_{\text{CTX-M-15}}$ (25/29).

Only one isolate had a chromosomal *ampC* promoter mutation previously associated with significant resistance (T-32A)(18). This isolate was resistant only to amoxicillin and co-amoxiclav, with no other mechanism identified to explain this, and was phenotypically susceptible to ceftriaxone and ceftazidime (on duplicate BD Phoenix testing and gradient diffusion analysis).

5.3.4.1.2. Quinolone resistance

Ciprofloxacin resistance was invariably associated with S83L/D87N mutations in *gyrA*; almost all (23/26; 88%) ciprofloxacin-resistant isolates also had S80I/E84V mutations in *parC*. The presence of *aac-6'-Ib-cr* was also common in ciprofloxacin-resistant isolates, although not universal (23/26; 88%); *aac-6'-Ib-cr* was also identified in one ciprofloxacin-susceptible isolate without any resistance-conferring chromosomal mutations. A single isolate had a *gyrB* quinolone-resistance determining region (QRDR) mutation (S463A) with a *parE* truncation; this isolate was phenotypically susceptible. No *qnr* variants or *qepA* or *oqxAB* loci were found.

5.3.4.1.3. Aminoglycoside resistance

Four (5%) isolates had four or five different aminoglycoside resistance-conferring elements; 15 (20%) had three, 12 (16%) two, 11 (15%) one, and 32 (43%) none. All gentamicin resistance was associated with the presence of *aac(3')-II*-like enzymes, mostly *aac(3')-IIe* variants (13/14), with one isolate containing *aac(3')-IId*. Other aminoglycoside resistance loci included: *aac(6')-Ib-cr* (24 isolates), *aadA1a* (3), *aadA4* (17), *aadA5* (17), *aph(6')-Id* (16), *aph(6')-Id*-like loci (> 80% but < 95% sequence homology; 3 isolates), and *aph(3')-Ia* (4).

5.3.4.2. Resistance gene profiles – *K. pneumoniae*

Genotypic profiles associated with resistance in *K. pneumoniae* are summarised in Tables IV (beta-lactam resistance, 24 profiles), V (fluoroquinolone resistance, 20 profiles), and VI (aminoglycoside resistance, 17 profiles) at the end of this chapter.

5.3.4.2.1. Beta-lactam resistance

Twenty-one (30%) isolates had three or four beta-lactam resistance-conferring elements; 8 (12%) had two, 29 (42%) one, and 11 (16%) none. All *bla*$_{TEM}$ were *bla*$_{TEM-1}$, with P3 (n = 24) or Pa/Pb (n = 3) promoters, all *bla*$_{CTX-M}$ were *bla*$_{CTX-M-15}$, and all *bla*$_{OXA}$ were *bla*$_{OXA-1}$ (only observed with *bla*$_{CTX-M-15}$).

Most *K. pneumoniae* isolates (61/69; 88%) contained *bla*$_{SHV}$ genes encoding beta-lactamases, with six containing *bla*$_{LEN}$ (two *bla*$_{LEN-7}$, four novel variants;, one *bla*$_{OKP-B-6}$, and one *bla*$_{LAP-2}$ in conjunction with *bla*$_{SHV-11}$. The commonest *bla*$_{SHV}$ beta-lactamase variant was *bla*$_{SHV-1}$ (n = 28), with additional variants in order of frequency as follows: *bla*$_{SHV-11}$ (19), *bla*$_{SHV-28}$ (4), *bla*$_{SHV-33}$ (2), *bla*$_{SHV-121}$ (2), *bla*$_{SHV-27}$ (1), *bla*$_{SHV-60}$ (1), and *bla*$_{SHV-135}$ (1). Three novel amino acid *bla*$_{SHV}$ variants were identified (Y7F + S14F, Y7F + M211L, and D101H; assigned allele numbers 169, 170, and 171, respectively, in the Lahey database)(22). One of the 69 *K. pneumoniae* isolates contained none of these resistance loci; its beta-lactam resistance was explained by the presence of *bla*$_{TEM-1}$, *bla*$_{OXA-1}$, and *bla*$_{CTX-M-15}$.

5.3.4.2.2. Quinolone resistance

Isolates with wild-type amino acids or only single amino acid mutations in the QRDRs of *gyrA*, *gyrB*, *parC*, and *parE*, and no more than one plasmid-mediated resistance mechanism (*aac-6'-Ib-cr*, *qnr* or *qepA*), were all ciprofloxacin-susceptible (41 wild-type isolates, four single amino acid mutations). In contrast, isolates with both *aac-6'-Ib-cr* and *qnrB1* (n = 15) were invariably resistant, irrespective of underlying chromosomal mutations. Likewise, isolates with single *gyrA* and *parC* mutations and a plasmid-mediated resistance mechanism (n = 2: S83I + S80I + *aac-6'-Ib-c;* S83T + S80I + *qnrS1*), or a double mutation in *gyrA*, a single mutation in

parC, and a plasmid-mediated resistance mechanism (n = 1: S83F + D87N + S80I + *aac-6'-Ib-cr*), were also resistant.

There were no observed mutations compared to wild type in the quinolone resistance-determining region (QRDR) of *gyrB*. All isolates contained *oqxAB*, which is commonly found located chromosomally in *K. pneumoniae*, although its association with ciprofloxacin resistance in this context is unclear(26).

5.3.4.2.3 Aminoglycoside resistance

One (1%) isolate had five different resistance-conferring elements; 16 (23%) had three, nine (13%) two, seven (10%) one, and 36 (52%) none. As in non-enteropathogenic *E. coli*, gentamicin resistance in *K. pneumoniae* was typically associated with the presence of *aac(3')-II*-like enzymes, mostly *aac(3')-II-e* (19/23). Three isolates had an *aac(3')-IId* enzyme, and one an *aac(3')-Ia* variant. Other aminoglycoside resistance loci included: *aph(6')-Id* (25 isolates), *aph(3')-Ia* (4), *aadA2* (4), *aadA1* (2) and *aadA16* (1).

5.4. DISCUSSION

In this study, I determined the resistance gene profile, and sensitivity and specificity of a genotypic prediction algorithm using clinical non-enteropathogenic *E. coli* and *K. pneumoniae* isolates and WGS. In our centre, the epidemiology of these organisms has been found to be similar to the wider national and European contexts(8, 9). Using publicly available resources, I determined the presence/absence of published variants (including genes and resistance-determining mutations) in over 100 resistance-associated gene families, with particular reference to those relevant to commonly used

antimicrobials. Relative to a comparison standard phenotype based on BD Phoenix plus gradient diffusion testing, genotype-based resistance prediction yielded overall sensitivity and specificity values of 0.96 and 0.97, respectively, with a very major error rate of 1.2% and major error of 2.1%. This assessment of error rate gives a population-based representation of the overall performance of the test compared with the FDA error rate calculations, which is essentially a representation of sensitivity and specificity.

Applying genetic 'resistotyping' to Gram-negative species is not new, with polymerase-chain reaction (PCR)-based methods having been widely used in the epidemiological assessment of both non-enteropathogenic *E. coli* and *K. pneumoniae* collections. However, the number of resistance mechanisms involved is extremely large, limiting the use of comprehensive PCR methods in any real-time diagnostic capacity. One response to this challenge has been to develop microarray-based approaches to assess a much larger panel of resistance mechanisms than is feasible with PCR; this method, however, has issues with sensitivity and cannot easily identify numerous mutation-based mechanisms of resistance(27). In addition, microarrays are expensive to develop and difficult to upgrade flexibly in response to the evolution of resistance mechanisms. In this pilot study, I have demonstrated that WGS provides a viable alternative approach.

This study has also demonstrated that novel variants of known resistance-associated loci can be easily identified using my approach. To expand on this, BLASTn-based cut-offs could be made less stringent to facilitate the discovery of putative, distantly related resistance genes, or a tBLASTx-based approach could be used to identify

protein homologues with different underlying coding sequences. Similar approaches have been used in the past, although in a limited fashion(28).

My data highlight some known issues with the accuracy of some phenotypic methods commonly used in diagnostic microbiology – particularly with the assessment of beta-lactam/beta-lactamase inhibitor susceptibilities. Duplicate BD Phoenix tests gave discordant results in 7 (5%) of 143 co-amoxiclav tests performed; on gradient diffusion 6/7 isolates had MICs more than one dilution away from the breakpoint. In 21 instances where the co-amoxiclav genotypic prediction disagreed with the Phoenix results (all involving genotypically susceptible, BD Phoenix resistant isolates), all isolates were susceptible according to gradient diffusion, suggesting that in 15% of tests automated phenotyping was overcalling resistance. Problems with the correct assessment of susceptibility by phenotyping for beta-lactam/beta-lactamase inhibitor combinations has been observed previously(29), particularly in the context of complex beta-lactamase genotypes(30), which were disproportionately represented in my dataset.

For extended-spectrum cephalosporins and non-enteropathogenic *E. coli*, I found certain genetic mechanisms known to be associated with resistance (such as the CTX-M enzyme family) in isolates considered susceptible, even in the context of using the new, lower EUCAST breakpoints following EUCAST's decision (mirrored by the CLSI) to report susceptibilities as observed without interpretative modifications for these drugs. This phenomenon has also been documented in other studies of CTX-M-producing organisms from China and New Zealand(31, 32), and highlights the controversy over whether an *in vitro* MIC or the presence of a genetic mechanism is

more predictive of clinical outcomes(33) – whether this is a limitation or a strength of genotypic resistance prediction methods is therefore unclear. The large-scale clinical outcome data needed to resolve this quandary are currently lacking, but could be obtained by using integrated routine clinical, phenotyping, and antibiotic prescribing data, combined with whole genome-based, comprehensive assessments of resistance mechanisms.

Overall, among 1001 isolate-antimicrobial combinations tested, I found 12 instances of phenotypic resistance that were supported by gradient diffusion analysis without any resistance mechanism being identified, indicating deficits with my initial gene reference database and/or genotypic prediction algorithm. I have yet to systematically investigate potential contributions made by other known resistance mechanisms such as porin genes or efflux pumps, in part because associations of the latter with phenotypic resistance are incompletely defined. Assessing the performance of my approach in determining all known mechanisms of resistance, including rare variants, is clearly important future work; for this study, however, I was particularly focused on characterising the potential of genotypic resistance prediction for organisms typically isolated in our clinical practice. Of interest, given the absence of any initial mechanism identified for carbapenem resistance in the single meropenem-resistant *K. pneumoniae* isolate, I subsequently studied the *ompK35* and *ompK36* loci as possible candidate loci using my BLASTn-based approach, and identified a 5bp deletion in *ompK36* leading to a truncation at position 227. Although I did not measure protein expression, porin deficiencies associated with prematurely truncated *ompK36*, coupled with the presence of $bla_{\text{CTX-M-15}}$, have been associated previously with carbapenem resistance(34), and could plausibly explain resistance in this isolate. This

demonstrates that once an isolate's genome sequence is available, it can be re-assessed rapidly for additional resistance gene mechanisms as necessary, without the need for further laboratory work.

There are several limitations to my approach as described. Establishing the sequencing and computational infrastructure required to process large volumes of sequencing data in real-time involves a substantial initial investment in terms of time and money. My study was a retrospective, proof-of-principle experiment, and further work would be required to assess its performance and cost-effectiveness in a routine diagnostic setting on a larger dataset. In addition, it remains to be seen whether predictions would be equally successful for all antimicrobials currently incorporated in phenotypic susceptibility testing strategies. The bioinformatic strategy used does not determine plasmid copy number and therefore cannot quantify the possible contribution of multiple gene copies (e.g., of bla_{TEM}), which might lead to hyper-production of certain enzymes, and phenotypic resistance by a gene dosage effect. Another limitation is that the phenotypic manifestations of certain allelic variants and promoter/attenuator mechanisms are not fully determined (e.g., for some of the bla_{SHV} variants), precluding reliable predictions. Importantly, resistance mechanisms evolve; approaches based on genotypic prediction rely on a resistance locus reference database requiring regular updating based on a scheme incorporating ongoing phenotyping, albeit in a more limited number of samples, such as those isolated from treatment failures. Phenotyping would also be needed to validate any novel genetic resistance mutations/mechanisms. Finally, epigenetic and expression-associated mechanisms cannot be determined using my DNA-based analysis, thus highlighting the intrinsic limitation of approaches based on gene/mutation identification with no

direct evidence of functional resistance. However, alternative sequencing-based methods could be explored to address this shortcoming, such as RNA-seq, chromatin immunoprecipitation sequencing (chIP-Seq) or methylation analysis(35), although the pilot data presented here suggests the contributions of these mechanisms may be relatively minor.

Despite these limitations, my approach achieved high sensitivity and specificity in proof-of-principle experiments using typical clinical isolates, and performed comparably to some phenotyping methods currently in routine use. Whole genome sequencing-based approaches may well become part of routine microbiology workflows in some settings within the next five years. This would afford the ability to undertake species identification, strain typing for epidemiological purposes or infection prevention and control, and prediction of antimicrobial susceptibilities reliably and quickly using a single method for around £40/isolate(1). On the basis of this work, methods are currently being developed to ascertain what the best automated bioinformatic resistotyping approach would be, and a large set of additional non-enteropathogenic *E. coli* and *K. pneumoniae* clinical isolates has been processed ready for the validation of this approach to be made.

Table I. Genotypic profiles for loci associated with beta-lactam resistance among 74 non-enteropathogenic *E. coli* isolates.

Profile[a]	Number of isolates with profile (discrepant phenotypes)	Total number of mechanisms[b]	Predicted susceptibility for profile[c]				*bla*$_{TEM}$-promoter variant	*bla*$_{TEM}$ variant	*bla*$_{OXA-1}$	*bla*$_{CTX-M}$ variant	Relevant chromosomal *ampC* promoter/attenuator mutations
			AMP	AMC	CRO/CAZ	MEM					
1	23	0	S	S	S	S	-	-	-	-	-
2	12[d] (2)	3	R	R	R	S	P3	1	+	15	-
3	12[e] (2)	1	R	S	S	S	P3	1	-	-	-
4	10 (6)	2	R	R	R	S	-	-	+	15	-
5	5	1	R	S	S	S	novel	1	-	-	-
6	2 (1)	1	R	S	R	S	-	-	-	1	-
7	2	2	R	S	R	S	P3	1	-	15	-
8	1	2	R	R	R	S	-	-	+	14	-
9	1	1	R	R	S	S	-	-	+	-	-
10	1 (1)	1	R	R	R	S	-	-	-	-	T-32A (promoter)
11	1	2	R	R	R	S	novel	1	+	15	-
12	1[e]	1	R	S	S	S	novel	1	-	-	-
13	1	1	R	R	S	S	novel	30/IRT-2	-	-	-
14	1 (1)	2	R	S	R	S	P3	1	-	14	-
15	1	1	R	R	S	S	Pa/Pb	1	-	-	-

[a] Profiles are arranged in order of frequency in the dataset
[b] *bla*$_{TEM}$-promoter and *bla*$_{TEM}$ considered as a single element
[c] AMP = amoxicillin, AMC = co-amoxiclav, CRO = ceftriaxone, CAZ = ceftazidime, MEM; S = susceptible, R = resistant
[d] 8 *bla*$_{TEM}$ promoter sequences aligned from partial sequence matches across contigs
[e] One *bla*$_{TEM}$ sequence aligned from partial sequence matches across contigs

Table II. Genotypic profiles for loci associated with ciprofloxacin susceptibility/resistance among non-enteropathogenic *E. coli* isolates.

Profile[a]	Number of isolates with profile	Number of QRDR mutations in *gyrA*	Number of QRDR mutations in *parC*	Predicted ciprofloxacin susceptibility for profile	All *gyrA* mutations[b]	All *gyrB* mutations[b]	All *parC* mutations[b]	All *parE* mutations[b]	*aac-6'-Ib-cr*
1	20	2	2	R	S83L[c]; D87N; A828S	A618T	S80I; E84V; A471G; Q481H	V136I; I529L	+
2	15	0	0	S	A828S	-	-	V136I	-
3	6	0	0	S	-	E703D	-	-	-
4	3	0	0	S	A828S	A618T	A471G; Q481H	V136I	-
5	3	0	0	S	-	-	-	V136I	-
6	2	0	0	S	A828S	A618T	-	V136I; I529L	-
7	2	1	0	S	S83L	-	-	-	-
8	2	2	1	R	S83L; D87N; A828S	A618T	E84V; A471G; Q481H	V136I; I529L	+
9	2	2	2	R	S83L; D87N; A828S	A618T	S80I; E84V; A471G; Q481H	V136I; I529L	-
10	1	0	0	S	A828S	M605L; E703D	-	V136I	-
11	1	0	0	S	A828S	T590M	-	V136I	-
12	1	0	0	S	A828S	-	-	V136I; D475E	-
13	1	0	0	S	A828S	-	-	V136I; I529L	-
14	1	0	0	S	A828S	A618T	-	V136I; K146T D475E	-
15	1	0	0	S	A828S	-	-	-	-

16	1	1	0	S	S83L	S492N	-	-	-
17	1			S	S83L; A828S	A618T	-	V136I; I529L	-
18	1	1	0	R	S83L; D87N	-	S80I	S458A	+
19	1	2	2	R	S83L; D87N; A828S	-	S80I; E84V; A471G; Q481H	V136I; I529L	-
20	1	0	0	S	-	E159D; E161D; R206K; S463A; V522I; T565S; T590A; N595V; Q737K; E703D	-	V136I frameshift at position 307; truncation at position 351	-
21	1	0	0	S	-	A618T; T692A	-	-	-
22	1	0	0	S	-	E703D	-	-	+

[a] Profiles are arranged in order of frequency in the dataset
[b] Mutations compared to wild type locus of *E. coli* MG1655 (RefSeq: NC_000913)
[c] Mutations in QRDR regions underlined

Table III. Genotypic profiles for genes associated with aminoglycoside resistance among 74 non-enteropathogenic *E. coli* isolates.

Profile[a]	Number of isolates with profile	Total number of mechanisms conferring aminoglycoside resistance (number conferring gentamicin resistance)	Predicted gentamicin susceptibility for profile	*aac-3'-IId*	*aac-3'-IIe*	*aac-6'-Ib-cr*	*aadA1a*	*aadA4*	*aadA5*	*aph-3'-Id*	*aph-6'-Id*[b]
1	32	0 (0)	S	-	-	-	-	-	-	-	-
2	8	1 (0)	S	-	-	-	-	-	-	-	+[c]
3	10	3 (0)	S	-	-	+	-	+	+	-	-
4	8	2 (1)	R	-	+	+	-	-	-	-	-
5	3	3 (0)	S	-	-	-	-	+	+	-	+
6	3	2 (0)	S	-	-	-	-	-	-	+[d]	+
7	2	5 (1)	R	-	+	+	-	+	+	-	+[e]
8	2	3 (1)	R	-	+	+	-	-	-	-	+[e]
9	3	1 (0)	S	-	-	-	+[f]	-	-	-	-
10	1	4 (1)	R	-	+	+	-	+	+	-	-
11	1	2 (0)	S	-	-	+	-	-	-	+	-
12	1	4 (1)	R	+	-	-	-	+	+	-	+[e]

[a] Profiles are arranged in order of frequency in the dataset
[b] Two *aph-6'-Id* sequences aligned from partial matches across contigs
[c] One isolate with 99.9% sequence homology to reference gene; three with 87.6%
[d] Two isolates with 99.4% sequence homology; one isolate with 99.1% sequence homology
[e] One isolate with 99.9% sequence homology
[f] Two isolates with 99.7% sequence homology to database reference sequence and 100% homology to JX494725 (*C. freundii* plasmid with aadA1 gene [position 633-1424])

Table IV. Genotypic profiles for loci associated with beta-lactam resistance among 69 *K. pneumoniae* isolates.

Profile[a]	Number of isolates with profile (discrepant phenotypes)	Total number of mechanisms[b]	Predicted susceptibility for profile[c]				*bla*$_{TEM}$-promoter variant	*bla*$_{TEM}$-variant	*bla*$_{OXA-1}$	*bla*$_{CTX-M}$ variant	*bla*$_{SHV}$[d] mutations; allelic variant	*bla*$_{LAP}$ variant	*bla*$_{LEN}$	*bla*$_{OKP}$
			AMP	AMC	CRO/CAZ	MEM								
1	13 (1)	1	R	S	S	S	-	-	-	-	L35Q; 11	-	-	-
2	14	4	R	R	R	S	P3	1	1	15	WT[e]; 1	-	-	-
3	11 (1)	1	R	S	S	S	-	-	-	-	WT; 1	-	-	-
4	6 (2)	1	R	S	S	S	-	-	-	-	-	-	2 *bla*$_{LEN-7}$, 4 novel	-
5	3	3	R	S	R	S	P3	1	-	15	WT; 1	-	-	-
6	2	1	R	S	S	S	-	-	-	-	Y7F; 28	-	-	-
7	2 (1)	4	R	S	S	S	P3	1	1	15	T18A, A22V, L35Q, M129V; 121	-	-	-
8	2 (1)	2	R	R	S	S	Pa/Pb	1	-	-	L35Q; 11	-	-	-
9	1	3	R	R	R	S	-	-	+	15	L35Q; 11	-	-	-
10	1	1	R	S	S	S	-	-	-	-	D101H; 169[f]	-	-	-
11	1 (1)	1	R	S	R	S	-	-	-	-	G156D; 27	-	-	-
12	1	1	R	S	S	S	-	-	-	-	L35Q, A187T; 60	-	-	-
13	1	1	R	S	S	S	-	-	-	-	P226S; 33	-	-	-
14	1	1	R	S	S	S	-	-	-	-	Y7F, M211L; 170[f]	-	-	-
15	1	1	R	S	S	S	-	-	-	-	Y7F, S14F; 171[f]	-	-	-
16	1	1	R	S	S	S	-	-	-	-	P226S; 33	-	-	-

17	1	1	R	S	S	S	-	-	-	-	-	-	-	6
18	1	4	R	R	R	S	P3	1	1	15	Y7F; 28	-	-	-
19	1 (1)	3	R	R	R	S	P3	1	1	15	-	-	-	-
20	1	3	R	S	R	S	P3	1	-	15	Y7F; 28	-	-	-
21	1	2	R	S	S	S	P3	1	-	-	E31K, L35Q, A187T; 135	-	-	-
22	1	2	R	S	S	S	P3	1	-	-	L35Q; 11	-	-	-
23	1	3	R	R	R	S	Pa/Pb	1	-	15	L35Q; 11	-	-	-
24	1	2	R	S	S	S	-	-	-	-	L35Q; 11	2	-	-

[a] Profiles are arranged in order of frequency in the dataset
[b] bla_{TEM}-promoter and bla_{TEM} considered as a single element; one bla_{TEM}-promoter aligned from partial matches across several contigs
[c] AMP = amoxicillin, AMC = co-amoxiclav, CRO = ceftriaxone, CAZ = ceftazidime, MEM; S = susceptible, R = resistant
[d] One bla_{SHV} sequence aligned from partial matches across contigs
[e] WT – wild type; equivalent to bla_{SHV-1}
[f] Novel bla_{SHV}-variant identified in this study

Table V. Genotypic profiles for loci associated with ciprofloxacin susceptibility/resistance among 69 *K. pneumoniae* isolates.

Profile[a]	Number of isolates with profile (discrepant phenotypes)	Number of *gyrA* QRDR mutations	Number of *parC* QRDR mutations	Number of *parE* QRDR mutations	Predicted ciprofloxacin susceptibility for profile	Amino acid mutations in quinolone resistance determining regions (QRDR) only					*aac-6'-Ib-cr*	*qnr* variant
						gyrA		*parC*		*parE*		
						position 83; wild type S	position 87; wild type D	position 80; wild type S	position 84; wild type E	position 458; wild type S		
1	37	-	-	-	S	S	D	S	E	S	-	-
2	12	-	-	-	R	S	D	S	E	S	+	B1
3	2	2	-	-	R	Y	N	S	E	S	+	B1
4	2 (2)	-	1	-	S	S	D	I	E	S	-	-
5	1 (1)	2	-	-	R	F	A	S	E	S	-	-
6	1	2	1	-	R	F	N	I	E	S	+	-
7	1	1	1	-	R	I	D	I	E	S	+	-
8	1	1	-	-	S	I	D	S	E	S	-	B1
9	1	1	-	-	S	I	D	S	E	S	-	-
10	1 (1)	2	-	-	R	I	N	S	E	S	-	-
11	1	-	1	-	R	S	D	I	E	S	+	B1
12	1 (1)	-	1	-	S	S	D	I	E	S	+	-
13	1	-	-	-	S	S	D	S	E	S	+	-
14	1	-	-	-	S	S	D	S	E	S	-	S1
15	1	-	-	-	S	S	D	S	E	S	-	B1
16	1	-	-	-	S	S	D	S	E	S	-	B6
17	1 (1)	-	1	1	S	S	D	S	K	T	-	-
18	1	1	1	-	R	T	D	I	E	S	-	S1
19	1	1	-	-	S	T	D	S	E	S	-	-
20	1	1	-	-	S	Y	D	S	E	S	-	-

[a] Profiles are arranged in order of frequency in the dataset

Table VI. Genotypic profiles associated with aminoglycoside resistance among 69 *K. pneumoniae* isolates.

Profile[a]	Number of isolates with profile (discrepant phenotypes)	Total number of mechanisms conferring aminoglycoside resistance (number conferring gentamicin resistance)	Predicted gentamicin susceptibility for profile	*aac-3'-Ia*	*aac-3'-IId*	*aac-3'-IIe*	*aac-6'-Ib-cr*	*aadA1a*	*aadA2*	*aadA4*	*aadA16*	*aph-3'-Ia*	*aph-6'-Id*
1	36 (1)	0	S	-	-	-	-	-	-	-	-	-	-
2	14	3 (1)	R	-	-	+	+	-	-	-	-	-	+
3	4	1 (0)	S	-	-	-	-	-	-	-	-	-	+
4	2	2 (1)	R	-	-	+	+	-	-	-	-	-	-
5	1	2 (1)	R	-	+	-	-	-	-	-	-	-	+
6	1	3 (1)	R	-	+	-	-	-	-	-	-	+[b]	+
7	1	1 (1)	R	-	-	+[c]	-	-	-	-	-	-	-
8	1	5 (1)	R	-	-	+	+	-	+[d]	-	-	+[b]	+
9	1	2 (1)	R	-	+	-	-	-	+[d]	-	-	-	-
10	1	2 (1)	R	+	-	-	-	+[e]	-	-	-	-	-
11	1	2 (0)	S	-	-	-	+	-	-	-	-	-	+
12	1	2 (0)	S	-	-	-	+	-	-	-	+	-	-
13	1	2 (0)	S	-	-	-	-	+[f]	-	-	-	-	+
14	1	2 (0)	S	-	-	-	-	-	+[d]	-	-	-	+
15	1	1 (0)	S	-	-	-	-	-	+[d]	-	-	-	-
16	1	4 (1)	R	-	-	+	+	-	-	-	-	+[b]	+
17	1	1 (0)	S	-	-	-	-	-	-	-	-	+[c]	-

[a] Profiles are arranged in order of frequency in the dataset
[b] 99.4% sequence homology over full length of gene in reference database
[c] 99.8% sequence homology over full length of gene
[d] 99.9% sequence homology over full length of gene
[e] 99.7% sequence homology over full length of gene
[f] 100% sequence homology, but over 86% of gene

CHAPTER 5 REFERENCES

1. **Didelot X, Bowden R, Wilson DJ, Peto TE, Crook DW.** 2012. Transforming clinical microbiology with bacterial genome sequencing. Nature reviews. Genetics **13:**601-612.
2. **Jorgensen JH, Ferraro MJ.** 2009. Antimicrobial susceptibility testing: a review of general principles and contemporary practices. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America **49:**1749-1755.
3. **Sader HS, Fritsche TR, Jones RN.** Accuracy of three automated systems (MicroScan WalkAway, VITEK, and VITEK 2) for susceptibility testing of Pseudomonas aeruginosa against five broad-spectrum beta-lactam agents. J Clin Microbiol. 2006 Mar;44(3):1101-4.
4. **Andrews JM.** 2001. The development of the BSAC standardized method of disc diffusion testing. The Journal of antimicrobial chemotherapy **48 Suppl 1:**29-42.
5. **Snyder JW, Munier GK, Johnson CL.** 2008. Direct comparison of the BD phoenix system with the MicroScan WalkAway system for identification and antimicrobial susceptibility testing of Enterobacteriaceae and nonfermentative Gram-negative organisms. Journal of clinical microbiology **46:**2327-2333.
6. **Zankari E, Hasman H, Kaas RS, Seyfarth AM, Agerso Y, Lund O, Larsen MV, Aarestrup FM.** 2013. Genotyping using whole-genome sequencing is a realistic alternative to surveillance based on phenotypic antimicrobial susceptibility testing. The Journal of antimicrobial chemotherapy **68:**771-777.

7. **Wilson J, Elgohari S, Livermore DM, Cookson B, Johnson A, Lamagni T, Chronias A, Sheridan E.** 2011. Trends among pathogens reported as causing bacteraemia in England, 2004-2008. Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases **17:**451-458.

8. **Gagliotti C, Balode A, Baquero F, Degener J, Grundmann H, Gur D, Jarlier V, Kahlmeter G, Monen J, Monnet DL, Rossolini GM, Suetens C, Weist K, Heuer O, Participants EA-N.** 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 surveillance : bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin **16**.

9. **Schlackow I, Stoesser N, Walker AS, Crook DW, Peto TE, Wyllie DH, Infections in Oxfordshire Research Database T.** 2012. Increasing incidence of Escherichia coli bacteraemia is driven by an increase in antibiotic-resistant isolates: electronic database study in Oxfordshire 1999-2011. The Journal of antimicrobial chemotherapy **67:**1514-1524.

10. **Webster DP, Young BC, Morton R, Collyer D, Batchelor B, Turton JF, Maharjan S, Livermore DM, Bejon P, Cookson BD, Bowler IC.** 2011. Impact of a clonal outbreak of extended-spectrum beta-lactamase-producing Klebsiella pneumoniae in the development and evolution of bloodstream infections by K. pneumoniae and Escherichia coli: an 11 year experience in Oxfordshire, UK. The Journal of antimicrobial chemotherapy **66:**2126-2135.

11. **Paterson DL.** 2006. Resistance in Gram-negative bacteria: Enterobacteriaceae. American journal of infection control **34:**S20-28; discussion S64-73.

12. **EUCAST** 2013, posting date. Breakpoint Tables for Interpretation of MICs and Zone Diameters; Version 3.0 (1 January 2013). [Online.]

13. **Chemotherapy TBSfA** 2012, posting date. Use of Gradient Tests for Determination of MICs by BSAC Methodology. [Online.]

14. **Ramirez MS, Tolmasky ME.** 2010. Aminoglycoside modifying enzymes. Drug resistance updates : reviews and commentaries in antimicrobial and anticancer chemotherapy **13:**151-171.

15. **Lartigue MF, Leflon-Guibout V, Poirel L, Nordmann P, Nicolas-Chanoine MH.** 2002. Promoters P3, Pa/Pb, P4, and P5 upstream from bla(TEM) genes and their relationship to beta-lactam resistance. Antimicrobial agents and chemotherapy **46:**4035-4037.

16. **Mulvey MR, Bryce E, Boyd DA, Ofner-Agostini M, Land AM, Simor AE, Paton S.** 2005. Molecular characterization of cefoxitin-resistant Escherichia coli from Canadian hospitals. Antimicrobial agents and chemotherapy **49:**358-365.

17. **Caroff N, Espaze E, Berard I, Richet H, Reynaud A.** 1999. Mutations in the ampC promoter of Escherichia coli isolates resistant to oxyiminocephalosporins without extended spectrum beta-lactamase production. FEMS microbiology letters **173:**459-465.

18. **Caroff N, Espaze E, Gautreau D, Richet H, Reynaud A.** 2000. Analysis of the effects of -42 and -32 ampC promoter mutations in clinical isolates of

Escherichia coli hyperproducing ampC. The Journal of antimicrobial chemotherapy **45:**783-788.

19. **Jacoby GA.** 2005. Mechanisms of resistance to quinolones. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America **41 Suppl 2:**S120-126.

20. **Cattoir V, Nordmann P.** 2009. Plasmid-mediated quinolone resistance in Gram-negative bacterial species: an update. Current medicinal chemistry **16:**1028-1046.

21. **Zankari E, Hasman H, Cosentino S, Vestergaard M, Rasmussen S, Lund O, Aarestrup FM, Larsen MV.** 2012. Identification of acquired antimicrobial resistance genes. The Journal of antimicrobial chemotherapy **67:**2640-2644.

22. **Jacoby GB, K.** 2012, posting date. β-Lactamase Classification and Amino Acid Sequences for TEM, SHV and OXA Extended-spectrum and Inhibitor-resistant Enzymes. [Online.]

23. **Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ.** 1990. Basic local alignment search tool. Journal of molecular biology **215:**403-410.

24. **Gouy M, Guindon S, Gascuel O.** 2010. SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Molecular biology and evolution **27:**221-224.

25. **Sutcliffe JG.** 1978. Nucleotide sequence of the ampicillin resistance gene of Escherichia coli plasmid pBR322. Proceedings of the National Academy of Sciences of the United States of America **75:**3737-3741.

26. **Yuan J, Xu X, Guo Q, Zhao X, Ye X, Guo Y, Wang M.** 2012. Prevalence of the oqxAB gene complex in Klebsiella pneumoniae and Escherichia coli clinical isolates. The Journal of antimicrobial chemotherapy **67:**1655-1659.

27. **Card R, Zhang J, Das P, Cook C, Woodford N, Anjum MF.** 2013. Evaluation of an expanded microarray for detecting antibiotic resistance genes in a broad range of Gram-negative bacterial pathogens. Antimicrobial agents and chemotherapy **57:**458-465.

28. **Davis MA, Baker KN, Orfe LH, Shah DH, Besser TE, Call DR.** 2010. Discovery of a gene conferring multiple-aminoglycoside resistance in Escherichia coli. Antimicrobial agents and chemotherapy **54:**2666-2669.

29. **Bond A, Plumb H, Turner P.** 2012. Susceptibility testing of Escherichia coli isolates from urines: are we at risk of reporting false antibiotic resistance to co-amoxiclav? The Journal of antimicrobial chemotherapy **67:**1557-1558.

30. **Canton R, Loza E, Del Carmen Conejo M, Baquero F, Martinez-Martinez L, Group MC.** 2003. Quality control for beta-lactam susceptibility testing with a well-defined collection of Enterobacteriaceae and Pseudomonas aeruginosa strains in Spain. Journal of clinical microbiology **41:**1912-1918.

31. **Wang P, Hu F, Xiong Z, Ye X, Zhu D, Wang YF, Wang M.** 2011. Susceptibility of extended-spectrum-beta-lactamase-producing Enterobacteriaceae according to the new CLSI breakpoints. Journal of clinical microbiology **49:**3127-3131.

32. **Williamson DA, Roberts SA, Smith M, Heffernan H, Tiong A, Pope C, Freeman JT.** 2012. High rates of susceptibility to ceftazidime among globally prevalent CTX-M-producing Escherichia coli: potential clinical implications of the revised CLSI interpretive criteria. European journal of clinical

microbiology & infectious diseases : official publication of the European Society of Clinical Microbiology **31:**821-824.

33. **Livermore DM, Andrews JM, Hawkey PM, Ho PL, Keness Y, Doi Y, Paterson D, Woodford N.** 2012. Are susceptibility tests enough, or should laboratories still seek ESBLs and carbapenemases directly? The Journal of antimicrobial chemotherapy **67:**1569-1577.

34. **Novais A, Rodrigues C, Branquinho R, Antunes P, Grosso F, Boaventura L, Ribeiro G, Peixe L.** 2012. Spread of an OmpK36-modified ST15 Klebsiella pneumoniae variant during an outbreak involving multiple carbapenem-resistant Enterobacteriaceae species and clones. European journal of clinical microbiology & infectious diseases : official publication of the European Society of Clinical Microbiology **31:**3057-3063.

35. **Guell M, Yus E, Lluch-Senar M, Serrano L.** 2011. Bacterial transcriptomics: what is beyond the RNA horiz-ome? Nature reviews. Microbiology **9:**658-669.

CHAPTER 6: OUTBREAK EPIDEMIOLOGY - USE OF WGS TO INVESTIGATE AN EXTENDED SERIES OF NDM-1-POSITIVE *KLEBSIELLA PNEUMONIAE* INFECTIONS IN A SINGLE INSTITUTION IN NEPAL

6.1. BACKGROUND

Klebsiella pneumoniae is a major pathogen in neonatal critical care(1), and in association with extended-spectrum beta-lactamases (ESBLs) and/or carbapenemases, results in increased hospital costs, longer stays and high mortality(2, 3). Furthermore, the success of controlling outbreaks with carbapenemase-producing *Klebsiella* spp. is mixed(4, 5). In resource-limited settings, such as South Asia, multi-drug resistant Enterobacteriaceae are widespread, and to delineate the relative contribution of within-hospital transmission of resistant strains and mobile genetic elements versus recurrent importation of these from the community is challenging.

The New Delhi Metallo-beta-lactamase family of carbapenemases (NDM; conferred by *bla*NDM variants) were first identified in a *K. pneumoniae* strain cultured from a patient originally hospitalised in India(6). Since then, at least eleven different gene variants have emerged globally(7), in several different plasmid backgrounds and bacterial species(8). South Asia is one potential reservoir for these genes, with an NDM carriage prevalence of 14% in out-patients and 27% in hospitalized patients reported in Pakistan(9), and NDM-1 genes detected in 4% of drinking water and 30% of waste seepage samples from New Delhi, India(10).

Several studies have used lower-resolution typing methods, such as pulsed field gel electrophoresis (PFGE), to characterize outbreaks caused by carbapenem-resistant

Klebsiella spp., but these are limited in their capacity to accurately define genetic relationships, particularly with respect to closely related isolates, or in contexts where mobile genetic elements may play a contributory role to transmission of resistance genes. Whole genome sequencing (WGS) has already been shown to markedly alter classical interpretations of transmission at a bacterial strain level, with respect to *Clostridium difficile* and *Mycobacterium tuberculosis* (11, 12). Two published studies have successfully utilized WGS to describe carbapenemase-producing outbreaks of *K. pneumoniae*, but both of these again focused only on the relationships of bacterial strains in the presence of a resistance gene without reference to its genetic context (13, 14). The first study described a six-month outbreak involving a different carbapenemase gene, *Klebsiella pneumoniae* carbapenemase (bla_{KPC}), (22 isolates) using genetic and epidemiological data, supporting both patient-patient and ventilator-patient transmission. The second study was smaller (n=8) and revealed likely inter-patient transmission of NDM-K. pneumoniae at a US hospital over four months.

This chapter expanded on previously published work, by using Illumina and Pacific Biosciences single molecule real-time sequencing (PacBio SMRT) to recover the near-complete genetic information, including that of episomal structures, of a series of disease-causing, multi-drug resistant *K. pneumoniae* isolates collected in a hospital neonatal service in Nepal. In addition to sequencing strains obtained from epidemiologically defined "outbreak" cases, other *K. pneumoniae* and Enterobacteriaceae isolates were sequenced in an attempt to capture the diversity of genes, plasmids and strains potentially involved in bla_{NDM} transmission networks. In a setting where NDM is thought to be endemic, this was undertaken to investigate the hypothesis that these drug-resistant cases were occurring as a result of independent

introductions of strains/plasmids from the wider community rather than within-hospital transmission.

6.2. METHODS

Patan hospital is a 450-bedded hospital in Kathmandu, Nepal, with a paediatric unit handling ~9,500 live births, 35,000 out-patient reviews and 2,600 admissions/year. Neonatal care is managed in two nurseries, one "clean", for non-sepsis-related supportive care, and one "septic", for patients who have major risk factors for, or a diagnosis of, sepsis. The neonatal and paediatric intensive care units (NICU and PICU) are located close to the nurseries.

The suspected index case was a low-birth weight, premature baby born in the hospital on 09/Aug/2011. Post-natally he was admitted to the NICU for respiratory distress. Early blood cultures were negative, but on day eight, *K. pneumoniae* was cultured, initially susceptible only to carbapenems. The child died four days later, despite treatment.

Cases of sepsis associated with similar drug-resistant *K. pneumoniae*-positive cultures susceptible only to one/more of the carbapenems/colistin/tigecycline were defined as part of an epidemiological cluster, with the last case identified on 30/Jun/2012. To contextualize genetic variation in cluster isolates, and identify wider resistance gene/element spread, we requested a randomly sampled subset of stored Enterobacteriaceae/other Gram-negative bacilli, stratified by susceptibility profile, age (adults/children), and hospital- (sampled ≥48 hours after admission) versus

community-associated infections (sampled <48 hours after admission, no previous admission to Patan hospital and delivery outside of a healthcare facility).

Infection control was enhanced from December 2011 and included changes in cleaning, equipment and surveillance protocols as follows:

1. 1% Virkon to clean equipment; previously, Savlon (0.3g chlorhexidine gluconate and 3g cetrimide/100mLs; diluted 1:30 in 70% alcohol)
2. Surface cleaning of floors, walls, ceilings, furniture and medical equipment following the identification of each new case.
3. Fumigation of the neonatal units on 02/Nov/2011 and 15/Dec/2011.
4. Inspection of and chlorination of water sources (27/Nov/2011)
5. Water purification by reverse osmosis and the use of purified water in ventilator humidifiers
6. Paper towels to replace cloth towels in all neonatal units
7. Routine change of staff and visitor gowns every morning
8. Hand washing implemented for all visitors; importance of hand hygiene emphasized to all staff on the neonatal units
9. Use of disposable tubing for ventilators
10. Cleaning protocols for ventilators implemented
11. Bedside chlorhexidine hand cleansing solution (Microshield) deployed in all neonatal units (equivalent to 0.5% w/v chlorhexidine gluconate + 70% ethanol v/v)
12. Implementation of microbiological surveillance

13. Closure of original clean nursery and repair of a leaking toilet in the ward above, which was being used by women undergoing vesico-vaginal fistula repair surgery.

6.2.1. Laboratory/sequencing methods

All samples were originally processed locally. Species identification was based on biochemical profiling. Antimicrobial susceptibility testing used disk-diffusion(15), and was reassessed using broth microdilution in Oxford (BD Phoenix Automated Microbiology System [Oxford, UK]), as per the thesis methods section.

Environmental sampling is performed routinely as part of infection control surveillance at Patan hospital, and this included the study period. Typically, paired swab samples are collected fortnightly before and after intensive cleaning from five different randomly selected sites in the NICU, and the established clean and septic nurseries, including laryngoscope blades, ventilators, stethoscopes, equipment trolleys, incubators, central lines, endotracheal/suction tubes, tap water, floors and door handles. If a patient was considered part of a *K. pneumoniae* case cluster, then a set of these surveillance swabs was taken from their incubator or bed. Rectal swabs were taken from all nursing and medical staff in the neonatal nurseries, NICU and PICU (between 9-11th July/2012). Samples were also collected from purified water used for clinical purposes (26/September/2011), disinfectants (13/December/2011), and air conditioners in the nurseries and ICUs (12/December/2011). Swabs/samples were plated on MacConkey agar and incubated aerobically; *Klebsiella* isolates were identified as above.

DNA was extracted from sub-cultured frozen isolate stocks as per the standard thesis method. All extracts were sequenced on the HiSeq 2000 platform (Illumina, San Diego, USA), generating 100 base paired-end reads. For the suspected outbreak index case (PMK1), PacBio SMRT sequencing was also undertaken.

6.2.2. Sequence data analysis

Species identification was confirmed with Kraken(16), and Illumina reads were mapped to species-specific reference genomes (*E. coli* CFT073, GenBank: AE014075.1; *Klebsiella pneumoniae* subsp. *pneumoniae* MGH78578, GenBank: CP000647.1; *Klebsiella oxytoca* E718, RefSeq: NC_018106.1; *Enterobacter cloacae* subsp. *cloacae* ATCC 13047, RefSeq: CP001918.1; *Pseudomonas aeruginosa* PA01, RefSeq: NC_002516.2; *Serratia marcescens* WW4; RefSeq: NC_020211; *Pantoea* sp. At9b, RefSeq: NC_014837.1).
Mapping/variant calling methodologies were done through the group pipeline; reads were also de novo assembled using Velvet and VelvetOptimiser. BLASTn was used to identify the presence/absence of resistance gene variants in assemblies using the resistance reference database defined in Chapter 5; at this stage, the identification process had been partly automated using an in-house python script. A similar BLASTn-based approach was used to infer multi-locus sequence types of *K. pneumoniae* isolates(17).

A maximum-likelihood phylogeny of the first *K. pneumoniae* isolate per case based on single nucleotide variants (SNVs) distributed over the core, mapped genome was constructed in PhyML(18). I used PhyML version 3.0, with a generalized time-reversible (GTR) nucleotide substitution model, a gamma-distribution with four rate

categories to estimate among-site variation in substitution rates, and 100 bootstrap replicates. Sites where at least one sample had a null/missing call were excluded from the input. For the input alignment, the variant sites derived from mapping to the MGH78578 reference were "padded" with invariant sites in a proportion consistent with the GC content and length of the reference genome (5.69Mb, 57.1% GC content).

PacBio sequencing of the isolate obtained from the first observed infected case (PMK1) was undertaken at the Department of Genetics and Genomics at the Icahn Medical Institute in New York, and assembled into contigs using their HGAP and Quiver software(19). Chromosomal and plasmid contigs were identified using BLASTn(20) against the NCBI nucleotide sequence database. Circular overlaps in plasmid contigs were identified with nucmer(21). Plasmid reference sequences were closed and corrected with visual inspection of BWA-generated Illumina and PacBio read alignments to plasmid contigs(22), and annotated with Prokka(23).

Reads from each strain were mapped to the PacBio-generated reference sequence (chromosome+plasmids) with BWA. Structural variation in plasmids was identified by plotting the mean read coverage for each 1000bp of the reference sequence divided by the mean read coverage across the whole reference sequence, capped to a maximum of one. Precise breakpoints were identified from the inspection of mapped reads. Presumptive plasmid structures generated are denoted pPMK[isolate number]-[plasmid reference suffix] e.g. pPMK17-NDM. NDM copy number was estimated by read counts to the NDM regions divided by the total number of reads, scaled to pPMK1-NDM, and then rounded to the nearest whole number.

The genetic outbreak was then defined as any closely genetically-related cases from the epidemiological cluster, plus other genetically related isolates, forming a distinct group on the maximum-likelihood tree. Comparing these to the PacBio-generated reference chromosome, a mutation rate for the outbreak strains was estimated using a time-scaled analysis in BEAST(24), including longitudinal isolates from individuals. Three separate runs on the dataset were undertaken using a strict molecular clock model with the following priors: (i) a GTR nucleotide substitution model with estimated base frequencies; (ii) a discrete gamma distribution with four categories to account for variable substitution rates at each site; (iii) a constant population size; (iv) a random starting tree; and (v) a Monte Carlo Markov Chain (MCMC) length of 30000000 with sampling logged every 1000 iterations. The output of the three runs with respect to mixing and convergence was compared using Tracer v1.5 [33]; good mixing and convergence were observed and effective sampling sizes for all parameters were above 300. Log and tree files for the respective runs were combined with down-sampling using LogCombiner v1.7.5; mutation rates and the phylogeny were determined from these. TreeAnnotator v1.7.5 was used to select the maximum clade credibility tree.

For the input alignment, 51 variant sites derived from mapping to the Pacbio-derived chromosomal reference for the outbreak strain (excluding positions 3126261 and 3137776 which had been affected by the large deletion in PMK13b) were “padded” with invariant sites (in proportions consistent with the ACGT content of reference chromosome) to the length of the called genome (5,038,898/5,317,001 bases; represents sites where bases were called in all sequences). The molecular clock

generated by BEAST was then multiplied by the called genome length to give a mutation rate/genome/year.

The most likely transmission chain was inferred using the R Outbreaker package(25). An alignment of the reference sequence with inserted SNVs for each of the first isolates obtained from each outbreak case was uploaded into R. The probability distribution of the generation time (the time between colonization of a primary case and transmission to a secondary case) was set up to follow an exponential decline, with the highest probability for transmission estimated for cases closer in time to the primary case. I tried to run the Outbreaker package with two different mean generation times, namely 50 days and 10 days. Four parallel MCMC runs were performed for 30,000,000 iterations. Good convergence on visualization of the trace of log-posterior values was observed (3,000,000 iterations were removed as burn-in) only for the first model (with a mean generation time of 50 days). This would be consistent with observations in the literature of colonisation with *bla*NDM-positive *Enterobacteriaceae* persisting for up to a year(26, 27); a mean generation time estimate of 50 days was therefore used for the Outbreaker analysis.

Posterior support for the edges of the transmission tree and estimated mean times to infection were obtained from the posterior ancestries ("alpha" columns in the Outbreaker MCMC output) and the dates of infection ("Tinf" columns in the Outbreaker MCMC output). The posterior distribution of R values (the R value being the number of secondary cases per infected individual), used to size the nodes in the transmission graph were obtained using the get.R function in the package.

6.3. RESULTS

Of 102 strains sequenced, 55 were confirmed as *K. pneumoniae*, 43 as other Enterobacteriaceae and 4 as other Gram-negative bacilli. Of 55 *K. pneumoniae* sequenced from 47 individuals, 34 isolates sampled from 26 individuals were part of the genetically-defined outbreak. All outbreak isolates were cultured from blood, except PMK9, cultured from cerebrospinal fluid. Figure 6.1. shows the timeline of cases in the epidemiologically-defined cluster and/or the genetically-defined outbreak. The prolonged time intervals between some cases (PMK1-to-PMK3, 96 days; PMK9-to-PMK10, 57 days; PMK12-to-PMK15, 34 days) demonstrate the challenge of distinguishing between multiple importations and on-going spread. The outbreak affected all units, including the overflow nurseries established towards the end of the outbreak in a different hospital building.

Outbreak-associated mortality was high, with 16 (64%) in-patient deaths in 25 neonates (6 with unknown outcome e.g. referred elsewhere, left against medical advice), in contrast with a hospital-wide contemporaneous neonatal death rate of 46/6908 (0.7%). Neonatal critical care mortality was 46% (45/98 cases) during the outbreak period, and 27% (32/117) in the year following the final case (p=0.007, Fisher's Exact Test).

Approximately 30 sites per month were sampled from the three neonatal units surveyed; the heaviest environmental contamination with *Klebsiella* spp. was observed at the onset of the outbreak, between August 2011 and January 2012 (mean: 2 environmental swabs positive [range: 0-3]), as opposed to the preceding and subsequent six-month periods (no sites positive at any time-point, and mean: 0 sites

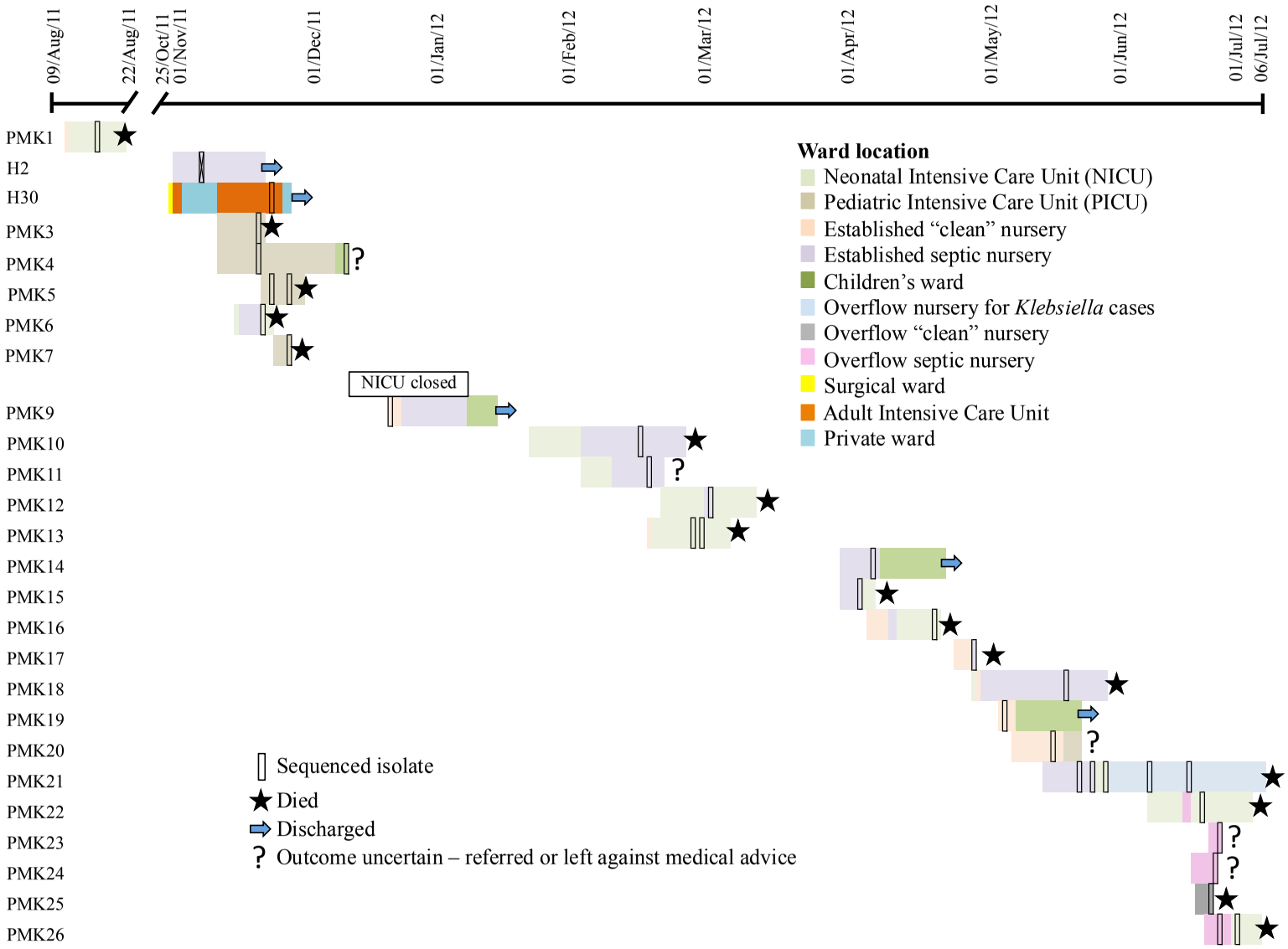


Figure 6.1. Representation of the timeline of *Klebsiella pneumoniae* cases, including individuals both part of epidemiologically-defined case clusters and genetically-linked outbreak strains. H2 was found to be genetically unrelated to the other outbreak isolates; H30 was thought to share a relatively recent common ancestor, but was not involved in the immediate transmission network. There were no clinical details available for H1460, which is therefore not shown.

positive [range: 0-1] respectively) (Figure 6.2.). *Klebsiella* spp. were isolated from a number of environmental cultures including from laryngoscope blades, suctioning equipment, purified water containers, tap water, soap dispensers, a burette set surface swab and a health care worker's hands; none of these underwent susceptibility testing. Eighteen of 69 (26%) rectal swabs cultured *Klebsiella* spp., but none were ESBL- or carbapenemase-producing. None of these isolates were stored unfortunately, and therefore none could be sequenced to test for the environmental presence/carriage of any genetic elements potentially involved in transmission.

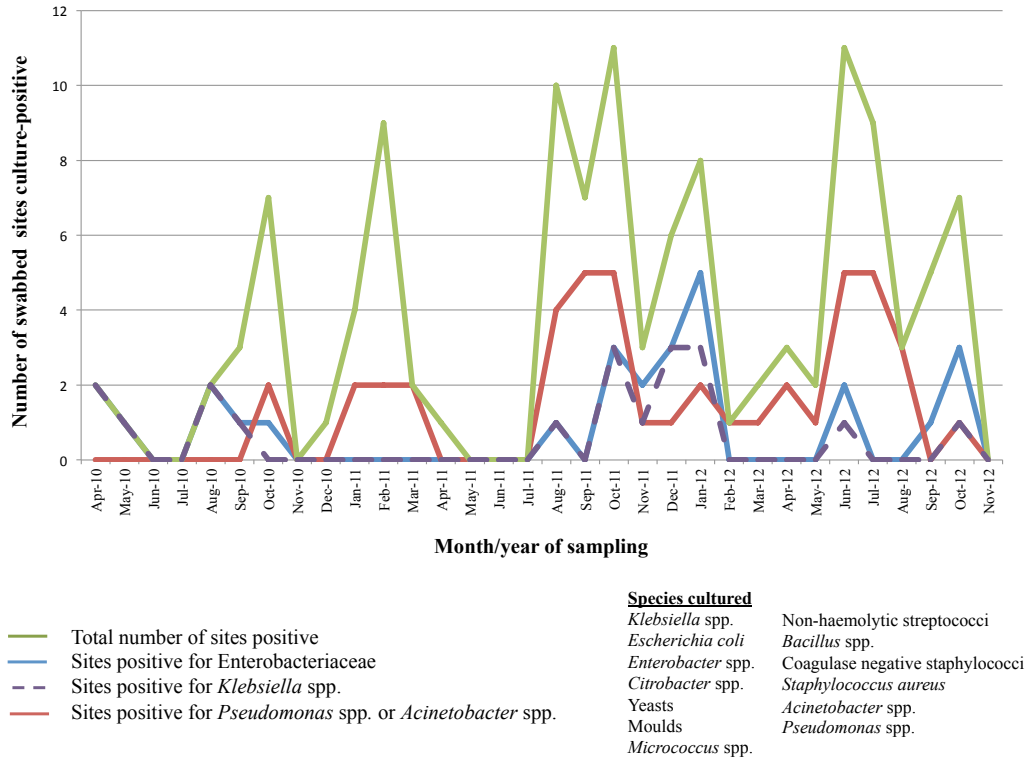


Figure 6.2. Number of environmental samples positive, stratified by organism by month, April 2010-November 2012. The NDM *K. pneumoniae* outbreak lasted from August 2011-July 2012.

All sequenced *K. pneumoniae* represented 14 STs; six were novel (Figure 6.3.). Strains exhibited considerable diversity (56947 core variable sites; mapped to reference strain MGH78578, mean call rate of 87%). H2, part of the epidemiologically-defined cluster, was distantly related to the outbreak cases (16557

SNVs from PMK1). Conversely, H1460, not identified by the clinical teams in Patan Hospital as being part of the case clusters, was part of the genetically defined outbreak (1 SNV from PMK26a), as was an adult hospital-associated strain, H30 (21 SNVs from PMK1). All outbreak isolates were ST15. Three additional smaller, closely genetically related clusters were observed; two neonatal hospital-associated clusters and a pair of community-associated infections within three weeks of each other (Figure 6.3.)

We also sequenced 47 other clinical isolates representing different species (Figures 6.4.a. and 6.4.b.); the earliest isolate was from 05/Jan/2008, and the latest from 01/Aug/2012. Within the wider hospital, several isolates causing bloodstream infection in neonates remained unsequenced, including five *K. pneumoniae*, four isolates defined as *Klebsiella* spp., two *E. coli*, seven *Acinetobacter* spp. and two *K. oxytoca*. Genotypic and phenotypic multi-drug resistance was very common in both hospital- and community-associated isolates: 13/21 (62%) non-outbreak *K. pneumoniae*, 12/14 (86%) *Enterobacter cloacae*, and 6/8 (75%) *Klebsiella oxytoca* contained at least one variant of each of *aac*, bla_{TEM}, bla_{OXA}, $bla_{\text{CTX-M}}$ and *qnr* in combination. Three non-enteropathogenic *E. coli* isolates and one *E. cloacae* isolate contained $bla_{\text{NDM-1}}$; a further non-enteropathogenic *E. coli* isolate contained $bla_{\text{NDM-6}}$.

Figure 6.3. Maximum-likelihood phylogeny of all sequenced *Klebsiella pneumoniae* study isolates (de-duplicated by individual). Isolates in red represent those originally considered part of case clusters on the basis of clinical suspicion and susceptibility testing. Additional closely genetically related clusters of non-outbreak strains are inidicated with curly brackets. "CA" denotes community-associated and "HA" hospital-associated strains; the time interval listed is that spanning the isolation dates of clustered isolates. Size of circle reflects bootstrap support at nodes.

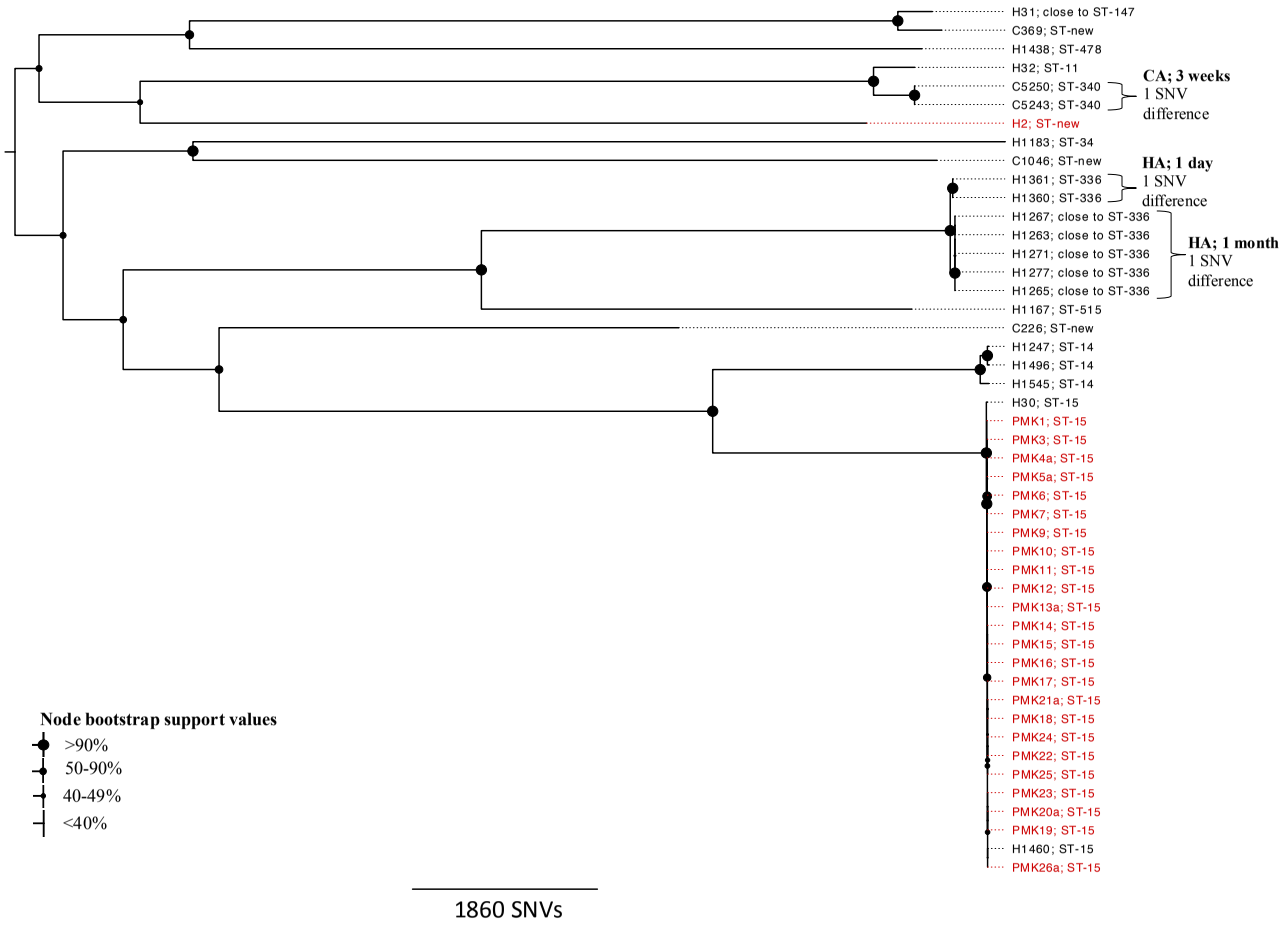


Figure 6.4.a. Resistance genotype and susceptibility phenotype by Phoenix for all sequenced *K. pneumoniae* isolates and copy number calculation for bla_{NDM}, where present. "S" susceptible, "I" intermediate, "R" = resistant – susceptibility category called by Phoenix in accordance with EUCAST breakpoints; "NA" – no result available. Antibiotic abbreviations as follows: AMP – ampicillin/amoxicillin; AMC – amoxicillin-clavulanate; TZP – piperacillin-tazobactam; CXM – cefuroxime; CRO – ceftriaxone; CAZ – ceftazidime; CIP – ciprofloxacin; ERT – ertapenem; MEM – meropenem; ATM – aztreonam; GEN – gentamicin; AMK – amikacin; SXT – trimethoprim-sulfamethoxazole; CST – colistin.

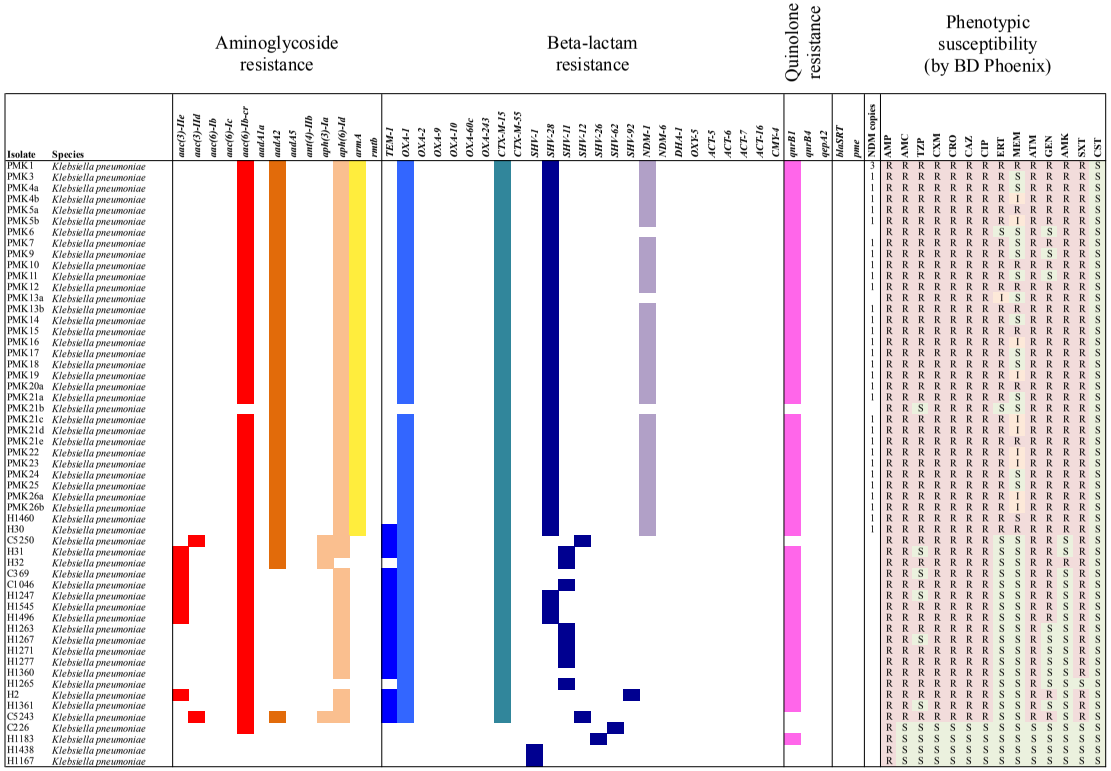
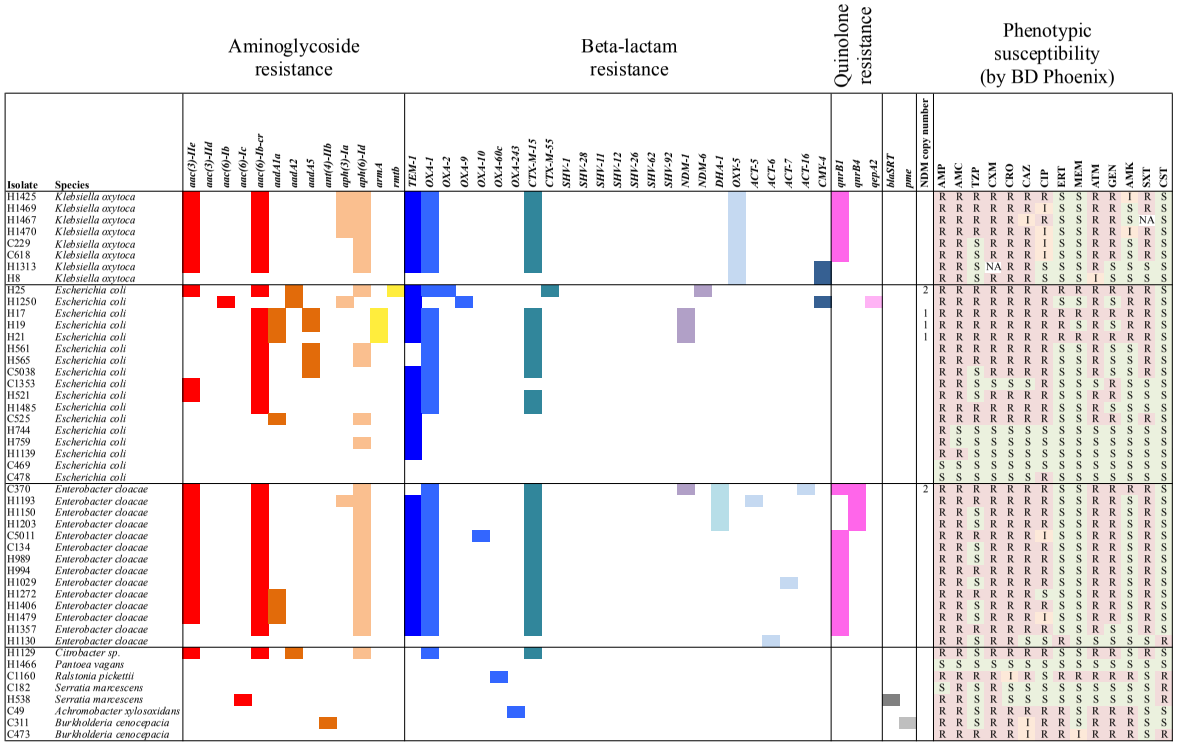


Figure 6.4.b. Resistance genotype and susceptibility phenotype by Phoenix for all sequenced non-*K. pneumoniae* isolates and copy number calculation for *bla*$_{NDM}$, where present. "S" susceptible, "I" intermediate, "R" = resistant – susceptibility category called by Phoenix in accordance with EUCAST breakpoints; "NA" – no result available. Antibiotic abbreviations as follows: AMP – ampicillin/amoxicillin; AMC – amoxicillin-clavulanate; TZP – piperacillin-tazobactam; CXM – cefuroxime; CRO – ceftriaxone; CAZ – ceftazidime; CIP – ciprofloxacin; ERT – ertapenem; MEM – meropenem; ATM – aztreonam; GEN – gentamicin; AMK – amikacin; SXT – trimethoprim-sulfamethoxazole; CST – colistin.



6.3.1. Detailed outbreak strain analysis

Fifty-three high-confidence chromosomal SNVs were identified during the outbreak (Figures 6.5.a. and 6.5.b.), 21 uniquely in the adult H30 strain. Sixteen SNVs emerged and persisted in more than one isolate, fourteen in coding sequences. Of these, eight resulted in non-synonymous mutations (Figure 6.5.a.). There was a large 121366bp deletion in PMK13b (reference positions: 3047928-3169294).

From the time-scaled phylogeny (6.6.), the outbreak strain's mutation rate was estimated at 3.65x10-6 (95% credibility intervals [CI]: 2.45x10-6 to 4.89x10-6) mutations per called site per year, equating to 18.4 (95% CI: 12.3–24.6) mutations per genome per year. The time-to-most-recent-common-ancestor of the adult-associated H30 strain and the index neonatal case PMK1 was estimated at between 1-7 months before the neonatal outbreak was observed.

The transmission network inferred by Outbreaker (Figure 6.7.) demonstrates uncertainty around the specific transmission links for early strains (PMK3-9). Four individuals harboring these strains shared ward space and time (PMK3-5, PMK7), and the network is consistent with direct transmissions. PMK9 may have contaminated the ward or equipment in the established septic nursery, or colonized an unsampled asymptomatic host, leading to the infections represented by PMK10, and possibly also to PMK11 (although this was also consistent with a direct transmission from PMK10), and later spread to the individual harboring PMK14. The link between PMK11 and PMK12 again most likely represents an indirect transmission through either an environmental or asymptomatic source, and between PMK11 and PMK13 an event across wards. Based on the available epidemiologic data (Figure 6.1.), the

Figure 6.5.a. Nucleotide and amino acid level variation for outbreak isolates with respect to the PMK1 reference chromosome, pPMK1-B and pPMK1-NDM. "del" represents a deletion. "NA" denotes an un-annotated hypothetical protein, "-" a stop codon. The order of the isolates approximates the outbreak timeframe, but also accommodates the grouping of isolates that share identical genetic sequence based on mapping to the PMK1 chromosomal reference (i.e. 0 SNV differences; shown in boxes with black borders). Brown-shaded chromosomal positions represent non-coding positions, gray-shaded positions represent positions in plasmids. Non-synonymous, persistent mutations occurred in *ppx* (an exopolyphosphatase), *azlC* (an azaleucine resistance protein), *adaA* (a methyltransferase), *gcd* (a quinoprotein glucose dehydrogenase), *bmR* (a transcriptional repressor), *oprD* (an outer membrane porin), *ssuC* (an alkane sulfonate transporter subunit), and *virB4* (a type IV secretion system).

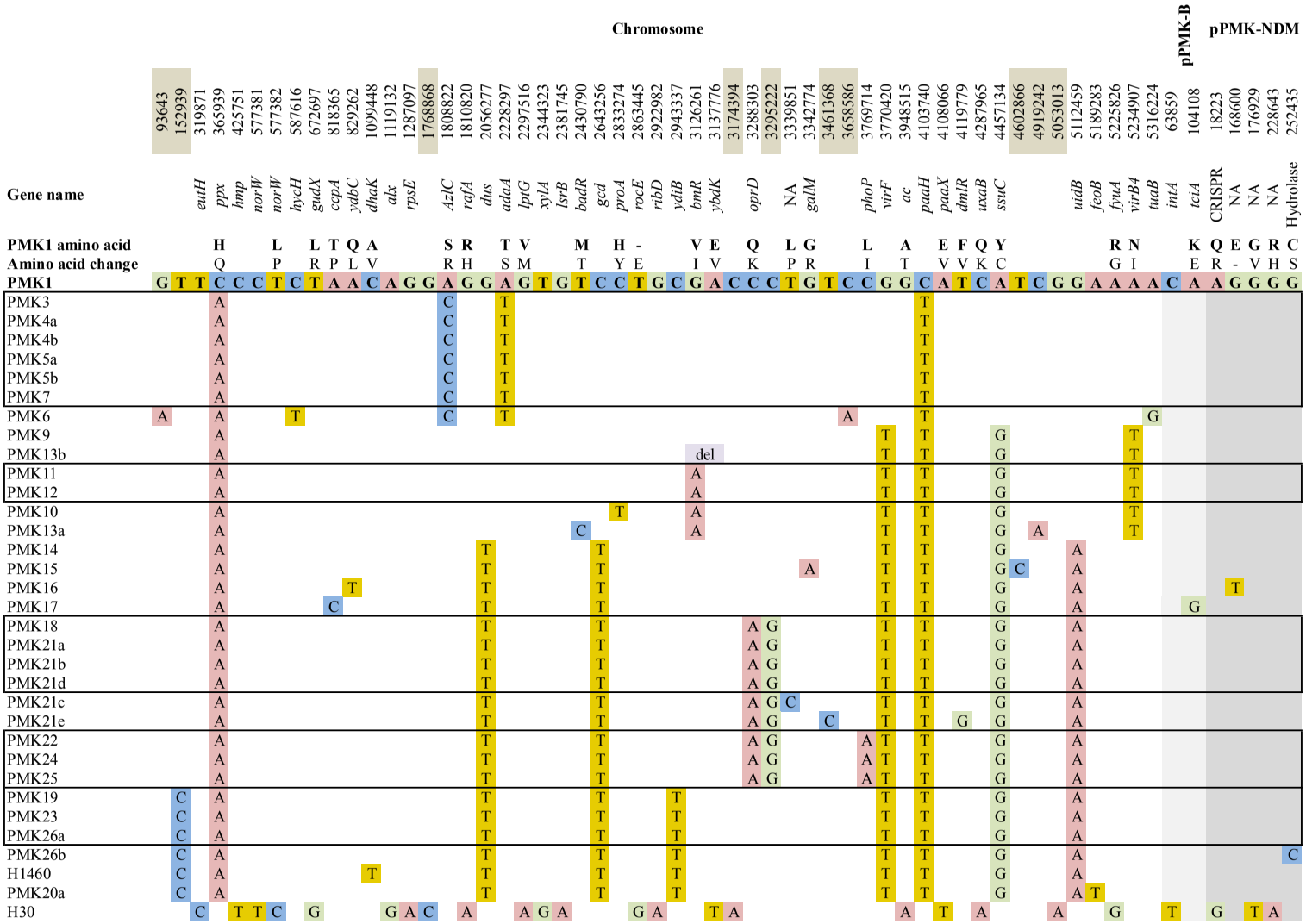


Figure 6.5.b. Phylogenetic tree summarizing chromosomal genetic relationships between all outbreak isolates. Colored nodes represent sampled isolates, black nodes unsampled intermediates; coloring reflects the sampling date in days after the first sample. Each solid branch represents a single SNV, with branch colors indicating mutation types: orange – non-synonymous, green – synonymous, gray – intergenic; gray, dashed line – 121,366bp deletion; black, dashed line – 21 SNVs.

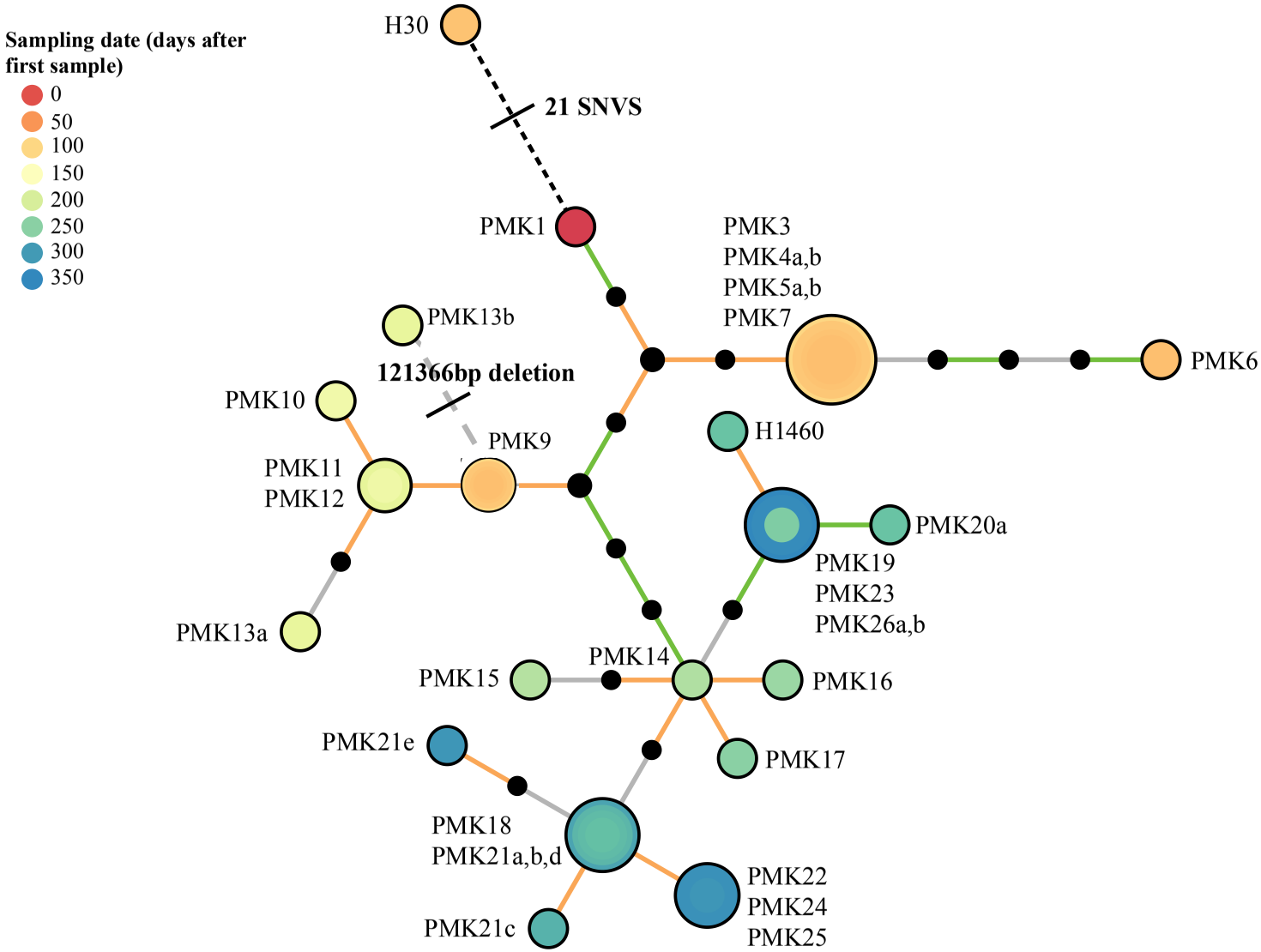


Figure 6.6. The time-scaled phylogeny inferred in BEAST. Coloured isolates represent longitudinally sampled isolates from the same individual/colour. Blue bars around the node represent the 95% credibility interval around the node height, and in this time-scaled context, the uncertainty around the time-to-most-recent-common-ancestor (TMRCA). Starred nodes have posterior support values >98%.

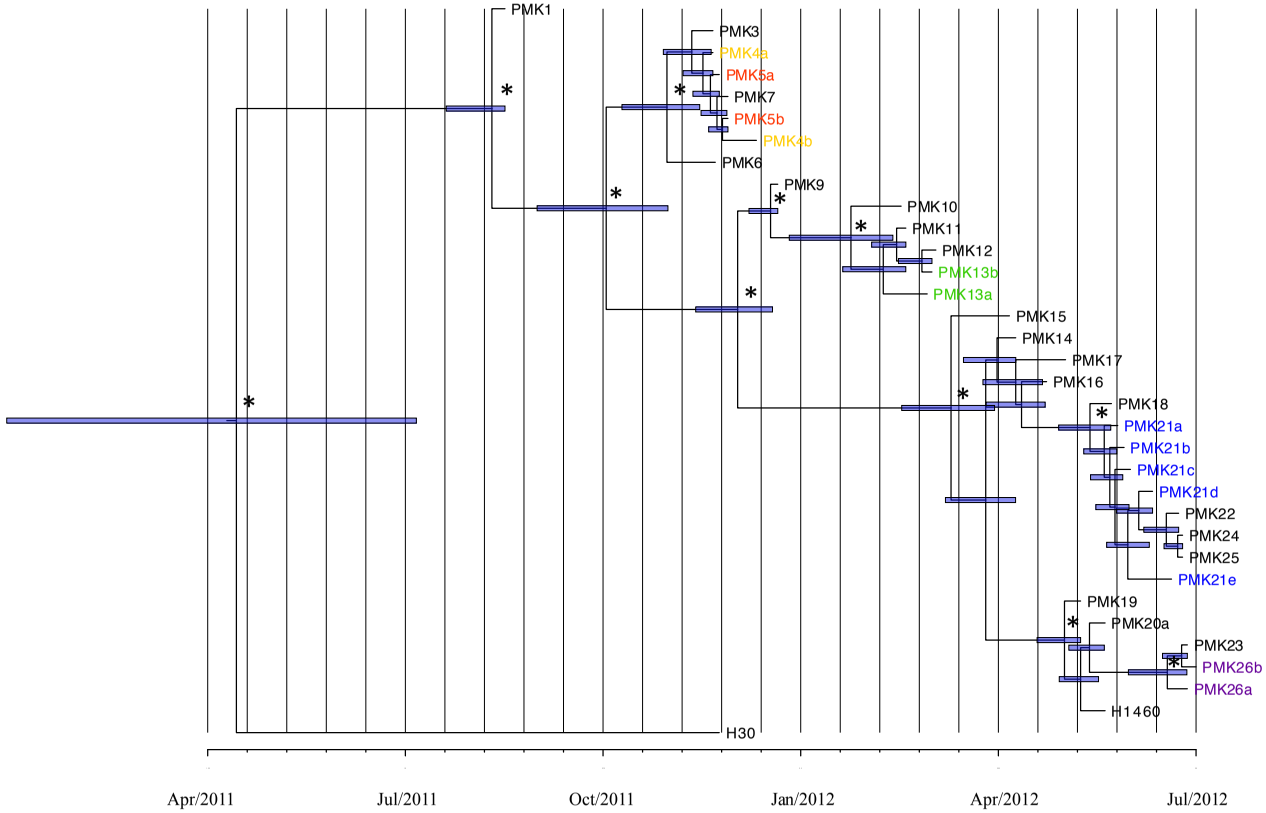
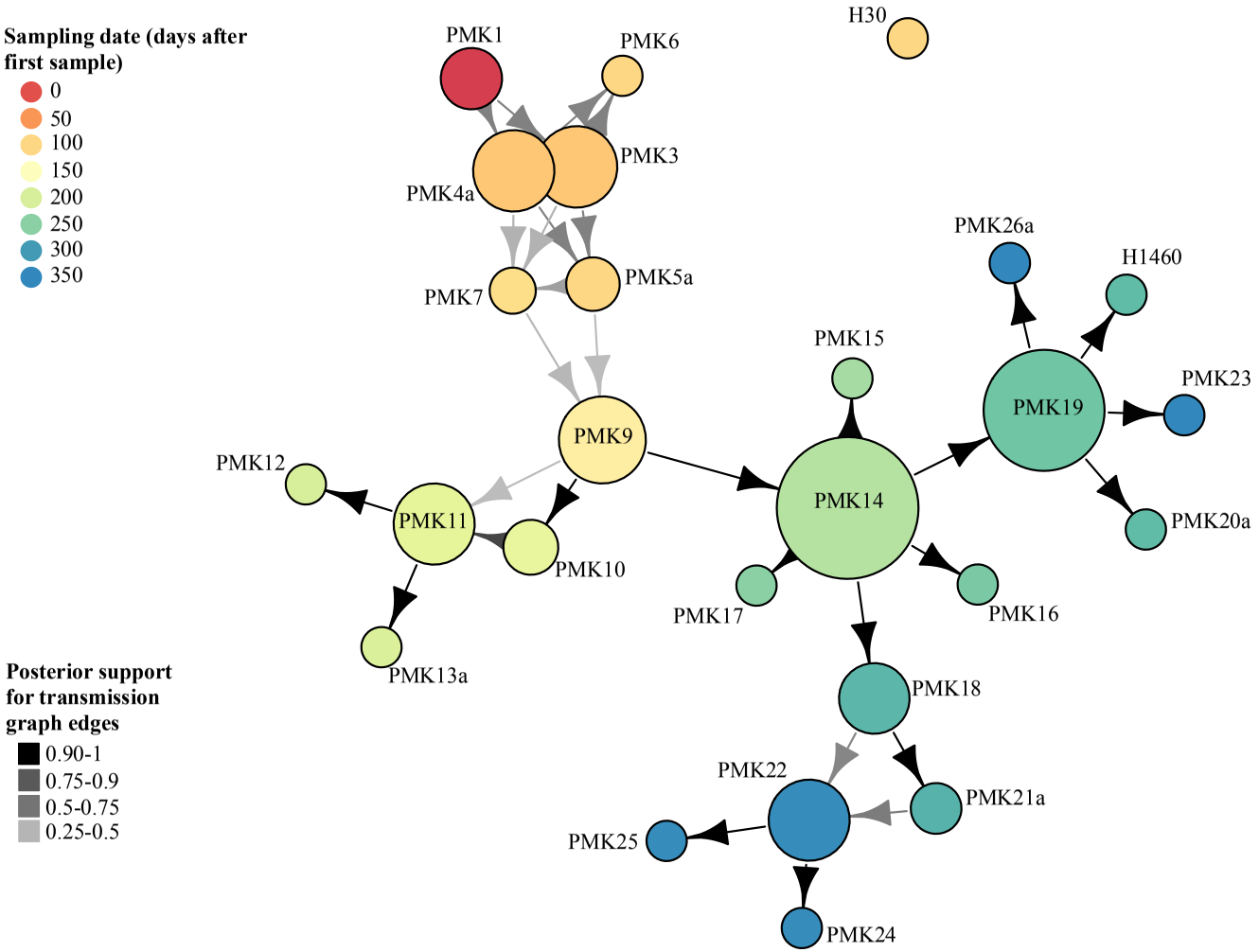


Figure 6.7. Transmission network inferred by Outbreaker. Nodes are coloured by sampling date, and sized according to the number of secondary cases they are inferred to have caused through onward transmission. Estimates of support for the certainty of transmission pathways are represented by the grey-scale shading of the graph edges.



spread from PMK14 most likely occurred within the established septic nursery (possibly directly to PMK15, otherwise indirectly via equipment/colonized asymptomatics). The links between PMK18, PMK21a, PMK22, PMK24 and PMK25 could potentially have been established through sequential transfer from the established septic nursery to the NICU and then to the overflow "septic" nursery. PMK19 to PMK20 may represent a direct transmission event, but the nature of the link between PMK19 and cases PMK23 and PMK26a is less evident. The model identifies PMK14 and PMK19 as contributing to the largest number of secondary outbreak cases, with H30 excluded from the inferred transmission network. I had limited epidemiological data on cases, so it was impossible to identify any features that may have been associated with the inferred increased contribution of PMK14 and PMK19 to transmission.

Four complete plasmid sequences were identified in PMK1. The $bla_{\text{NDM-1}}$-containing plasmid, pPMK1-NDM (304526bp), was a multi-replicon (IncHI1B/IncFIB) plasmid with antibiotic resistance determinants including *aac(6')-Ib-cr, aadA2*, $bla_{\text{CTX-M-15}}$, $bla_{\text{OXA-1}}$, *folP*, *catA1*, *dfrA12*, *armA*, and a large conjugative transfer module. A number of mercury resistance (*mer*) genes were present (Figure 6.8.).

pPMK1-A (187571bp), also a multi-replicon (IncFIIK/IncFIBK) and likely conjugative plasmid, contained tetracycline resistance genes *tetA* and *tetR*, and iron (*fec*), arsenic (*ars*), copper (*cop*), tellurite and silver (*sil*) resistance gene cassettes. pPMK1-B (111693 bp) was a colE1-type, IncFIB plasmid, containing a tellurite resistance gene, but lacking in obvious conjugative transfer genes. pPMK1-C (69947 bp) contained *aph(6)-Id* and *aph(3')-Ib*-like resistance genes encoding for streptomycin and kanamycin/neomycin resistance respectively (Figures 6.9.-6.11.).

Figure 6.8. pPMK1-NDM, the NDM-carrying plasmid from isolate PMK1, the observed index infected case in the outbreak.

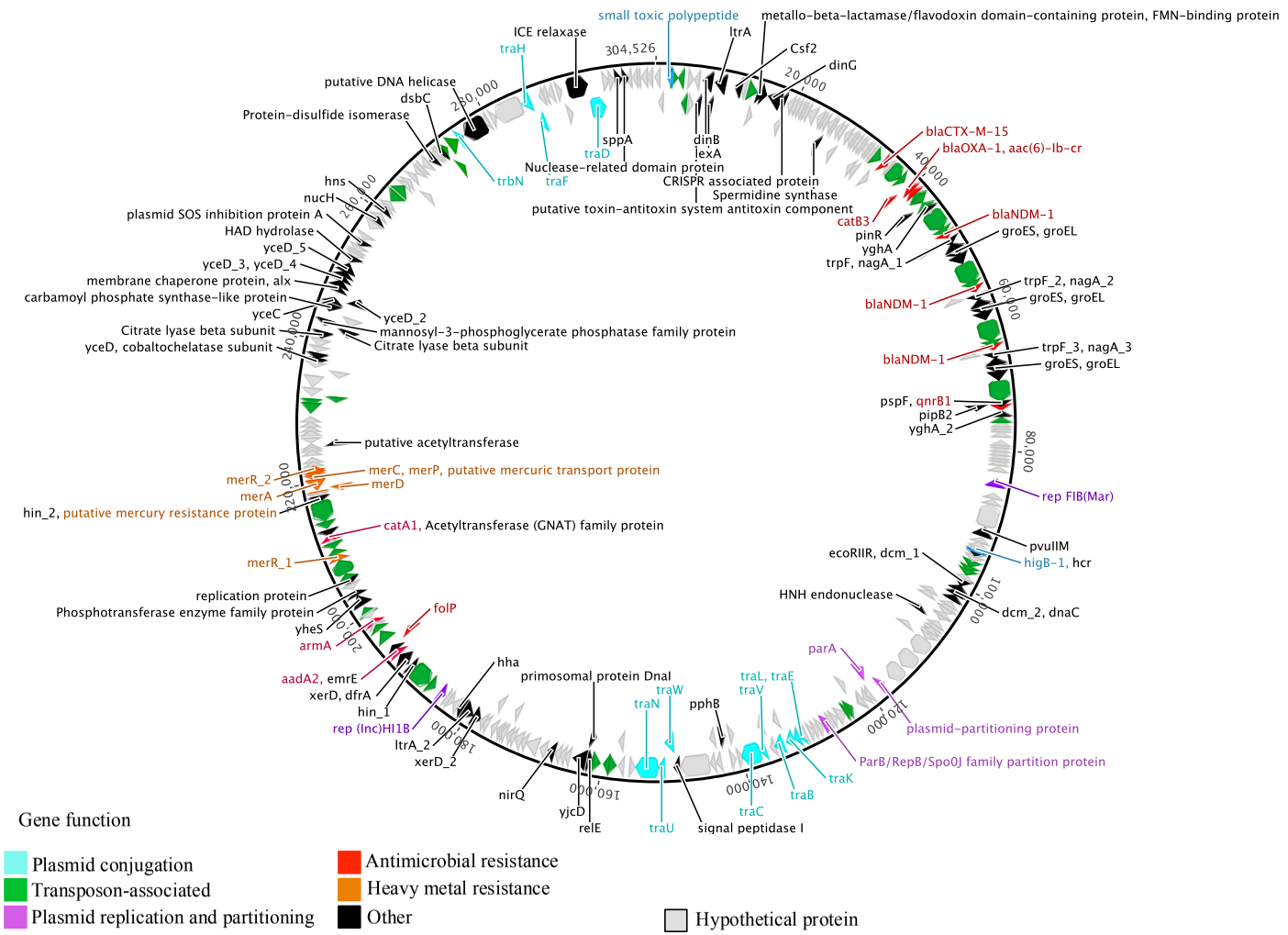


Figure 6.9. pPMK1-A, the second plasmid from isolate PMK1, the observed index infected case in the outbreak.

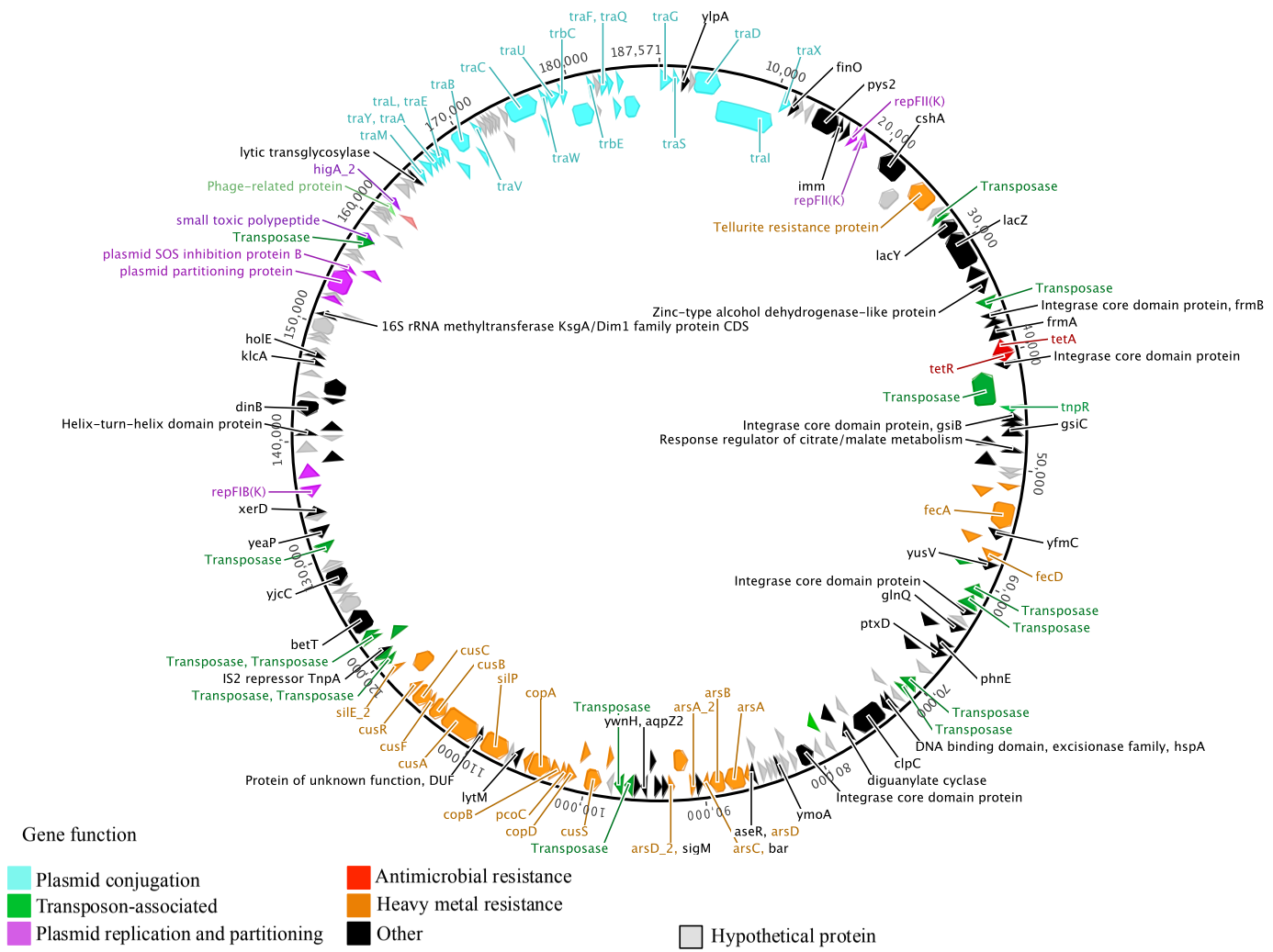


Figure 6.10. pPMK1-B, the third plasmid from isolate PMK1, the observed index infected case in the outbreak.

pPMK1-B (111693bp)

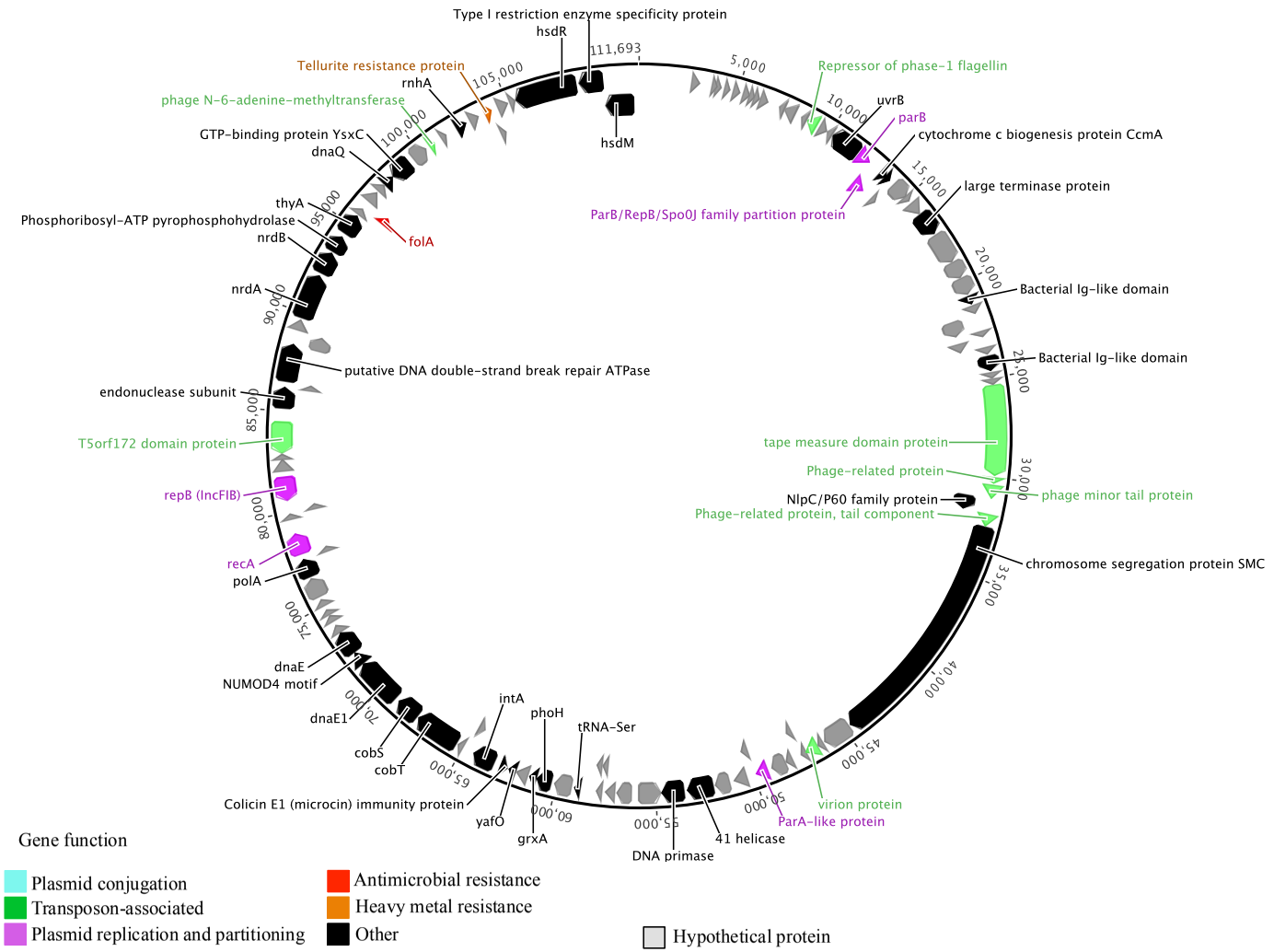
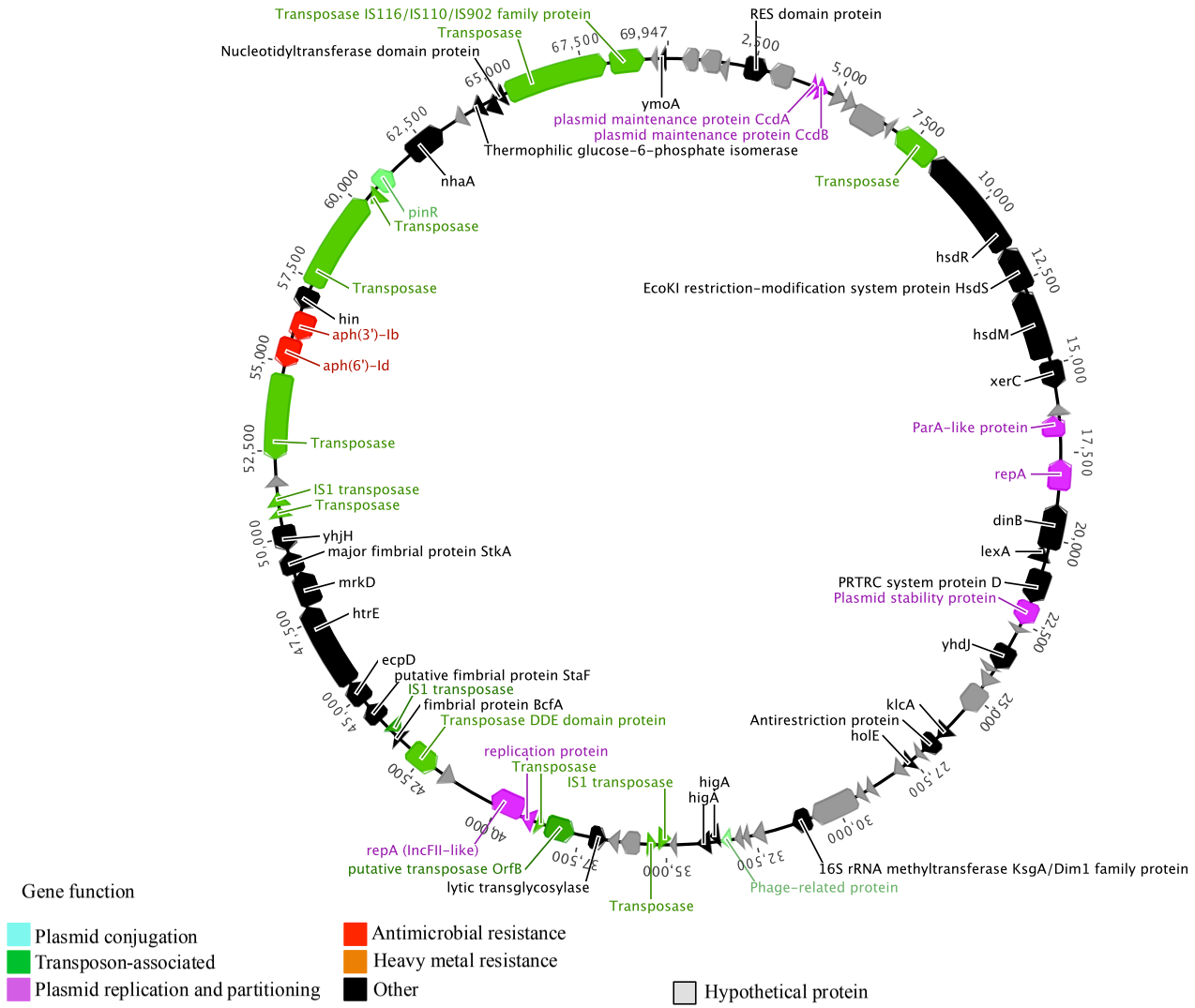


Figure 6.11. pPMK1-A, the second plasmid from isolate PMK1, the observed index infected case in the outbreak.

pPMK1-C (69947bp)



All four plasmids were highly conserved across the outbreak (Figure 6.12.). There were just five SNVs in the NDM-containing plasmid, and two in pPMK-B, three and one respectively only occurring in H30. No SNV-level variation was observed for pPMK-A or pPMK-C (Figure 6.5.a. above). Plasmid pPMK1-NDM contained two regions not present in other outbreak strains. The first was a tandem duplication of the $bla_{\text{NDM-1}}$ region (positions:56-76kb) resulting in three $bla_{\text{NDM-1}}$copies; the second an acquisition of two transposases (positions:272471-275170bp) in the plasmid structure. pPMK6-NDM and pPMK13a-NDM shared a 26kb deletion (positions:454501-70264bp), including $bla_{\text{NDM-1}}$. pPMK21b-NDM contained a larger 83kb deletion (positions:13135-96114bp), involving $bla_{\text{NDM-1}}$and other antibiotic resistance genes including *aac(6')-Ib-cr*, $bla_{\text{OXA-1}}$ and *catB3*. $bla_{\text{NDM-1}}$ gene deletions correlated with reversion to ertapenem susceptibility. pPMK24-NDM contained a small 2401bp deletion representing three phage-related open reading frames. The only other variation in plasmid gene/presence was a deletion in positions 36552-40801bp of pPMK17-A, involving tetracycline resistance genes (*tetA/tetR*).

Comparing plasmid structures across non-outbreak *Klebsiella* strains and non- *K. pneumoniae*, outbreak plasmids were essentially restricted to the outbreak *Klebsiella* strains (Figure 6.12.). Partial exceptions were: (i) pPMK1-B, large tracts of which were also found in a community-associated *K. pneumoniae* strain (C226); and (ii) pPMK1-A, regions of which were found in several nosocomial and *K. pneumoniae* and *K. oxytoca* strains. pPMK1-NDM however was not observed in any non-outbreak strains, and bla_{NDM} in the five other NDM-positive isolates was located in non-pPMK1-NDM genetic backgrounds (four *E. coli*: H17, H19, H21 [closely genetically related; data not shown], H25; one *E. cloacae*: C370).

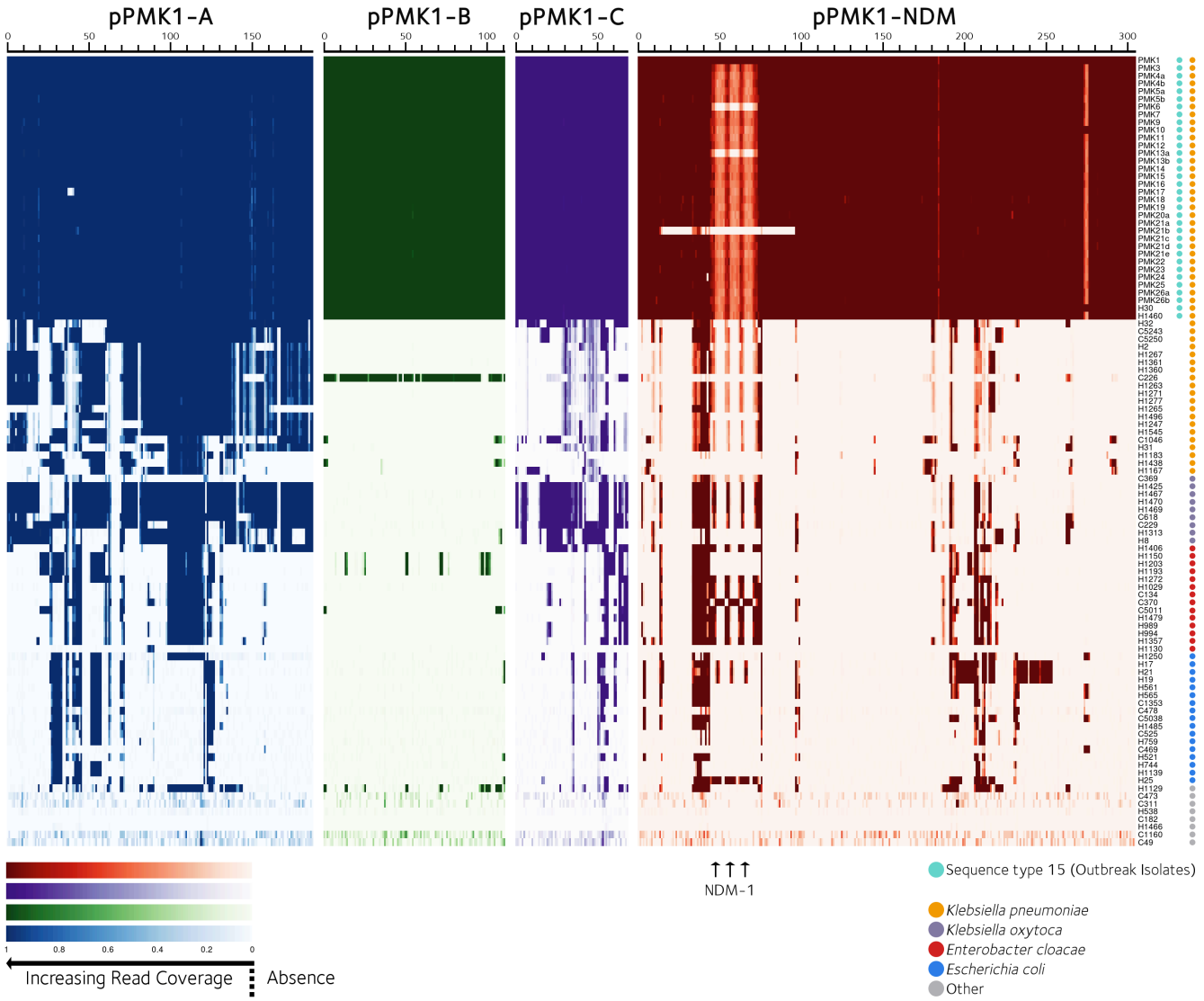


Figure 6.12. Mapping coverage of plasmid references in outbreak and non-outbreak strains. Plasmid reference coordinates in are displayed in kilobases. Heatmap values represent the number of bases mapped, scaled by the number of bases mapped to the whole reference. Values of zero represent absence. Values of one represent presence at the average coverage or greater. Values between zero and one represent presence at a lower than average coverage.

6.4. DISCUSSION

Despite time lapses between isolates and initial uncertainty as to whether all cases were linked, WGS clearly demonstrated that the outbreak was caused by a single, clonal strain of *K. pneumoniae* ST15 in association with a highly conserved population of four plasmids, including a $bla_{\text{NDM-1}}$-containing plasmid. The intervals between case clusters suggest persistence of the NDM-1-*K. pneumoniae* strain in the unit environment or in asymptomatic carriage, potentially supported by the isolation of *Klebsiella* spp. from environmental samples/rectal swabs taken from staff. Both environmental contamination and asymptomatic colonization are likely contributors to the transmission of drug-resistant *K. pneumoniae* in the nosocomial setting(13), although the relative contribution of each route is unknown, and could feasibly vary amongst lineages. The assessment of combined epidemiologic data and the transmission network inferred from sampling dates and the genomic data in this study strongly supports the view that even wider sampling frames are needed to fully understand the dynamics of these outbreaks, given that both direct and indirect human and environmental transmissions are likely to be occurring. The time-scaled analysis of the outbreak isolates and H30 suggests an ancestral strain predating PMK1 and H30 was present somewhere between 1-7 months before the identification of the first, infected, neonatal case – this ancestral strain may have been present in the parents, other patients, hospital staff or the hospital environment.

Whilst it is impossible to exclude repeated introductions of either the strain or the $bla_{\text{NDM-1}}$-containing plasmid into the pediatric critical care setting from the community or elsewhere in the hospital, the extraordinary degree of similarity between outbreak strains in contrast to other contemporaneous strains from both

locations makes this less likely. Regarding the selection of the wider set of isolates for sequencing, I specifically avoided characterizing only phenotypically carbapenem-resistant organisms given the known lack of sensitivity of phenotypic methods in the presence of carbapenemases(28), aiming to determine whether there was any evidence for wider dissemination of the outbreak plasmids or the outbreak *K. pneumoniae* strain in the absence of bla_{NDM}. The wider sampling and detailed plasmid analysis is a major strength of this study and expands on the two previous WGS outbreak investigations(13, 14); without it, uncertainty about transmission versus repeated importations of strains or dissemination of drug resistance plasmids remains, particularly in high-prevalence contexts.

This study exemplifies the potential of using WGS to benefit both outbreak management and antimicrobial treatment. In particular, long-read (30kb) PacBio sequencing enabled the production of reference assemblies of the isolate cultured from the first, infected, neonatal case, including both chromosome and plasmids, which were then used as a comparator for Illumina short-read datasets for the other isolates, allowing fine-scale definition of genetic differences between strains and probabilistic interpretation of likely transmission pathways. As well as resolving the temporally distinct clusters of NDM-1-*K. pneumoniae* into a single year-long outbreak, phylogenetic analysis identified several other unrecognized clusters of antimicrobial-resistant isolates indicating both nosocomial (two separate CTX-M-15-*K. pneumoniae* outbreaks; one NDM-1-*E. coli* outbreak; two *K. oxytoca*, and three *E. cloacae* clusters [data not shown]) and community transmission (CTX-M-15-*K. pneumoniae*).

Laboratory susceptibility phenotyping SIR categories were inconsistent in some cases between isolates with the same complement of resistance genes, even for the highly genetically related outbreak strains, highlighting the challenges of relying on this for cluster identification (Figure 6.3.a. and 6.3.b.; observed for meropenem in the presence of NDM-1, and gentamicin in the presence of *armA*). Although sequence-based susceptibility prediction has yet to be correlated with patient-level clinical outcomes, it appears sensitive and specific as shown in Chapter 5(29, 30), and potentially indicates the future value of WGS in managing patients when current routine laboratory turnaround times are matched. Single strand sequencing platforms and new, fast genome assemblers(31) are likely to make resistance prediction from clinical specimens, such as blood cultures, possible within hours of sampling, providing that the currently high error-rates can be reduced.

This NDM-*K. pneumoniae* outbreak terminated after a year, and hospital-wide neonatal critical care deaths almost halved after it ended, consistent with its suppression or eradication. This is of interest for two reasons: Firstly, bla_{NDM} is known to be locally prevalent, and was found in 5/68 (7%) of sequenced, non-outbreak clinical isolates in this study; and secondly, the hospital is of older construction and potentially less amenable to infection control. Carbapenemase-associated *K. pneumoniae* outbreaks have been shown to be difficult to control in some settings, even those where resources are less restricted(4). More recently, a separate NDM-1-*Enterobacter cloacae* outbreak was observed in the hospital, albeit in association with a different plasmid vector (data not shown), and was terminated following fumigation of a number of the clinical units, demonstrating the need for on-going surveillance and comprehensive infection control.

K. pneumoniae ST15 is one of the dominant global clones, associated with a range of beta-lactamases, including NDM/CTX-M-15(32, 33). Its success may partly relate to accumulation of resistance without fitness costs(34). The outbreak strain stably supported several plasmids totaling nearly 700kb, 13% of the chromosomal size. Outbreak plasmids contained resistance genes, plasmid addiction modules, and genes encoding for systems protecting host bacteria against bacteriophages/plasmids, which may have contributed to the strain's dominance and stable persistence.

pPMK1-NDM showed substantial homology to pNDM-MAR (JN420336.1), first identified from ST15-*K. pneumoniae* isolates in Morocco, but differed from other NDM-containing plasmids(35). The major differences were the presence of additional NDM-1 copies and further resistance genes in pPMK1-NDM (Figure 6.8). Carbapenem exposure affects NDM-1 copy number, and lack of selection pressure can lead to complete gene deletion(36), providing further evidence to support minimizing unnecessary carbapenem therapy.

The mutation rate estimate for the outbreak *K. pneumoniae* strain is higher than for other *Enterobacteriaceae*, such as *E. coli* (~1.1/genome/year)(37). There are no published data on *K. pneumoniae* mutation rates to my knowledge; it is therefore impossible to ascertain whether this strain was adaptable because of hyper-mutability, or whether this is a species-level phenomenon. Mutation rates define the plausible timeframe of acquisition events in bacteria, and are critical now genomic data are increasingly being relied upon to refine transmission epidemiology. Some SNVs were intergenic, highlighting the potential additional resolution achieved using the

complete mapped genome rather than extended multi-locus sequence typing (sometimes known as "super-MLST") for transmission analyses(38).

There are several limitations to this study. NDM-1 copy number variation may have been due to variable selection pressure in strain culture and storage. Selective culture would have aided environmental detection of NDM-containing isolates, and the identification of an environmental source or presence of a susceptible variant of the outbreak clone was unfortunately impossible given the lack of further susceptibility testing on and/or storage of environmental/fecal carriage isolates for sequencing. Routine surveillance of patients admitted to at-risk units would have been ideal, but was not possible given the local resource constraints. It was not entirely clear what method had been used to identify the wider set of non-outbreak isolates – it may have ben that this did not represent a truly random sample. The wider set of clinical isolates also did not contain other ST15-*K. pneumoniae*; these were either present at low frequencies and thus unsampled, or not present at all. Despite the wider sampling, it was still impossible to identify the source of the outbreak, a limitation which is likely to be overcome if it becomes possible to implement WGS as a high-resolution, real-time typing method and use it to refine and extend ongoing outbreak investigations.

In summary, this study highlights challenges in managing clusters of resistant *K. pneumoniae* and demonstrates that, despite lack of obvious transmission pathways, the same strain can persist in hospital environments for months with a highly stable and conserved set of plasmids, sporadically causing disease. Whether this is a consequence of particular combinations of host bacterial strains, and NDM-containing and/or other plasmids, representing clinically successful entities with a significant

impact on patient outcome, is unclear. However, WGS provides a high-resolution mechanism by which the contribution of these different elements can start to be unravelled.

CHAPTER 6 REFERENCES

1. **Downie L, Armiento R, Subhi R, Kelly J, Clifford V, Duke T.** 2013. Community-acquired neonatal and infant sepsis in developing countries: efficacy of WHO's currently recommended antibiotics--systematic review and meta-analysis. Archives of disease in childhood **98:**146-154.
2. **Gupta N, Limbago BM, Patel JB, Kallen AJ.** 2011. Carbapenem-resistant Enterobacteriaceae: epidemiology and prevention. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America **53:**60-67.
3. **Giske CG, Monnet DL, Cars O, Carmeli Y, ReAct-Action on Antibiotic R.** 2008. Clinical and economic impact of common multidrug-resistant Gram-negative bacilli. Antimicrobial agents and chemotherapy **52:**813-821.
4. **Tofteland S, Naseer U, Lislevand JH, Sundsfjord A, Samuelsen O.** 2013. A long-term low-frequency hospital outbreak of KPC-producing Klebsiella pneumoniae involving Intergenus plasmid diffusion and a persisting environmental reservoir. PloS one **8:**e59015.
5. **Munoz-Price LS, Quinn JP.** 2013. Deconstructing the infection control bundles for the containment of carbapenem-resistant Enterobacteriaceae. Current opinion in infectious diseases **26:**378-387.
6. **Yong D, Toleman MA, Giske CG, Cho HS, Sundman K, Lee K, Walsh TR.** 2009. Characterization of a new metallo-beta-lactamase gene, bla(NDM-1), and a novel erythromycin esterase gene carried on a unique genetic structure in Klebsiella pneumoniae sequence type 14 from India. Antimicrobial agents and chemotherapy **53:**5046-5054.

7. **Jacoby GB, K.** 2012, posting date. β-Lactamase Classification and Amino Acid Sequences for TEM, SHV and OXA Extended-spectrum and Inhibitor-resistant Enzymes. [Online.]

8. **Johnson AP, Woodford N.** 2013. Global spread of antibiotic resistance: the example of New Delhi metallo-beta-lactamase (NDM)-mediated carbapenem resistance. Journal of medical microbiology **62:**499-513.

9. **Perry JD, Naqvi SH, Mirza IA, Alizai SA, Hussain A, Ghirardi S, Orenga S, Wilkinson K, Woodford N, Zhang J, Livermore DM, Abbasi SA, Raza MW.** 2011. Prevalence of faecal carriage of Enterobacteriaceae with NDM-1 carbapenemase at military hospitals in Pakistan, and evaluation of two chromogenic media. The Journal of antimicrobial chemotherapy **66:**2288-2294.

10. **Walsh TR, Weeks J, Livermore DM, Toleman MA.** 2011. Dissemination of NDM-1 positive bacteria in the New Delhi environment and its implications for human health: an environmental point prevalence study. The Lancet infectious diseases **11:**355-362.

11. **Eyre DW, Cule ML, Wilson DJ, Griffiths D, Vaughan A, O'Connor L, Ip CL, Golubchik T, Batty EM, Finney JM, Wyllie DH, Didelot X, Piazza P, Bowden R, Dingle KE, Harding RM, Crook DW, Wilcox MH, Peto TE, Walker AS.** 2013. Diverse sources of C. difficile infection identified on whole-genome sequencing. The New England journal of medicine **369:**1195-1205.

12. **Walker TM, Ip CL, Harrell RH, Evans JT, Kapatai G, Dedicoat MJ, Eyre DW, Wilson DJ, Hawkey PM, Crook DW, Parkhill J, Harris D, Walker AS, Bowden R, Monk P, Smith EG, Peto TE.** 2013. Whole-genome

sequencing to delineate Mycobacterium tuberculosis outbreaks: a retrospective observational study. The Lancet infectious diseases **13:**137-146.

13. **Snitkin ES, Zelazny AM, Thomas PJ, Stock F, Group NCSP, Henderson DK, Palmore TN, Segre JA.** 2012. Tracking a hospital outbreak of carbapenem-resistant Klebsiella pneumoniae with whole-genome sequencing. Science translational medicine **4:**148ra116.

14. **Epson EE, Pisney LM, Wendt JM, MacCannell DR, Janelle SJ, Kitchel B, Rasheed JK, Limbago BM, Gould CV, Kallen AJ, Barron MA, Bamberg WM.** 2014. Carbapenem-resistant Klebsiella pneumoniae producing New Delhi metallo-beta-lactamase at an acute care hospital, Colorado, 2012. Infection control and hospital epidemiology : the official journal of the Society of Hospital Epidemiologists of America **35:**390-397.

15. **Standards NCfCL.** 2001. Performance Standards for Antimicrobial Disk Susceptibility Tests.

16. **Wood DE, Salzberg SL.** 2014. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome biology **15:**R46.

17. **Diancourt L, Passet V, Verhoef J, Grimont PA, Brisse S.** 2005. Multilocus sequence typing of Klebsiella pneumoniae nosocomial isolates. Journal of clinical microbiology **43:**4178-4182.

18. **Guindon S, Delsuc F, Dufayard JF, Gascuel O.** 2009. Estimating maximum likelihood phylogenies with PhyML. Methods in molecular biology **537:**113-137.

19. **Chin CS, 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. Nature methods **10:**563-569.

20. **Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ.** 1990. Basic local alignment search tool. Journal of molecular biology **215:**403-410.

21. **Delcher AL, Phillippy A, Carlton J, Salzberg SL.** 2002. Fast algorithms for large-scale genome alignment and comparison. Nucleic acids research **30:**2478-2483.

22. **Li H, Durbin R.** 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics **25:**1754-1760.

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

24. **Drummond AJ, Rambaut A.** 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC evolutionary biology **7:**214.

25. **Jombart T, Cori A, Didelot X, Cauchemez S, Fraser C, Ferguson N.** 2014. Bayesian reconstruction of disease outbreaks by combining epidemiologic and genomic data. PLoS computational biology **10:**e1003457.

26. **Poirel L, Herve V, Hombrouck-Alet C, Nordmann P.** 2011. Long-term carriage of NDM-1-producing Escherichia coli. The Journal of antimicrobial chemotherapy **66:**2185-2186.

27. **D'Andrea MM, Venturelli C, Giani T, Arena F, Conte V, Bresciani P, Rumpianesi F, Pantosti A, Narni F, Rossolini GM.** 2011. Persistent carriage and infection by multidrug-resistant Escherichia coli ST405 producing NDM-1 carbapenemase: report on the first Italian cases. Journal of clinical microbiology **49:**2755-2758.

28. **Doyle D, Peirano G, Lascols C, Lloyd T, Church DL, Pitout JD.** 2012. Laboratory detection of Enterobacteriaceae that produce carbapenemases. Journal of clinical microbiology **50:**3877-3880.

29. **Gordon NC, Price JR, Cole K, Everitt R, Morgan M, Finney J, Kearns AM, Pichon B, Young B, Wilson DJ, Llewelyn MJ, Paul J, Peto TE, Crook DW, Walker AS, Golubchik T.** 2014. Prediction of Staphylococcus aureus antimicrobial resistance by whole-genome sequencing. Journal of clinical microbiology **52:**1182-1191.

30. **Stoesser N, Batty EM, Eyre DW, Morgan M, Wyllie DH, Del Ojo Elias C, Johnson JR, Walker AS, Peto TE, Crook DW.** 2013. Predicting antimicrobial susceptibilities for Escherichia coli and Klebsiella pneumoniae isolates using whole genomic sequence data. The Journal of antimicrobial chemotherapy **68:**2234-2244.

31. **Iqbal Z, Turner I, McVean G.** 2013. High-throughput microbial population genomics using the Cortex variation assembler. Bioinformatics **29:**275-276.

32. **Peirano G, Ahmed-Bentley J, Fuller J, Rubin JE, Pitout JD.** 2014. Travel-related carbapenemase-producing Gram-negative bacteria in Alberta, Canada: the first 3 years. Journal of clinical microbiology **52:**1575-1581.

33. **Quinones D, Valverde A, Rodriguez-Banos M, Kobayashi N, Zayaz A, Abreu M, Canton R, del Campo R.** 2014. High clonal diversity in a non-outbreak situation of clinical ESBL-producing Klebsiella pneumoniae isolates in the first national surveillance program in Cuba. Microbial drug resistance **20:**45-51.

34. **Toth A, Kocsis B, Damjanova I, Kristof K, Janvari L, Paszti J, Csercsik R, Topf J, Szabo D, Hamar P, Nagy K, Fuzi M.** 2014. Fitness cost

associated with resistance to fluoroquinolones is diverse across clones of Klebsiella pneumoniae and may select for CTX-M-15 type extended-spectrum beta-lactamase. European journal of clinical microbiology & infectious diseases : official publication of the European Society of Clinical Microbiology **33:**837-843.

35. **Villa L, Poirel L, Nordmann P, Carta C, Carattoli A.** 2012. Complete sequencing of an IncH plasmid carrying the blaNDM-1, blaCTX-M-15 and qnrB1 genes. The Journal of antimicrobial chemotherapy **67:**1645-1650.

36. **Huang TW, Chen TL, Chen YT, Lauderdale TL, Liao TL, Lee YT, Chen CP, Liu YM, Lin AC, Chang YH, Wu KM, Kirby R, Lai JF, Tan MC, Siu LK, Chang CM, Fung CP, Tsai SF.** 2013. Copy Number Change of the NDM-1 sequence in a multidrug-resistant Klebsiella pneumoniae clinical isolate. PloS one **8:**e62774.

37. **Reeves PR, Liu B, Zhou Z, Li D, Guo D, Ren Y, Clabots C, Lan R, Johnson JR, Wang L.** 2011. Rates of mutation and host transmission for an Escherichia coli clone over 3 years. PloS one **6:**e26907.

38. **Maiden MC, Jansen van Rensburg MJ, Bray JE, Earle SG, Ford SA, Jolley KA, McCarthy ND.** 2013. MLST revisited: the gene-by-gene approach to bacterial genomics. Nature reviews. Microbiology **11:**728-736.

CHAPTER 7: OUTBREAK EPIDEMIOLOGY – USE OF WGS TO INVESTIGATE AN EXTENDED SERIES OF KPC-*KLEBSIELLA PNEUMONIAE* INFECTIONS IN A SINGLE INSTITUTION IN THE USA

7.1. BACKGROUND

In the US, the public health threat represented by carbapenemase-producing *Enterobacteriaceae*(1) is currently predominantly represented by the presence of *K. pneumoniae*-carbapenemase (KPC)-producing *K. pneumoniae*, which have become endemic in many US healthcare settings(2). Like bla_{NDM}, the bla_{KPC} gene commonly resides in highly mobile plasmids that can move freely between genera of *Enterobacteriaceae*(3, 4), but are thought to have become predominantly associated with *K. pneumoniae*(2), and in particular, a global lineage, sequence type ST-258(5-10) and other highly related strains in the ST-258 clonal complex(11, 12). This association has not been explained. No current evidence supports increased virulence of ST-258 KPC- *K. pneumoniae*(13, 14), though some studies suggest that either the presence of the KPC plasmid and/or specific factors associated with motility or DNA repair may be a feature in promoting its evolutionary success(15, 16).

The bla_{KPC} gene is predominantly contained within Tn*4401*, a ten kilobase (kb) Tn*3*-family transposon capable of mobilisation through transposition(17). This transposon exists as a number of described "isoforms"(18-21), and inserted in a variety of plasmids(19). Little is known about the diversity of mobile genetic elements participating in the wider dissemination of KPC-mediated carbapenem resistance among other lineages of *K. pneumoniae*. This chapter describes the use of WGS to investigate the molecular epidemiology of KPC- *K. pneumoniae* occurring in a single

hospital ecosystem over a five-year period, dating from the introduction in 2007 of both a KPC- *K. pneumoniae* and KPC-producing *Klebsiella oxytoca* by the presumed index case in the study institution(22).

7.2. METHODS

Isolates for this study had been collected from August 2007 to September 2012, as part of a larger surveillance study of carbapenemase-producing *Enterobacteriaceae* undertaken by my colleague, Dr Amy Mathers, working at the Clinical Microbiology Laboratory of the University of Virginia Health System (UVaHS). UVaHS serves: (i) a 619-bed tertiary care hospital; (ii) outpatient clinics in central Virginia, and, since August 2010, a 40-bed long-term acute care hospital (LTACH). All clinical *Enterobacteriaceae* isolates that flagged as possible extended spectrum β-lactamase (ESBL)-producers or had an ertapenem MIC of ≥1 μg/mL by automated susceptibility profiling (VITEK2, Biomérieux, Durham, NC**)** underwent carbapenemase phenotypic testing using the modified Hodge test (August 2007-June 2008) or the indirect carbapenemase test (July 2008-September 2012)(23). In addition, weekly infection control surveillance by peri-rectal swab was performed on all inpatients on units where there was a patient who was known to be colonised or infected with carbapenemase-producing *Enterobacteriaceae* (CPE)(24-26), diagnosed either in the UVaHS labs, or in other centres, in the case of patients being transferred into UVaHS. Any isolates identified as *K. pneumoniae* by the VITEK2 with a positive carbapenemase phenotypic test and/or meropenem or imipenem MIC ≥1μg/mL underwent specific testing for the bla_{KPC} gene using PCR analysis(4, 23). These isolates had previously undergone pulsed field gel electrophoresis (PFGE) and multi-

locus sequence typing (MLST), and a subset of these results had been previously reported(23).

For this work, the KPC-*K. pneumoniae* isolate for the index patient(22), who was the first KPC-positive patient identified at UVaHS, was whole genome sequenced. In addition, a further 36 KPC-positive isolates from 64 patients identified as colonised/infected were selected by a laboratory technician blinded to the clinical data. These represented the first KPC-positive isolates cultured from each of these patients, from samples taken at timepoints throughout the study period (1 in 2007, 5 in 2008, 10 in 2009, 10 in 2010, 5 in 2011 and 5 in 2012). For a smaller cohort of individuals (n=11) in whom persistent colonisation with a KPC- *K. pneumoniae* of the same sequence type (ST) was observed, the last isolate cultured from the patient was sequenced in addition to the first, in order to determine a "molecular clock", or evolutionary rate for the *K. pneumoniae* being investigated.

The date, location in the hospital and anatomical site of each sample were collected at the time of sampling. For this study, in which the genetic mechanism of carbapenem-resistance across species in the UVaHS had been previously shown to be uniformly bla_{KPC}, the initial culture with any carbapenemase-producing *Enterobacteriaceae* strain by phenotypic carbapenemase testing (modified Hodge test and later indirect carbapenemase test) was considered the epidemiologic acquisition event of bla_{KPC} by a patient. KPC-positive patients were considered to be carriers for the entire duration of their hospital stay. KPC-*K. pneumoniae* isolates were classified as "imported" if cultured from patients without prior admission to UVaHS Medical Centre/LTACH (UVaMC) and culture-positive before or within 48 hours of transfer to UVaHS. For

the remaining patients, the risk for acquiring KPC-*K. pneumoniae* at UVaHS was based on any hospitalisation at UVaHS in the previous 90 days and classified as: "high", if there was a CPE carrier on the unit simultaneously prior to isolation of a new KPC-*K. pneumoniae*; and "indeterminate", if the patient was admitted to UVaHS for ≥48 hours at the time of acquisition but had never been on the same hospital unit at the same time as a CPE-carrier. The study was approved by the University of Virginia Institutional Review Board (#13558); however, I never had access to any patient-identifiable information.

DNA was extracted and sequenced as described in the thesis methods. Reads were mapped against the reference *K. pneumoniae* sequence MGH78578 (RefSeq: NC_009653) and *de novo* assembled using the standard group's bioinformatic pipeline.

In silico MLST was performed for all isolates using the MLST scheme developed at Institut Pasteur(27). Filtered single nucleotide variants (SNVs) for each isolate were reinserted into the reference MGH78578 sequence, resulting in an alignment of 37 modified reference sequences for comparisons across the whole dataset. A maximum-likelihood tree of the complete dataset of 37 strains was constructed using RaxML version 7.7.6, implementing: (i) a generalized time-reversible nucleotide substitution model, (ii) fixing the proportion of invariant sites at 0, (iii) allowing for four relative substitution rate categories, and (iv) using the rapid bootstrapping algorithm outputting the best ML tree with 100 replicates.

The molecular clock was determined by comparing the sequence variation between the first and last isolate taken from each patient in the smaller cohort of individuals in whom longitudinal colonization had been observed. The clock was calculated using the time interval between the two samplings, and applying a Bayesian model to estimate the evolutionary rate(28).

Additional analyses were undertaken for the two largest sub-groups identified by *in silico* MLST, using alignments of 13 sequences for the ST941 group and seven sequences for the ST258 group, respectively. For these two distinct STs, within-ST comparisons of evolutionary relationships were made using ClonalFrame (version 1.2)(29) to account for the impact of recombination, particularly segments which may have been imported from more divergent lineages. Three separate runs were performed for both the ST941 and the ST258 clusters, on the alignments of modified reference strains (with SNVs reinserted). The ClonalFrame settings were as follows:

- Number of MCMC iterations following the burn-in period = 2000
- Number of burn-in iterations = 2000
- Number of iterations between recording parameter values in the posterior sample (the thinning interval) = 2

Convergence of the runs was assessed by comparing the similarity of the run outputs, with reference to the global parameters (θ [the mutational rate, assumed to be constant on the branches of the topology]; R [the recombination rate, assumed to be constant on the branches of the topology], ν [the rate of new polymorphism introduced by recombination], and δ [the average tract length of a recombination event]) and the phylogeny. As a statistical measure of convergence, the Gelman & Rubin test statistic was calculated for the global parameters for pairs of runs(30). Consensus trees, based

on the union of posterior samples for each of the triplicate datasets, were determined. The molecular clock estimate derived from the paired, longitudinal samples was used to scale the ClonalFrame phylogenies to estimate dating and dating intervals around the time to most recent common ancestors (TMRCA) for groups of isolates.

Transformants were generated using plasmid DNA from a KPC-*K. pneumoniae* and a KPC-producing *K. oxytoca* (both isolated from the presumed index patient (30)) with subsequent electroporation into *E. coli* GeneHogs (Invitrogen Grand Island, NY). DNA extraction and Illumina sequencing were undertaken as for host strains, as described above. Unmapped reads (corresponding to presumed plasmid sequences) were extracted from the mapped .bam files using features available in Samtools and Picard (version 1.66). Unmapped, paired end reads represented in the resulting fastq files were then assembled with A5(31). This resulted in one and two A5-generated scaffolds respectively, corresponding to the electroporated KPC-containing plasmid sequences in the index patient's *K. pneumoniae* and *K. oxytoca*. Comparisons with the velvet assemblies derived from the complete sets of reads for each transformant were used to close gaps between the A5-generated scaffolds and create near-complete reference plasmid sequences. Mapping of the complete sets of reads to these near-complete reference plasmid sequences was used to confirm plasmid structure and correct minor sequence errors in the assemblies. This resulted in single, closed plasmid structures for each transformant, pKPC_UVA01 for the index case *K. pneumoniae* and pKPC_UVA02 for the *K. oxytoca* strains respectively.

For plasmid analysis across the wider dataset, reads were mapped to pKPC_UVA01 and pKPC_UVA02. At the time of the analysis, there were also a number of complete

bla$_{KPC}$-plasmid sequences publicly available in GenBank, namely: pKPC-NY79 [JX104759.1], pKP048 [FJ628167], pKpQIL [GU595196], pKPN101-IT [JX283456], pKpQIL-IT [JN233705], pKpS90 [JX461340], S15 [FJ223606], S9 [FJ223607], pKpQIL-LS6 [JX442975.1], pBK31551 [JX193301], pBK31567 [JX193302], pBK32179 [JX430448], pBK15692 [KC845573]). Pairwise comparisons to determine the extent of homology between these plasmids (using BLASTn version 2.2.27 with default parameters and including pKPC_UVA01 and pKPC_UVA02) was undertaken in order to select a non-redundant set; this included only those with homology over less than half the length of the shorter reference. Reads were then also mapped to the non-redundant, publicly available references.

Mapping to plasmid references was performed by my colleague, Dr Anna Sheppard, using BWA mem (version 0.7.5a; default parameters), and the depth of coverage for each position on the reference was ascertained with samtools mpileup version (0.1.19; -A flag and otherwise default parameters). A conservative, absolute value of 10 reads was used as a cut-off for presence at each position. Custom python scripts were used to calculate the proportion of positions with read depth above this threshold, presented as % plasmid presence. In the absence of well-defined and validated thresholds for plasmid presence at the time of the analysis, we defined an arbitrary cut-off of 90% of coverage of a plasmid reference with read depth ≥ 10 as representing plasmid presence in an isolate. This is similar to the arbitrary threshold used in a subsequent publication partly analysing the population structure of KPC plasmids in ST258 KPC-*K. pneumoniae*(32).

De novo assembled contigs containing the Tn*4401* transposon were identified using BLASTn against a Tn*4401* isoform b reference (no deletions upstream of the KPC gene; GenBank: EU176013.1) with an in-house python script, and the matching sequences were then extracted, aligned and visualised in Geneious(33). Annotations were added manually. Confirmation of the presence of large structural changes was done by mapping reads to the contig of interest with BWA(34), visualising coverage plots using R, and confirming presence/absence of structural contiguity.

7.3. RESULTS

In the complete surveillance dataset (August 2007-September 2012), of the 250 patients with suspected CPE by phenotypic testing for the presence of carbapenemase (modified Hodge/indirect carbapenemase test), 226 (90%) patients had evaluable isolates collected. Of these 226 patients, 202 (89%) had at least one carbapenemase-positive *Enterobacteriaceae* isolate, including 64 (28%) patients with at least one bla_{KPC}-positive *K. pneumoniae* isolate. Of the first culture-positive samples containing KPC-*K. pneumoniae* from 37 unique patients that underwent WGS for this study, twenty-three (62%) were sourced from clinical specimens derived from urine (10), respiratory (6), abdominal wound (4), and blood (3) samples. Fourteen (38%) isolates were cultured from peri-rectal surveillance specimens. Of these patients identified as initial carriers by surveillance, eight (57%) later yielded a clinical specimen with other bla_{KPC}-positive Enterobacteriaceae.

Seven patients imported *KPC-K. pneumoniae* into UVaMC, including the suspected index case (CAV1016 on Figure 7.1., ST45), of these: (i) three patients had a carbapenem-resistant *K. pneumoniae* positive clinical or surveillance culture obtained

at an outside hospital prior to transfer; (ii) two patients had a KPC-*K. pneumoniae*-positive urine culture obtained as an outpatient; (iii) a single patient had a KPC-*K. pneumoniae*-positive culture within 48 hours of admission. Of the 30 patients at risk of acquiring KPC-*K. pneumoniae* in UVaMC, the median duration of hospitalisation in the 90 days prior to first CPE isolation was 20 days (range 2-81). Eighty-three percent (25/30) were considered high-risk for acquisition at UVaMC and 17% (5/30) indeterminate risk.

7.3.1. Host-strain diversity, risk of acquisition and evolutionary clock of KPC-*K. pneumoniae* in UVaMC

Sixteen distinct STs were identified from the 37 *K. pneumoniae* isolates, two of which were novel (assigned new ST numbers: ST1521 and ST1522 [Pasteur scheme](27)). Forty-six percent (17/37) of the isolates belonged to discrete STs with two or fewer isolates per ST. There were two STs represented by more than two isolates: ST941 (n=13/37; 35%), which has not previously been described in association with bla_{KPC}, and the epidemic ST258 (n=7/37; 19%). None of the patients with ST258 KPC-KP isolates were considered high-risk for acquisition within UVaMC; five of them were imported and two were of indeterminate risk (Figure 7.1.). All of the patients with ST941 were considered high-risk for acquisition within UVaMC. Excluding the index and indeterminate acquisition cases, patients with ST258 were more likely to acquire KPC-*K. pneumoniae* outside UVaMC than non-ST-258 (Fisher's Exact; 5/5 versus 1/26; p=0.0001).

Eleven individuals had paired longitudinal samples sharing the same ST from which the molecular clock could be assessed, with a median of 18 days (range 0-274) and 1

SNV (range 0-10) between paired samples. The molecular clock was calculated as being $1.9x10^{-6}$ substitutions/called site/year (95% credibility interval [CI]: $1.1x10^{-6}$ - $2.9x10^{-6}$), equating to 10.1 substitutions/genome/year (95% CI: 5.7-15.6) (Figure 7.2.).

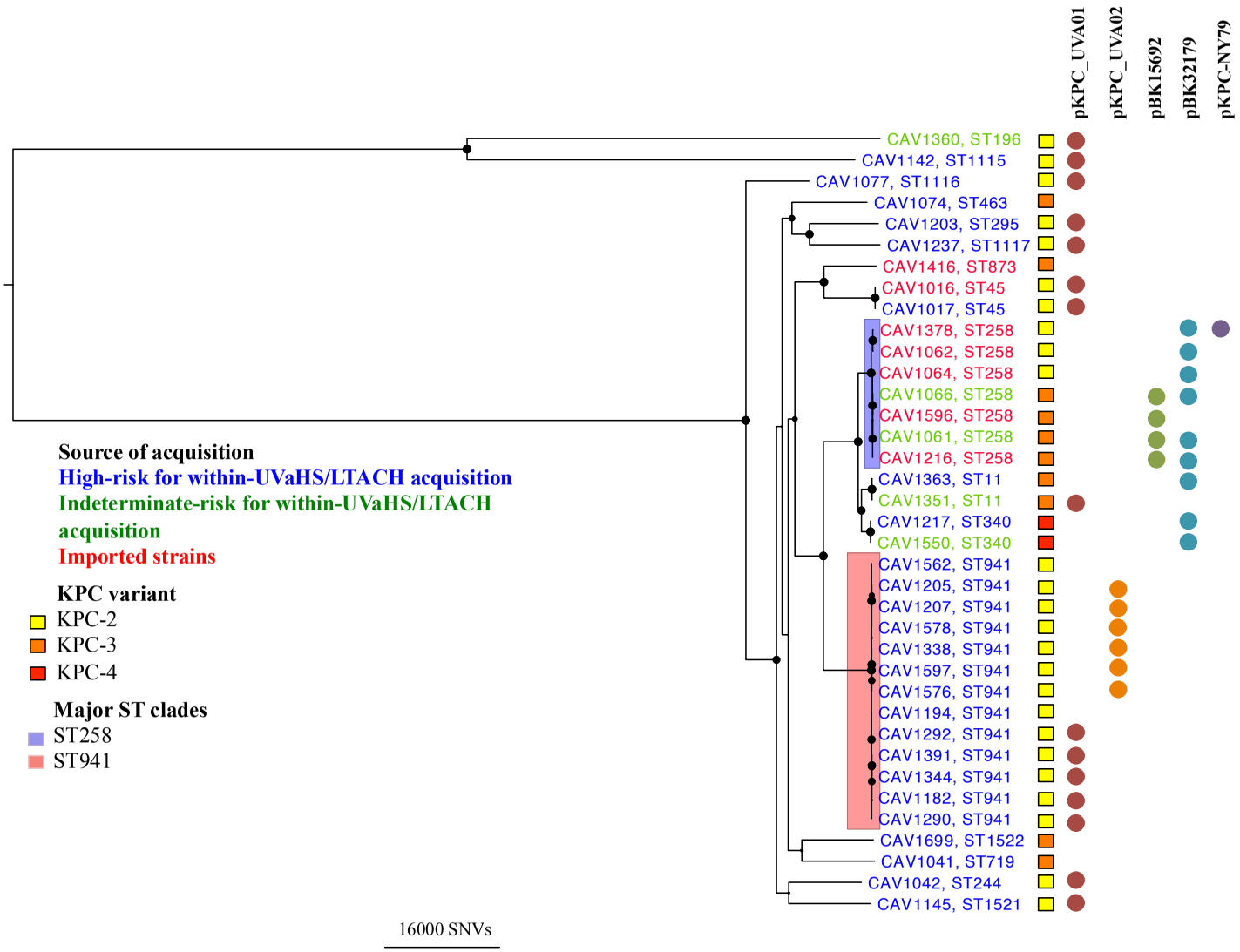


Figure 7.1. Maximum likelihood phylogeny of *Klebsiella pneumoniae* carbapenemase-*K. pneumoniae* core sequences, in association with bla_{KPC} plasmids (pKPC_UVA01, pKPC_UVA02, pBK15692, pBK32179, and pKPC-NY79), and risk of within-hospital acquisition. Black circles at the nodes represent bootstrap values >70% with the size of the circle reflecting the degree of support. The largest circles have bootstrap values of 100%. Sequence Type (ST), University of Virginia Health System/Long-term Acute Care Hospital (UVaMC), Single Nucleotide Variants (SNVs). Where there is no circle for a corresponding plasmid structure (e.g. CAV1074, CAV1416, CAV1562, CAV1194, CAV1699 and CAV1041), a significant alignment to a plasmid reference structure was not confirmed, and the location of bla_{KPC} in these isolates requires further assessment. CAV1016 (ST45) was the presumed index case.

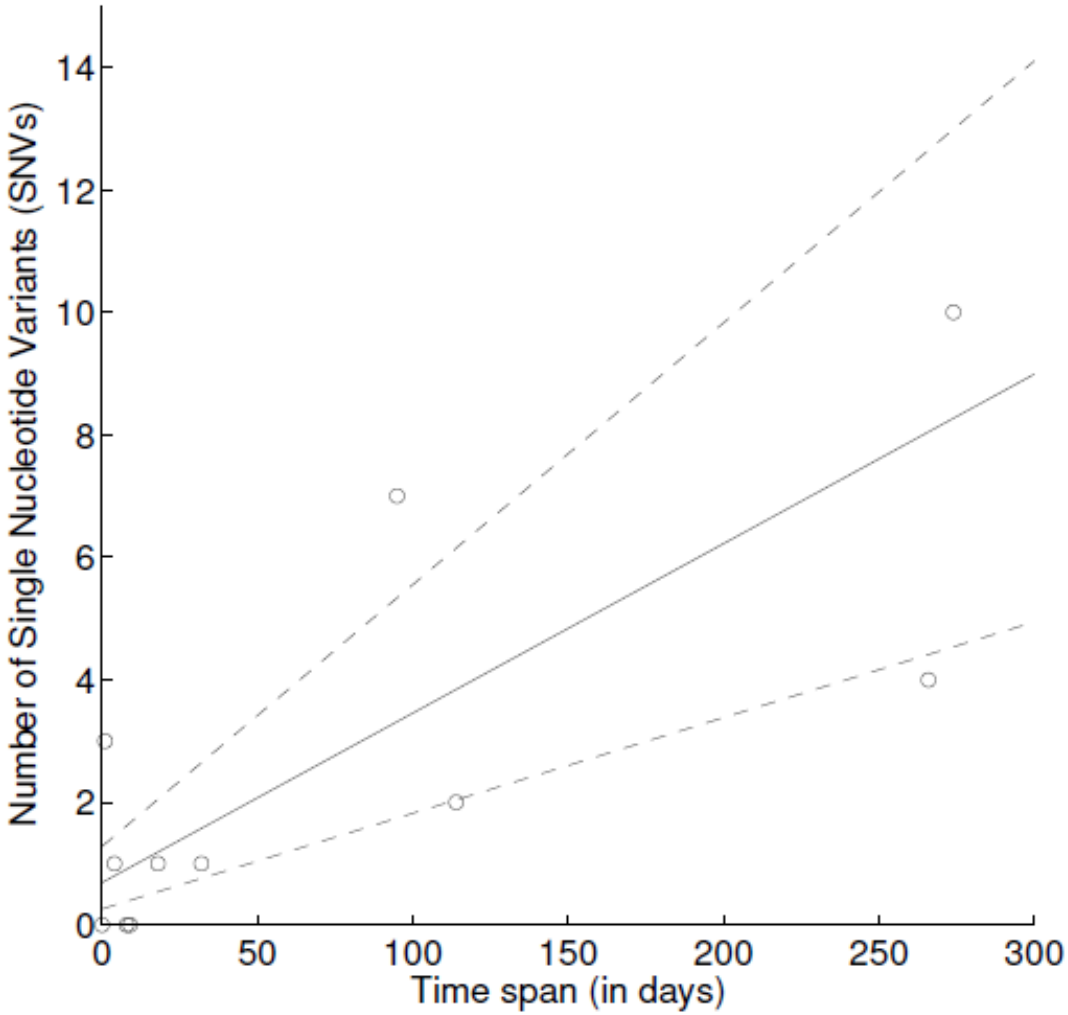


Figure 7.2. Molecular clock estimate calculated using Bayesian inference on genetic data from the first and last *Klebsiella pneumoniae* carbapenemase-*K. pneumoniae* isolates sampled from 11 study individuals. Solid line is the mean molecular clock estimate; dashed line is the 95% credibility interval around the estimate.

Based on a ClonalFrame host-strain analysis of the ST941 isolates, the TMRCA dated to mid-2006 with a 95% credibility interval ranging from 2003 to 2008 (Figure 7.3.). This includes the time of the first appearance in 2007 of bla_{KPC} in UVaMC and would be consistent with this lineage acquiring bla_{KPC} soon after its appearance in the local hospital ecosystem. A ClonalFrame analysis of the seven ST258 isolates demonstrated the TMRCA dating to 1997 (95% credibility interval: 1988-2002; Figure 7.4.). Four of these isolates (CAV1596, CAV1061, CAV1066 and CAV1216) were closely related (≤15 SNVs) and had a shared exposure to a single outside hospital. The remaining three ST258 KPC-*K. pneumoniae* were genetically divergent and imported to UVaMC from three separate outside hospitals in the mid-Atlantic region of the United States. There was no epidemiologic evidence of sustained ST258 transmission within UVaMC.

Twenty-five of the isolates contained KPC-2 (68%), ten KPC-3 (27%) and two KPC-4 (5%); the distribution of resistance genes was not entirely congruent with the host-strain phylogeny, suggestive either of horizontal transmission, or of on-going evolution of the gene within its plasmid/transposon context (KPC-3 is one SNV different from KPC-2; KPC-4 is two SNVs different from both KPC-2 and KPC-3). In the transposon, all isolates had a series of identical SNVs compared with the reference, but only 15 KPC-2 and 3 KPC-3 containing isolates were otherwise identical to the reference ([51%]; Figure 7.5.). Two isolates shared a 99bp deletion upstream of the KPC gene (CAV1378, CAV1062) consistent with a known isoform, Tn*4401a*; a third isolate (CAV1077) had a novel 188bp deletion, representing a novel isoform. One isolate (CAV1064) had a truncated transposon region involving the IS*Kpn7* insertion sequence and genes upstream of it. These large deletions and the

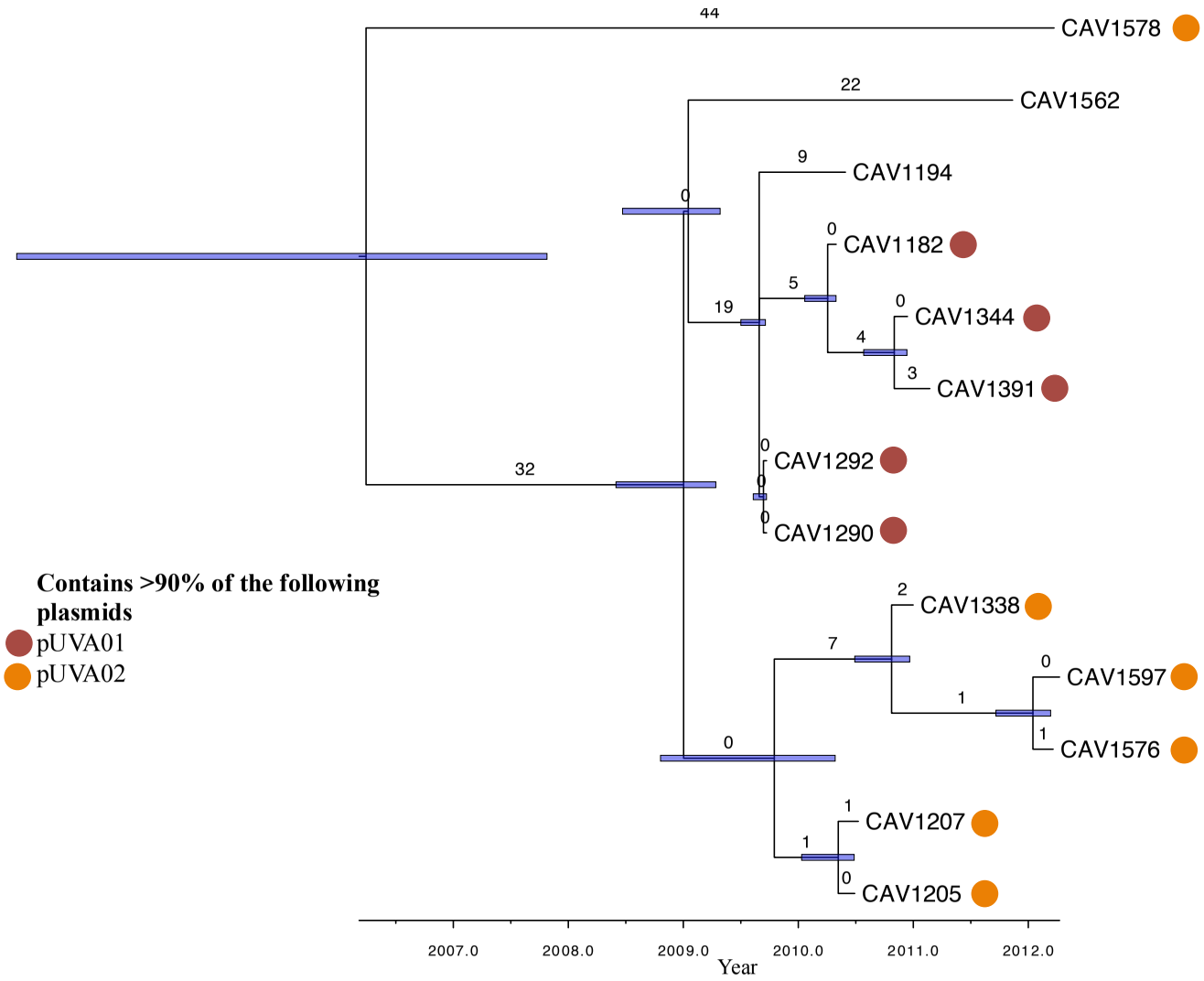


Figure 7.3. Time-scaled representation of genetic relationships between ST941 strains, in association with predicted *Klebsiella pneumoniae* carbapenemase-plasmid structures. The number of mutational substitutions are represented numerically on the branches, with bars at nodes indicating 95% credibility intervals around the estimates of the dating of the most recent common ancestors.

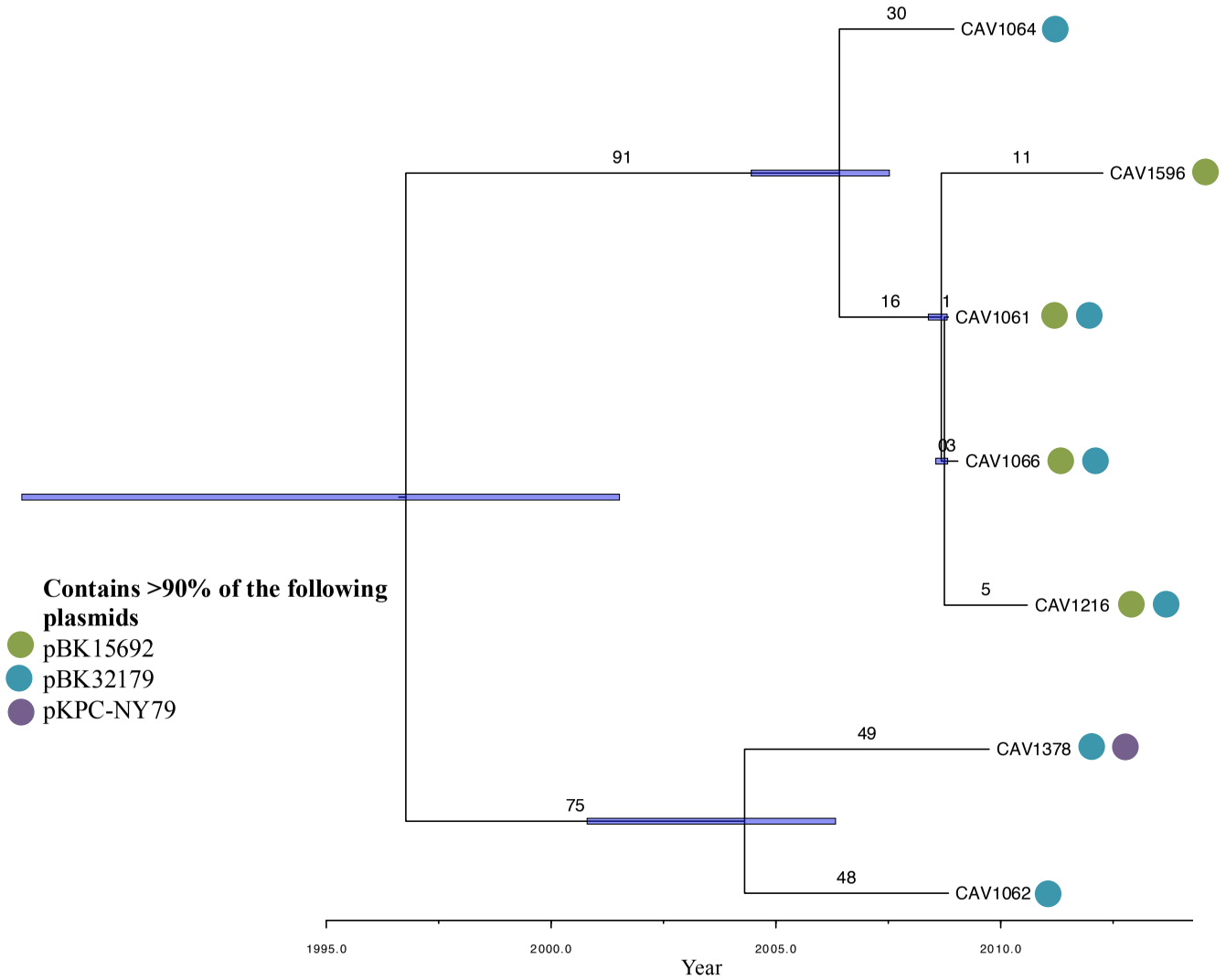


Figure 7.4. Time-scaled representation of genetic relationships between ST258 strains, in association with predicted *Klebsiella pneumoniae* carbapenemase-plasmid structures. The number of mutational substitutions are represented numerically on the branches, with grey bars at nodes indicating uncertainty 95% credibility intervals around the estimates of the dating of the most recent common ancestors.

truncation were confirmed by investigating read coverage depth across the region at these positions. The remaining strains all exhibited additional degrees of diversity in the transposon region. Much of this appeared to be in the *istA* gene, one of the transposase genes in the ISKpn7 (15/36 [42%] isolates, excluding the isolate with the truncated *Tn4401* [Fig. 7.5.]); this variability remains to be confirmed, as it may represent an artefact of the de novo assemblies. One isolate, CAV1074, had a single additional SNV (T>C at position 9621) in the *TnpA* gene of the ISKpn6 sequence; CAV1550 had an additional SNV (T>C at position 6800) in the *istB* gene of the ISKpn7.

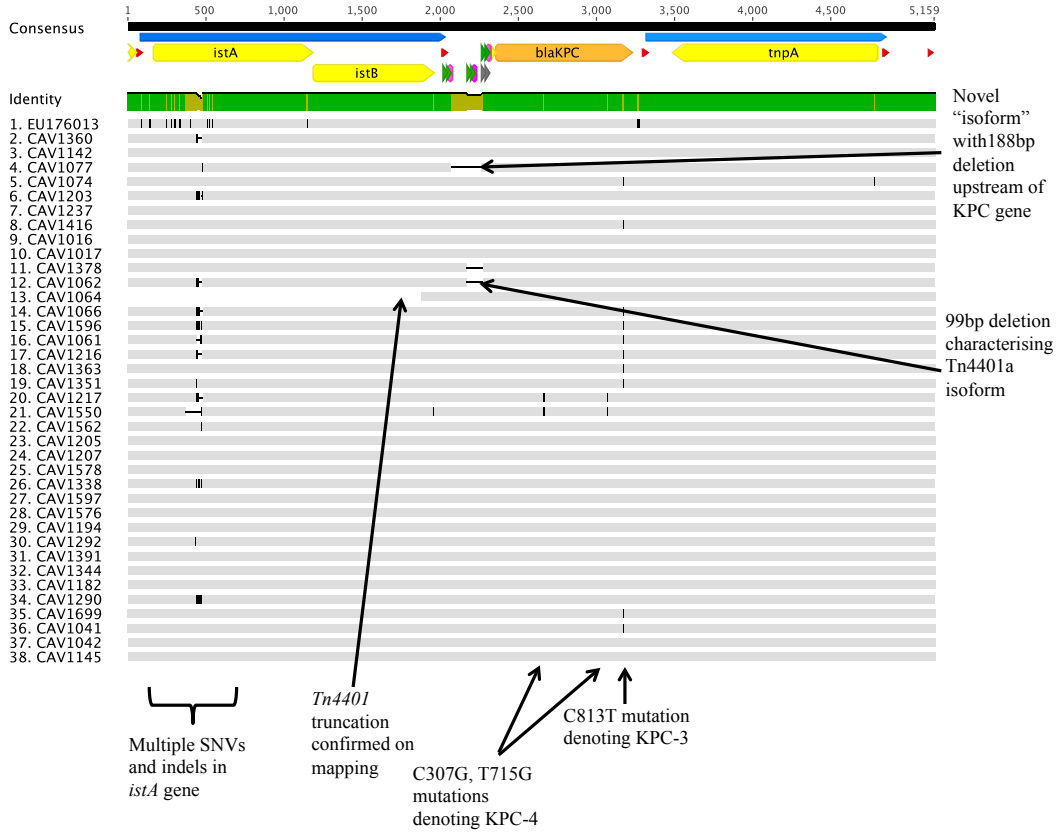


Figure 7.5. Alignment of ~5kb of the 3' end of the *Tn4401* transposon of 37 study sequences, and the Tn4401b reference (EU176013). Light grey denotes sequence homology, black vertical lines represent nucleotide differences and thin horizontal black lines in sequences represent deletions. Other features are marked. Annotations are labelled in the consensus sequence at the top of the figure.

De novo assembly of the index bla_{KPC} plasmid from CAV1016 generated a 43,621 base-pair (bp), closed, non-typable (by incompatibility group) Tn*4401* containing

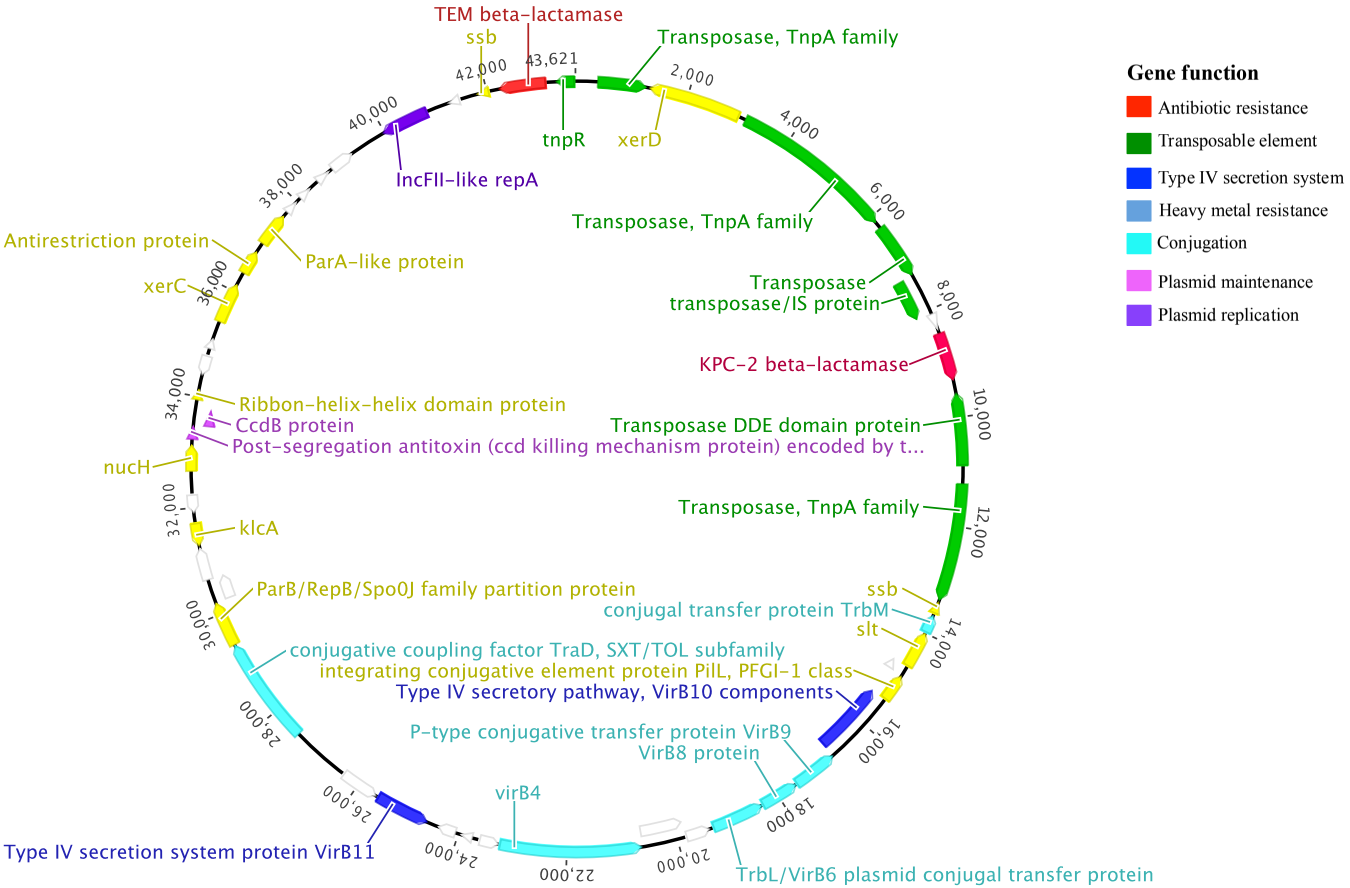
plasmid (pKPC_UVA01) with little homology to previously described plasmids (Figure 7.6.). Mapping reads from the 37 KPC-KP isolates against pKPC_UVA01 demonstrated that 41% (15/37) of the isolates, spanning ten STs over a 3.5 year period, contained >90% this plasmid sequence (Figure 7.1. above and Table 7.7.).

Assembly of plasmid reads from the CAV1015 *K. oxytoca* transformant resulted in two scaffolds totalling ~113kb, giving a non-typable plasmid structure designated pKPC_UVA02 (Figure 7.6.). A total of six isolates carried pKPC_UVA02, all of which were ST-941 (Table 7.7.), and all of which, except CAV1338, contained the same transposon in conjunction with a KPC-2 gene (Figure 7.5.).

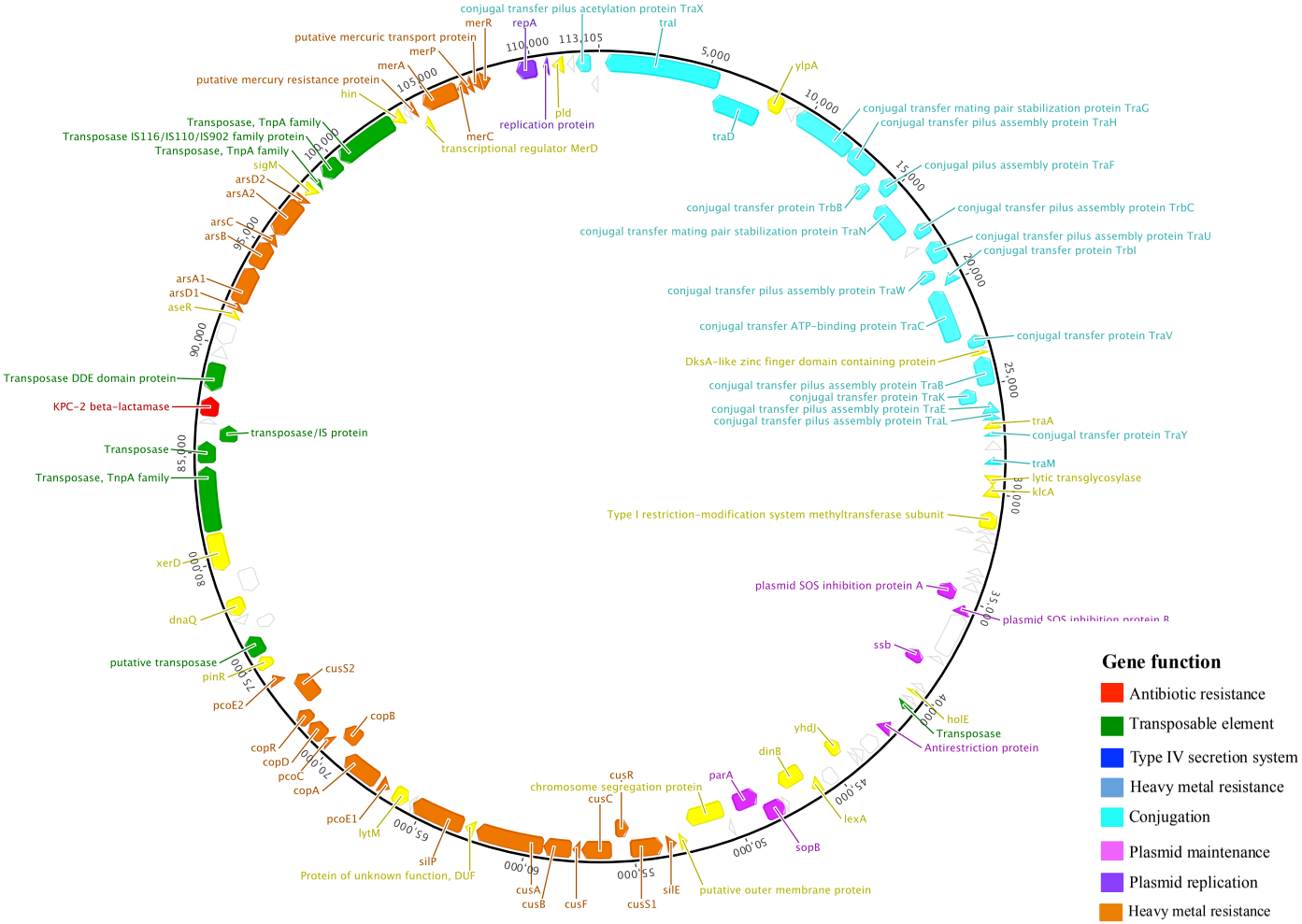
In the ST-941s, five isolates carried pKPC_UVA01 – these isolates also formed a host strain cluster on the tree (Figure 7.3. above), suggestive of localised transmission; however, although these were all KPC-2, 3 strains may have contained *istA*-level variation in their transposons. The absence of pKPC_UVA01 in CAV1562 and CAV1194 suggest the acquisition of pKPC_UVA01 by this group of strains some time in the second half of 2009-early 2010. The most parsimonious explanation for the presence of pKPC_UVA02 in divergent clusters on the ST-941 phylogeny, given the uniformity of the transposon in the most divergent strain, CAV1578, and in CAV1205/CAV1207/CAV1597/CAV1576, represents its acquisition on at least two separate occasions, also around 2009-2010. In the CAV1205/CAV1207/CAV1576/CAV1597 cluster, this was associated with clonal expansion and on-going spread; this cluster may also include CAV1338, although the variability in its *istA* region (Figure 7.5.) may reflect a separate acquisition event, further evolution within the plasmid, or an artefact in the de novo assembly. The exact

Figure 7.6. Two plasmids identified from transformant assemblies.

pUVA01 – 43,621bp



pUVA02 – 113,105bp



nature of the KPC-plasmid structures in CAV1562 and CAV1194 remain unresolved with currently available data, but they probably contain different, novel plasmid structures.

Plasmid analysis performed by mapping reads to fully sequenced KPC-*K. pneumoniae* plasmids available in GenBank demonstrated all strains that contained a published reference plasmid were members of clonal complex (CC)-258 (i.e. ST11, ST258, and ST340). pKPC-NY79 was found in a single ST258 isolate (CAV1378), and pBK15692 was found in four closely related ST258 isolates (CAV1596, CAV1061, CAV1066, and CAV1216), with exposure to the same referring hospital over a 41 month period (Figure 7.1. above). All ST258 isolates except CAV1596 carried pBK32179, as did one ST11 isolate (CAV1363) and two ST340 isolates (CAV1217 and CAV1550). Details of the percentage homology of all study isolates to the relevant reference plasmids is presented in Table 7.7.

Isolate name	Date of Isolation	Risk of in-hospital acquisition	Multi locus Sequence Type	KPC allelle	pKPC_UVA01 (%)	pKPC_UVA02 (%)	pBK15692 (%)	pBK32179 (%)	pKPC-N79 (%)
CAV1016	31-Aug-07	Index case	45	2	100	20	22	23	28
CAV1017	05-Oct-07	High	45	2	100	20	22	23	28
CAV1042	06-Apr-08	High	244	2	100	30	21	52	28
CAV1041	08-Apr-08	High	719	3	30	29	17	39	21
CAV1061	02-Nov-08	Indeterminate	258	3	34	35	100	97	31
CAV1062	03-Nov-08	Imported	258	2	33	33	22	95	84
CAV1064	16-Dec-08	Imported	258	2	18	31	14	93	81
CAV1066	18-Jan-09	Indeterminate	258	3	35	37	100	97	31
CAV1077	12-Feb-09	High	1116	2	100	28	20	65	23
CAV1074	13-Mar-09	High	463	3	23	34	13	60	21
CAV1142	10-Aug-09	High	1115	2	100	25	21	30	21
CAV1145	04-Sep-09	High	1521	2	100	38	22	67	30
CAV1292	22-Sep-09	High	941	2	100	38	23	68	25
CAV1290	22-Sep-09	High	941	2	100	38	24	68	25
CAV1378	29-Sep-09	Imported	258	2	34	34	24	100	100
CAV1360	28-Nov-09	Indeterminate	196	2	100	31	20	48	28
CAV1363	04-Dec-09	High	11	3	71	37	20	95	21
CAV1182	30-Apr-10		941		100	39	23	68	25
CAV1194	29-May-10	High	941	2	34	38	23	68	23
CAV1237	07-Jun-10	High	1117	2	100	30	20	57	21
CAV1203	10-Jun-10	High	295	2	100	15	20	23	21
CAV1205	28-Jun-10	High	941	2	34	100	23	69	23
CAV1207	09-Jul-10	High	941	2	34	100	23	69	23
CAV1217	04-Aug-10	High	340	4	25	43	14	94	21
CAV1216	05-Aug-10	Imported	258	3	34	34	100	97	31
CAV1344	13-Dec-10	High	941	2	100	37	23	67	25
CAV1338	31-Dec-10	High	941	2	34	100	23	69	21
CAV1351	07-Feb-11	Indeterminate	11	3	100	29	21	83	21
CAV1391	23-Feb-11	High	941	2	100	38	23	68	25
CAV1416	09-May-11	Imported	873	3	34	24	22	55	28
CAV1550	25-May-11	Indeterminate	340	4	28	33	18	97	28
CAV1562	12-Nov-11	High	941	2	34	38	23	67	23
CAV1576	19-Mar-12	High	941	2	34	100	23	69	21
CAV1597	09-Apr-12	High	941	2	34	100	23	69	21
CAV1596	09-Apr-12	Imported	258	3	34	15	100	15	31
CAV1699	24-Sep-12	High	1522	3	30	29	17	40	28
CAV1578	22-Mar-12	High	941	2	23	96	14	67	21

Table 7.7. Summary of acquisition status, ST, KPC-variant and % homology to reference KPC plasmids

7.4. DISCUSSION

This WGS-based analysis of a relatively small set of KPC-*K. pneumoniae* dating from August 2007 when bla_{KPC} was introduced into a single US hospital has demonstrated: (i) transfer of two newly-described plasmids likely containing Tn*4401*, pKPC_UVA01 and pKPC_UVA02, to other *K. pneumoniae* lineages, and in the case of pKPC_UVA01 to genetically diverse lineages; (ii) emergence of a new KPC-*K. pneumoniae* lineage ST941 which has acquired these two plasmids (and possibly others) independently and has then undergone local spread from a common source; (iii) Multiple independent importation events of KPC- *K. pneumoniae* ST258, which has not spread or become established in the local hospital ecosystem.

These observations are at variance with the prevailing and widely held view that successful dissemination of KPC- *K. pneumoniae* is most often restricted to ST258 *K. pneumoniae* or highly related strains in the clonal complex(2, 5, 35). Instead, they are consistent with the introduction by an index case in 2007 of two independent plasmids, pKPC_UVA01 and pKPC_UVA02 - present in a *K. pneumoniae* and *K. oxytoca,* strain respectively - to UVaMC(4, 22). These plasmids have then dispersed among diverse *K. pneumoniae* lineages, and may have donated the bla_{KPC}-containing Tn*4401* to other plasmid structures that could not be resolved in this analysis. In contrast to historical descriptions of KPC-*K. pneumoniae* nosocomial dissemination, typically associated with spread of a single lineage, (8, 36-39), this analysis describes a complex scenario including the emergence of a particularly promiscuous plasmid (pKPC_UVA01) and a novel *K. pneumoniae* host strain (ST941). Recently, similar multi-clonal, complex, KPC-*K. pneumoniae* outbreak scenarios have been observed in

Spain and Norway (40, 41), and may represent a highly worrying "lineage-escape" phenomenon for bla_{KPC}.

It is not clear why KPC-*K. pneumoniae* ST258 was not involved in on-going transmission as seen elsewhere, despite multiple introductions. The reason for the differences in transmission epidemiology between ST258 and ST941 is not evident, but may be related to the initial introduction, promiscuity and/or selective advantage afforded by resistance plasmids (in this case particularly pKPC_UVA01) rather than a single clonal complex-plasmid combination.

There are several limitations of this study. One obvious shortcoming was that I only chose to analyse a relatively small subset of the most globally relevant species of CPE, namely *K. pneumoniae*; current work however is investigating evidence for inter-species/inter-genus transfer of mobile genetic elements. In spite of this single species approach, which has undoubtedly left gaps in the total plasmid/transposon/resistance gene epidemiology, a convoluted dynamic of plasmid and strain interaction is already clearly observed.

In addition, with the exception of pKPC_UVA01 and pKPC_UVA02, we have not been able to obtain a closed sequence for any of the plasmid assemblies, leading to some remaining structural uncertainty regarding plasmids across the dataset, including determining the precise relationship with Tn*4401*. The increasing availability of long-read strand sequencing technology is likely to overcome this shortcoming, and I am currently involved in work trying to utilise PacBio sequencing to better define the plasmids involved. The decision to consider a plasmid as present

in an isolate if sequencing reads mapped to >90% of a plasmid with a minimum depth of ten reads at each position was an arbitrary cut-off which was chosen, but is a pragmatic approach to defining presence/absence and similar to cut-offs used in other work(32). As can be seen in Table 7.8., the presence/absence of plasmids in these isolates frequently represents a spectrum, which requires a degree of caution in its interpretation.

In summary, I describe a complex dynamic of plasmid and strain interaction among KPC-*K. pneumoniae* at a single institution over time. There has been remarkably multi-faceted dispersal of bla_{KPC}: namely, among plasmids by Tn*4401*, by plasmids across multiple lineages of *K. pneumoniae* and by the host *K. pneumoniae* strains themselves. Analysis of plasmid and host bacterial diversity has given new insights into the ways in which KPC-*K. pneumoniae* have become endemic in a single hospital over a relatively short timeframe.

CHAPTER 7 REFERENCES

1. **(CDC) CfDCaP.** 2013. *Antibiotic resistance threats in the United States*, p. 1-114. *In* (CDC) CfDCaP (ed.), 1 ed, vol. 1, Atlanta, GA USA.
2. **Munoz-Price LS, Poirel L, Bonomo RA, Schwaber MJ, Daikos GL, Cormican M, Cornaglia G, Garau J, Gniadkowski M, Hayden MK, Kumarasamy K, Livermore DM, Maya JJ, Nordmann P, Patel JB, Paterson DL, Pitout J, Villegas MV, Wang H, Woodford N, Quinn JP.** 2013. Clinical epidemiology of the global expansion of Klebsiella pneumoniae carbapenemases. The Lancet infectious diseases **13:**785-796.
3. **Sidjabat HE, Silveira FP, Potoski BA, Abu-Elmagd KM, Adams-Haduch JM, Paterson DL, Doi Y.** 2009. Interspecies spread of Klebsiella pneumoniae carbapenemase gene in a single patient. Clin Infect Dis **49:**1736-1738.
4. **Mathers AJ, Cox HL, Kitchel B, Bonatti H, Brassinga AK, Carroll J, Scheld WM, Hazen KC, Sifri CD.** 2011. Molecular dissection of an outbreak of carbapenem-resistant enterobacteriaceae reveals Intergenus KPC carbapenemase transmission through a promiscuous plasmid. MBio **2:**e00204-00211.
5. **Kitchel B, Rasheed JK, Patel JB, Srinivasan A, Navon-Venezia S, Carmeli Y, Brolund A, Giske CG.** 2009. Molecular epidemiology of KPC-producing Klebsiella pneumoniae isolates in the United States: clonal expansion of multilocus sequence type 258. Antimicrob Agents Chemother **53:**3365-3370.
6. **Leavitt A, Carmeli Y, Chmelnitsky I, Goren MG, Ofek I, Navon-Venezia S.** 2010. Molecular epidemiology, sequence types, and plasmid analyses of

KPC-producing Klebsiella pneumoniae strains in Israel. Antimicrobial agents and chemotherapy **54:**3002-3006.

7. **Baraniak A, Izdebski R, Herda M, Fiett J, Hryniewicz W, Gniadkowski M, Kern-Zdanowicz I, Filczak K, Lopaciuk U.** 2009. Emergence of Klebsiella pneumoniae ST258 with KPC-2 in Poland, p. 4565-4567, Antimicrob Agents Chemother, vol. 53, United States.

8. **Samuelsen O, Naseer U, Tofteland S, Skutlaberg DH, Onken A, Hjetland R, Sundsfjord A, Giske CG.** 2009. Emergence of clonally related Klebsiella pneumoniae isolates of sequence type 258 producing plasmid-mediated KPC carbapenemase in Norway and Sweden. J Antimicrob Chemother **63:**654-658.

9. **Mammina C, Bonura C, Di Bernardo F, Aleo A, Fasciana T, Sodano C, Saporito MA, Verde MS, Tetamo R, Palma DM.** 2012. Ongoing spread of colistin-resistant Klebsiella pneumoniae in different wards of an acute general hospital, Italy, June to December 2011. Euro Surveill **17**.

10. **Yoo JS, Kim HM, Yoo JI, Yang JW, Kim HS, Chung GT, Lee YS.** 2013. Detection of clonal KPC-2-producing Klebsiella pneumoniae ST258 in Korea during nationwide surveillance in 2011. J Med Microbiol **62:**1338-1342.

11. **Yang J, Ye L, Guo L, Zhao Q, Chen R, Luo Y, Chen Y, Tian S, Zhao J, Shen D, Han L.** 2013. A nosocomial outbreak of KPC-2-producing Klebsiella pneumoniae in a Chinese hospital: dissemination of ST11 and emergence of ST37, ST392 and ST395. Clin Microbiol Infect.

12. **Andrade LN, Curiao T, Ferreira JC, Longo JM, Climaco EC, Martinez R, Bellissimo-Rodrigues F, Basile-Filho A, Evaristo MA, Del Peloso PF, Ribeiro VB, Barth AL, Paula MC, Baquero F, Canton R, Darini AL, Coque TM.** 2011. Dissemination of blaKPC-2 by the spread of Klebsiella

pneumoniae clonal complex 258 clones (ST258, ST11, ST437) and plasmids (IncFII, IncN, IncL/M) among Enterobacteriaceae species in Brazil. Antimicrob Agents Chemother **55:**3579-3583.

13. **Tzouvelekis LS, Miriagou V, Kotsakis SD, Spyridopoulou K, Athanasiou E, Karagouni E, Tzelepi E, Daikos GL.** 2013. KPC-Producing, Multi-Drug Resistant Klebsiella pneumoniae ST258 as a Typical Opportunistic Pathogen. Antimicrob Agents Chemother.

14. **Lavigne JP, Cuzon G, Combescure C, Bourg G, Sotto A, Nordmann P.** 2013. Virulence of Klebsiella pneumoniae Isolates Harboring bla KPC-2 Carbapenemase Gene in a Caenorhabditis elegans Model. PLoS One **8:**e67847.

15. **Adler A, Paikin S, Sterlin Y, Glick J, Edgar R, Aronov R, Schwaber MJ, Carmeli Y.** 2012. A swordless knight: epidemiology and molecular characteristics of the blaKPC-negative sequence type 258 Klebsiella pneumoniae clone. J Clin Microbiol **50:**3180-3185.

16. **Chmelnitsky I, Shklyar M, Hermesh O, Navon-Venezia S, Edgar R, Carmeli Y.** 2013. Unique genes identified in the epidemic extremely drug-resistant KPC-producing Klebsiella pneumoniae sequence type 258. J Antimicrob Chemother **68:**74-83.

17. **Cuzon G, Naas T, Nordmann P.** 2011. Functional characterization of Tn4401, a Tn3-based transposon involved in blaKPC gene mobilization. Antimicrob Agents Chemother **55:**5370-5373.

18. **Naas T, Cuzon G, Truong HV, Nordmann P.** 2012. Role of ISKpn7 and deletions in blaKPC gene expression. Antimicrobial agents and chemotherapy **56:**4753-4759.

19. **Naas T, Cuzon G, Villegas MV, Lartigue MF, Quinn JP, Nordmann P.** 2008. Genetic structures at the origin of acquisition of the beta-lactamase bla KPC gene. Antimicrobial agents and chemotherapy **52:**1257-1263.

20. **Chen L, Chavda KD, Mediavilla JR, Jacobs MR, Levi MH, Bonomo RA, Kreiswirth BN.** 2012. Partial excision of blaKPC from Tn4401 in carbapenem-resistant Klebsiella pneumoniae. Antimicrobial agents and chemotherapy **56:**1635-1638.

21. **Cuzon G, Naas T, Truong H, Villegas MV, Wisell KT, Carmeli Y, Gales AC, Venezia SN, Quinn JP, Nordmann P.** 2010. Worldwide diversity of Klebsiella pneumoniae that produce beta-lactamase blaKPC-2 gene. Emerging infectious diseases **16:**1349-1356.

22. **Mathers AJ, Cox HL, Bonatti H, Kitchel B, Brassinga AK, Wispelwey B, Sawyer RG, Pruett TL, Hazen KC, Patel JB, Sifri CD.** 2009. Fatal cross infection by carbapenem-resistant Klebsiella in two liver transplant recipients. Transpl Infect Dis **11:**257-265.

23. **Mathers AJ, Carroll J, Sifri CD, Hazen KC.** 2013. Modified Hodge test versus indirect carbapenemase test: prospective evaluation of a phenotypic assay for detection of Klebsiella pneumoniae carbapenemase (KPC) in Enterobacteriaceae. Journal of clinical microbiology **51:**1291-1293.

24. **Mathers AJ, Poulter M, Dirks D, Carroll J, Sifri CD, Hazen KC.** 2014. Clinical Microbiology Costs for Methods of Active Surveillance for Klebsiella pneumoniae Carbapenemase-Producing Enterobacteriaceae. Infect Control Hosp Epidemiol **35:**350-355.

25. **Lewis JD, Bishop M, Heon B, Mathers AJ, Enfield KB, Sifri CD.** 2013. Admission surveillance for carbapenamase-producing enterobacteriaceae at a long-term acute care hospital. Infect Control Hosp Epidemiol **34:**832-834.

26. **Mathers AJ, Hazen KC, Carroll J, Yeh AJ, Cox HL, Bonomo RA, Sifri CD.** 2013. First clinical cases of OXA-48-producing carbapenem-resistant Klebsiella pneumoniae in the United States: the "menace" arrives in the new world. J Clin Microbiol **51:**680-683.

27. **Diancourt L, Passet V, Verhoef J, Grimont PA, Brisse S.** 2005. Multilocus sequence typing of Klebsiella pneumoniae nosocomial isolates. Journal of clinical microbiology **43:**4178-4182.

28. **Didelot X, Eyre DW, Cule M, Ip CL, Ansari MA, Griffiths D, Vaughan A, O'Connor L, Golubchik T, Batty EM, Piazza P, Wilson DJ, Bowden R, Donnelly PJ, Dingle KE, Wilcox M, Walker AS, Crook DW, A Peto TE, Harding RM.** 2012. Microevolutionary analysis of Clostridium difficile genomes to investigate transmission. Genome Biol **13:**R118.

29. **Didelot X, Falush D.** 2007. Inference of bacterial microevolution using multilocus sequence data. Genetics **175:**1251-1266.

30. **Gelman ARDB.** 1992. Inference from Iterative Simulation using Multiple Sequences. Statistical Science **7:**457-511.

31. **Tritt A, Eisen JA, Facciotti MT, Darling AE.** 2012. An integrated pipeline for de novo assembly of microbial genomes. PloS one **7:**e42304.

32. **Wright MS, Perez F, Brinkac L, Jacobs MR, Kaye K, Cober E, van Duin D, Marshall SH, Hujer AM, Rudin SD, Hujer KM, Bonomo RA, Adams MD.** 2014. Population Structure of KPC-Producing Klebsiella pneumoniae

Isolates from Midwestern U.S. Hospitals. Antimicrobial agents and chemotherapy **58:**4961-4965.

33. **Biomatters.** Geneious, 7.1 ed.

34. **Li H, Durbin R.** 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics **26:**589-595.

35. **Nordmann P, Naas T, Poirel L.** 2011. Global spread of Carbapenemase-producing Enterobacteriaceae. Emerg Infect Dis **17:**1791-1798.

36. **Navon-Venezia S, Leavitt A, Schwaber MJ, Rasheed JK, Srinivasan A, Patel JB, Carmeli Y.** 2009. First report on a hyperepidemic clone of KPC-3-producing Klebsiella pneumoniae in Israel genetically related to a strain causing outbreaks in the United States. Antimicrob Agents Chemother **53:**818-820.

37. **Nicoletti AG, Fehlberg LC, Picão RC, Machado AeO, Gales AC.** 2012. Clonal complex 258, the most frequently found multilocus sequence type complex in KPC-2-producing Klebsiella pneumoniae isolated in Brazilian hospitals. Antimicrob Agents Chemother **56:**4563-4564; author reply 4565.

38. **Giakkoupi P, Papagiannitsis CC, Miriagou V, Pappa O, Polemis M, Tryfinopoulou K, Tzouvelekis LS, Vatopoulos AC.** 2011. An update of the evolving epidemic of blaKPC-2-carrying Klebsiella pneumoniae in Greece (2009-10). J Antimicrob Chemother **66:**1510-1513.

39. **Snitkin ES, Zelazny AM, Thomas PJ, Stock F, Henderson DK, Palmore TN, Segre JA.** 2012. Tracking a hospital outbreak of carbapenem-resistant Klebsiella pneumoniae with whole-genome sequencing. Sci Transl Med **4:**148ra116.

40. **Ruiz-Garbajosa P, Curiao T, Tato M, Gijon D, Pintado V, Valverde A, Baquero F, Morosini MI, Coque TM, Canton R.** 2013. Multiclonal dispersal of KPC genes following the emergence of non-ST258 KPC-producing Klebsiella pneumoniae clones in Madrid, Spain. J Antimicrob Chemother.

41. **Tofteland S, Naseer U, Lislevand JH, Sundsfjord A, Samuelsen O.** 2013. A long-term low-frequency hospital outbreak of KPC-producing Klebsiella pneumoniae involving Intergenus plasmid diffusion and a persisting environmental reservoir. PloS one **8:**e59015.

CHAPTER 8: POPULATION GENETICS OF CTX-M-ASSOCIATED RESISTANCE WITH A COMMON GLOBAL *ESCHERICHIA COLI* LINEAGE, SEQUENCE TYPE (ST) 131

8.1. BACKGROUND

Resistance to extended-spectrum cephalosporins in non-enteropathogenic *E. coli* presents a major clinical challenge, given that beta-lactam antibiotics are still widely used in the management of infections caused by this organism. In 2008, it was suggested that the burden of disease caused by ESBL-*E. coli* was largely attributable to the emergence of a global pandemic clone, multi-locus sequence type (ST) 131, serogroup O25b(1, 2), a member of the B2 non-enteropathogenic *E. coli* phylogroup. Evidence for this hypothesis was presented in two separate studies, both of which sampled approximately 40 CTX-M-15 containing strains from a range of global locations, and found that the majority of them were ST131, had similar PFGE and virulence profiles, and were associated with plasmids containing IncFII-type replicons, although these were of variable size(3, 4). Some recent data suggest that the emergence and dissemination of ST131 has taken place in less than a decade(1, 5), and has been driven by the descendants of a single strain, subclone H30, that carry a specific *fimH* variant, *fimH*30(6). *fimH* is a gene encoding the mannose-specific type 1 fimbrial adhesin that is part of an alternative, recently developed partial sequence-based typing scheme shown to have equivalent if not better typing resolution than conventional, widely used MLST methods(7).

The rising rates of antimicrobial resistance to a number of other classes of antibiotic have also been reflected in members of the ST131 lineage, which in addition to its

association with the ESBL phenotype, is also particularly associated with resistance to fluoroquinolones(8, 9). CTX-M-15 has become the dominant ESBL in ST131, and along with CTX-M-14 has become widely prevalent as the cause of resistance to extended-spectrum beta-lactams since the 1990s, with almost contemporaneous identification in a number of geographically widespread locations possibly suggesting repeated acquisitions in multiple horizontal gene transfer events(10). Other data however support the notion that the clonal expansion of CTX-M-containing strains and globalisation make more important contributions to the wider transmission of these genes(11). Both CTX-M-15 and CTX-M-14 have been shown to be present on conjugative plasmids, with the former frequently being found on multireplicon IncFII plasmids additionally harbouring FIA and/or FIB replicons(12). Both variants have also been shown to integrate chromosomally, and previous studies have suggested that this can occur in different regions of the chromosome(13), although chromosomal acquisition and subsequent clonal expansion has not yet been demonstrated in *E. coli*.

The epidemiological analyses giving rise to theories around the relative contributions of horizontal gene transfer versus clonal spread of resistant isolates have traditionally been done on the basis of restriction fragment length polymorphism (RFLP)-based typing methods(3, 11). However, it has since been recognized that these methods are highly sensitive to horizontal gene transfer events and lack sufficient resolution to fully characterize the underlying genetic differences within a single lineage(13). In response, two important studies published in 2014 have harnessed the high-resolution typing possibilities afforded by WGS in the analysis of CTX-M and ST131. The first study, by Price *et al*, represented an analysis of 96 ST131 sequences, predominantly sampled from North America, and confirmed that the emergence of drug resistance in

ST131 *E. coli* in the USA has been driven by the successful clonal expansion of the *fimH*30 strain (the H30 lineage), which is associated with higher rates of severe clinical disease, and resistance to fluoroquinolones conferred by chromosomal *gyrA* mutations. Moreover, they also demonstrated that within the fluoroquinolone-resistant H30 sub-lineage (defined as H30-R), a subset of third generation cephalosporin-resistant isolates has emerged, the H30-Rx sub-lineage. ESBL resistance in these isolates is principally associated with the presence of CTX-M-15(13). However, because most isolates in this study were from North America, the authors were limited in making major global comparisons. They were also unable to calculate evolutionary rates, and did not look in great detail at the mobile elements potentially involved in the dissemination of resistance. The second study, by Petty *et al* extended the global range of the analysis, with a larger number (n=100) of isolates sampled from Europe and Australasia, and resolved the ST131 population structure into one of three clades, A, B and C, with the third clade composed of two sub-groups, C1 and C2 – the latter corresponding to the H30 and H30-Rx sub-groupings in the US study(14). This study likewise, however, did not investigate the plasmid component involved in much detail.

The apparent association of particular CTX-M variants within specific lineages in ST131 could be attributable to several possibilities: (i) acquisition of a CTX-M-containing plasmid or smaller mobile genetic unit (e.g. a transposon) at a specific time-point, and subsequent plasmid/transposon evolution within a host lineage by descent; (ii) multiple discrete CTX-M-containing plasmid acquisition events; (iii) chromosomal integration of CTX-M and evolution by descent; (iv) multiple chromosomal integration events; or, (v) a combination of the above. The scenario

described in (i) could be complicated, given that plasmid evolution might involve substantial genetic rearrangements even within single lineages, and transposons could conceivably move within the plasmid populations present in a host bacterial cell.

The work in this chapter builds on the two previously published pieces of work, by: (i) including a much larger number of strains from Asia (and larger subset of CTX-M-14/14-like-containing strains), representing a region of high ESBL resistance prevalence(15); (ii) estimating mutation rates of the lineage, to ascertain whether this was likely to be associated with its capacity to become prevalent; (iii) determining in greater detail the genetic context of CTX-M across a large dataset and the structure of some of the important plasmids involved in spreading CTX-M-associated resistance within the lineage, and their distribution within our global collection of isolates.

8.2. METHODS

8.2.1. Sample collection and sequencing

For this analysis, I included a novel subset of 108 ST131 isolates identified as part of my wider work in sequencing collections of clinical and carriage isolates from various locations in South-East Asia, Canada and Europe, and from Oxfordshire in the UK, between early 2005 and 2012. These strains were sub-cultured and the DNA extracted using the standard thesis methods protocol. Sequencing then took place on the Illumina HiSeq, generating100-150bp paired-end reads depending on the time of sequencing, except for 14 strains that were later acquisitions and were processed on the MiSeq, generating 150bp paired-end reads. We also obtained Illumina short-read data from a collaboration with a research group at AstraZeneca (n=11), and had early access to the data from the Price *et al* paper(13) which had been processed on the

Illumina HiSeq (n=96). The data from the more recent Petty *et al* paper(14) was downloaded from the European Nucleotide Archive (ENA) hosted by the European Bioinformatics Institute (EBI); these isolates had also been sequenced on the Illumina HiSeq (n=100).

Raw reads for all datasets were processed using the standard group pipeline. The data from the Price *et al* paper and from our collections were merged to form the main dataset of 215 whole genome sequences on which analyses were undertaken. The data from the Petty *et al* paper became available too late to be substantially included in the analysis, and also increased the dataset size to such an extent that it was not possible to repeat the analysis on the complete set of ST131 sequences given our software and hardware constraints. This dataset therefore only contributes to part of the analysis, and is specifically referenced in these contexts.

8.2.2. Sequence read processing

For phylogenetic analyses, Illumina reads were mapped to the genome of the reference ST131 strain, *E. coli* SE15 [(16); NC_013654]. This is a carriage strain, isolated from human faeces, and does not contain any ESBL genes. Alignments of core variable sites across the dataset were generated and reinserted into the reference to form a modified alignment of reference sequences for the dataset.

De novo assemblies were generated using the standard group pipeline approach using Velvet and VelvetOptimiser. Where assemblies appeared to be of poor quality, based on an inappropriate total assembly size, low/high number of contigs assembled or a

low N50, sequence data were re-processed using A5-MiSeq, to see if viable assemblies could be generated.

8.2.3. Characterisation of specific genetic sequences of interest

All assemblies were then analysed with BLASTn to identify genetic sequences of interest, specifically: (i) *bla*$_{\text{CTX-M}}$ presence and variant; (ii) genetic context for *bla*$_{\text{CTX-M}}$, by extracting and annotating contigs containing *bla*$_{\text{CTX-M}}$ variants using PROKKA; (iii) chromosomal *gyrA* mutations in the quinolone-resistance determining region known to be responsible for conferring most resistance to fluoroquinolones; and (iv) *fimH* presence and variant. I also confirmed that all strains to be included in the analysis were ST131 using *in silico* typing in accordance with the Achtman scheme(17).

8.2.4. ST131 chromosomal phylogenetic comparisons using RaxML, ClonalFrame and BEAST

Non-enteropathogenic *E. coli* are known to be recombinogenic, with rates of recombination generally estimated to be as frequent as mutation (recombination rate:mutation rate ratio = 1.024), and some recombination hotspots with higher rates(18). Recombination is known to obscure the true phylogenetic signal, and therefore needs to be addressed prior to undertaking any phylogenetic reconstructions. In order to do this, the alignment of modified reference sequences was initially analysed using ClonalFrame(19), to identify regions involved in recombination, which were then removed. For the ClonalFrame analysis, settings and an assessment of run quality were undertaken as per the analyses described in Chapter 7 [page 229].

When the Petty *et al* sequence data became publicly available, I attempted to repeat the ClonalFrame analysis on an extended dataset incorporating their strains. This however outstripped the hardware memory available to me. As a compromise, I included their data in the construction of a maximum-likelihood tree of all strains using RaxML, with the standard parameters described previously. This was done to broadly define the phylogenetic topology of the dataset including the Petty *et al* strains, and to confirm that there were no outlying clusters in their data that were unrepresented in our dataset, which was undergoing more detailed analysis. This tree reflects the data in the alignment of modified reference sequences without having removed any sites predicted to be involved in recombination (Figure 8.1.).

Using the alignment with regions of recombination removed, several models were set up in BEAST, primarily to calculate the evolutionary clock rate for the lineage, and to generate a time-scaled phylogeny. I explored four different models of population size change through time: (i) constant – i.e. no change in population size over time, (ii) exponential growth, (iii) logistic growth, and (iv) implementation of the Bayesian skyride, which in theory allows for fluctuations in population size, incorporating both increases and decreases in size(20). Each of these were implemented with two different molecular clock models: (i) the strict clock, estimating a uniform evolutionary rate across all branches of the tree, and (ii) the uncorrelated relaxed clock model, which co-estimates substitution parameters, relaxed-clock parameters (where an evolutionary rate is calculated along each tree branch and then presented in summary as a mean rate), and the ancestral phylogeny(21). For model selection, I implemented stepping stone sampling to determine the log marginal likelihood for

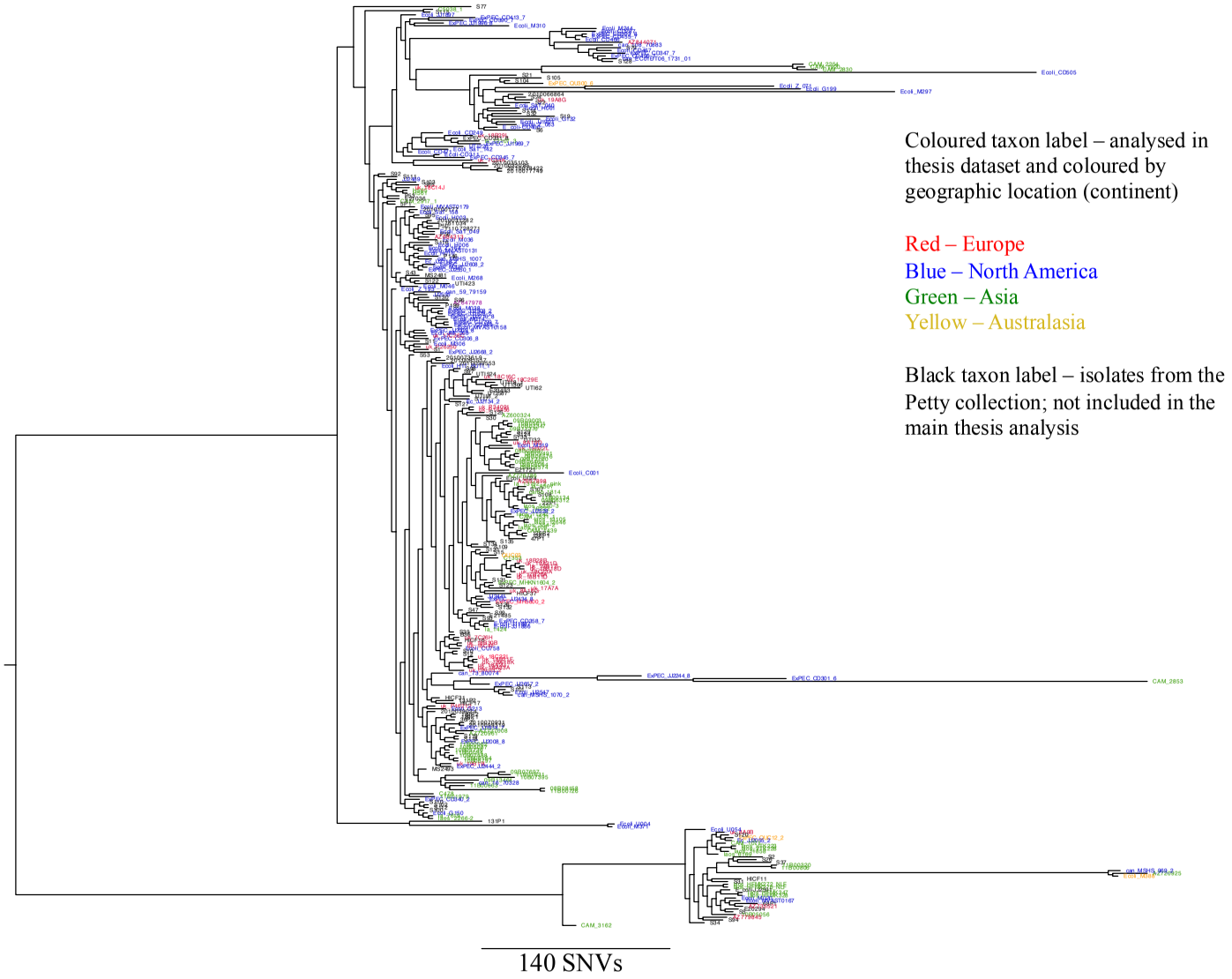


Figure 8.1. Maximum-likelihood phylogeny of all sequenced ST131 strains, including the Petty *et al* dataset(14), without accounting for recombination.

each run; this considerably increases run times, but had been shown to be a much more robust method of model selection than using the harmonic mean estimate (HME) or the Akaike Information Criterion through MCMC (AICM) estimate(22). Runs for each model were done in triplicate, with 30 million iterations, 10% of which were discounted as burn-in. Run convergence and mixing was assessed by inspecting the log files for each run in Tracer v1.5; adequate convergence of run statistics and mixing for each run and effective sample size (ESS) estimates for all parameters greater than 200 were required for an analysis to be considered adequate in line with recommendations in the BEAST tutorials on the developers' website (http://beast.bio.ed.ac.uk).

8.2.5. Plasmid transformations and analyses

Plasmid transformants were generated from a number of strains on the basis of tree topology and association with CTX-M variants. I aimed to transform at least one plasmid from each of the major CTX-M variant clusters where possible. I also transformed two CTX-M-15 containing plasmids from non-ST131 *E. coli* as an external comparison. DNA extraction, electroporation, sequencing, plasmid assembly and annotation were performed as per the thesis methods, using A5 or A5-MiSeq. Sequencing was performed on the Illumina HiSeq or MiSeq generating 150 or 300-base paired-end reads (Appendix Table 2). Sequencing reads from the isolate from which the transformed plasmid had been obtained were mapped back to the transformed plasmid assembly in order to ascertain the reliability of the assembly in each case.

Plasmid content across the dataset was investigated in a number of ways. Firstly, the transformants generated could be used as references, against which BLASTn-based comparisons for degree of presence/absence could be made for the whole dataset. Secondly, pairwise comparisons between each set of transformant assemblies were made by identifying the extent of shared homology of sequences across a number of subsets representing different groupings on the main host tree, again using BLASTn. For each pair, two % similarity statistics were generated, taking each member of the pair as a reference in turn, to account for differences in length. The mean % divergence was then plotted against the time to most recent common ancestor of the host strain (derived from the time-scaled tree). Finally, coding sequences across any transformant groups of interest were clustered using CD-Hit(23), to identify whether any coding sequences were shared, and whether there might be any biological significance associated with these on the basis of their annotations.

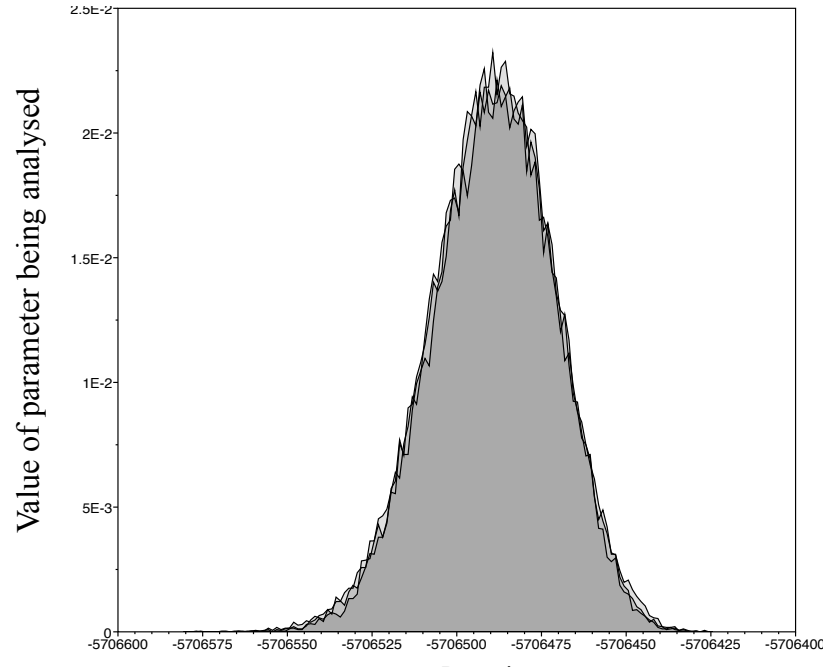
8.3. RESULTS

8.3.1. Chromosomal analysis

Of 4,717,338 sites in the reference, 40,057 (0.85%) were shown to be variable across the dataset, with 6,879 (0.15%) of the total shown to be core single nucleotide variants (core SNVs; i.e. no "N" or "-" calls for that position in the mapped data for any of the 215 isolates). The core SNVs were re-inserted into the reference genome, generating an alignment of modified reference sequences, and ClonalFrame was run on this dataset to identify regions involved in recombination (incorporating both variant and invariant sites). This resulted in the removal of 611,770 sites (13% of the genome), leaving 2,759 variant sites available for downstream analysis.

For the BEAST analysis, poor convergence across runs, and low ESS values (representing poor mixing) were observed for several parameters for all of the relaxed clock models. Run times had already stretched upwards of three weeks for these models, and repeating the analysis with extended run times was not feasible at this point. For the strict clock models, only the exponential model and the constant model had ESS values above 200 for all parameters across runs; the exponential model had the highest marginal likelihood value using the stepping stone sampling algorithm, but this may have been as a result of the impact of sampling, in which relatively larger numbers of dataset isolates had been sampled from more recent time periods. Data from the constant population size model, as the simplest of the models, were therefore used in downstream analyses. Figure 8.2. shows representative traces of posterior sampling and convergence of one of the parameters for the three combined runs for the strict clock, constant population size model, showing good mixing and good convergence.

The resulting global phylogeny of ST131 is consistent with the circulation of three major lineages of organisms, A-C, with C divisible into two subclades, C1 and C2, depending on the presence/absence of CTX-M-15 (A [n=25 isolates]; B [n=51]; C1 [n=59]; C2 [n=80]); on Figure 8.3.). The time to most recent common ancestor (TMRCA) varies from approximately 130 years ago to (for lineages A and B/C), to the more recently dated emergence of clade C as a distinct entity from clade B around 25 years ago, which would be contemporaneous with the widespread clinical use of ciprofloxacin and third generation cephalosporins. Approximately 250 non-recombinogenic, core SNVs separate clades A and B/C, 50-60 clades B and C1/C2, and 10-30 clades C1 and C2.



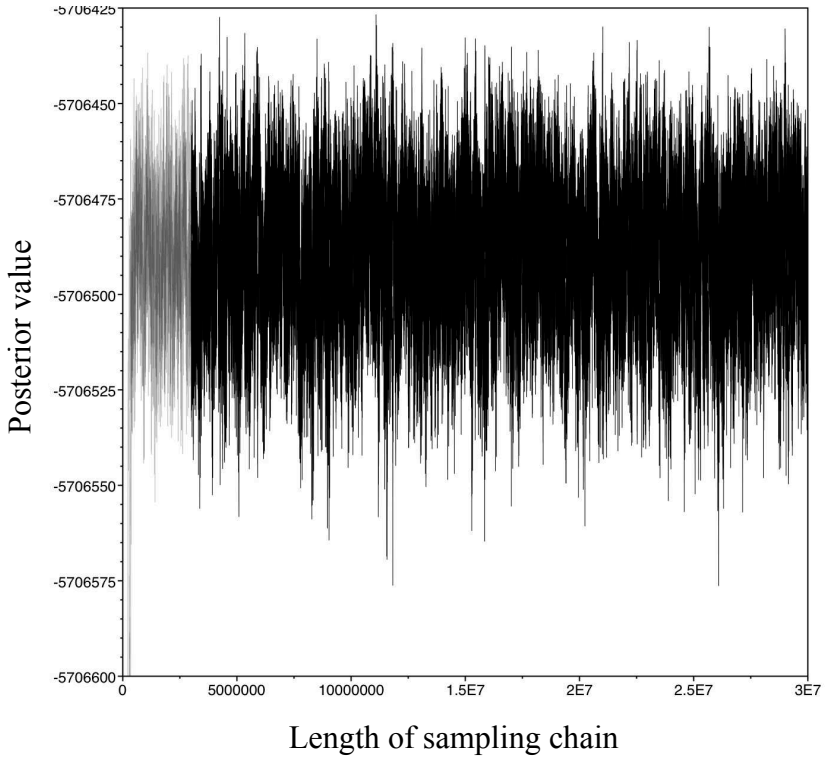


Figure 8.2. Example traces for parameters modelled in BEAST using a constant population size and strict growth model. The top panel shows marginal posterior densities for a parameter of the model for three independent runs showing near-identical distributions; the bottom panel demonstrates good mixing of the chains, with dense sampling of the posterior and a successful "hairy caterpillar" appearance.

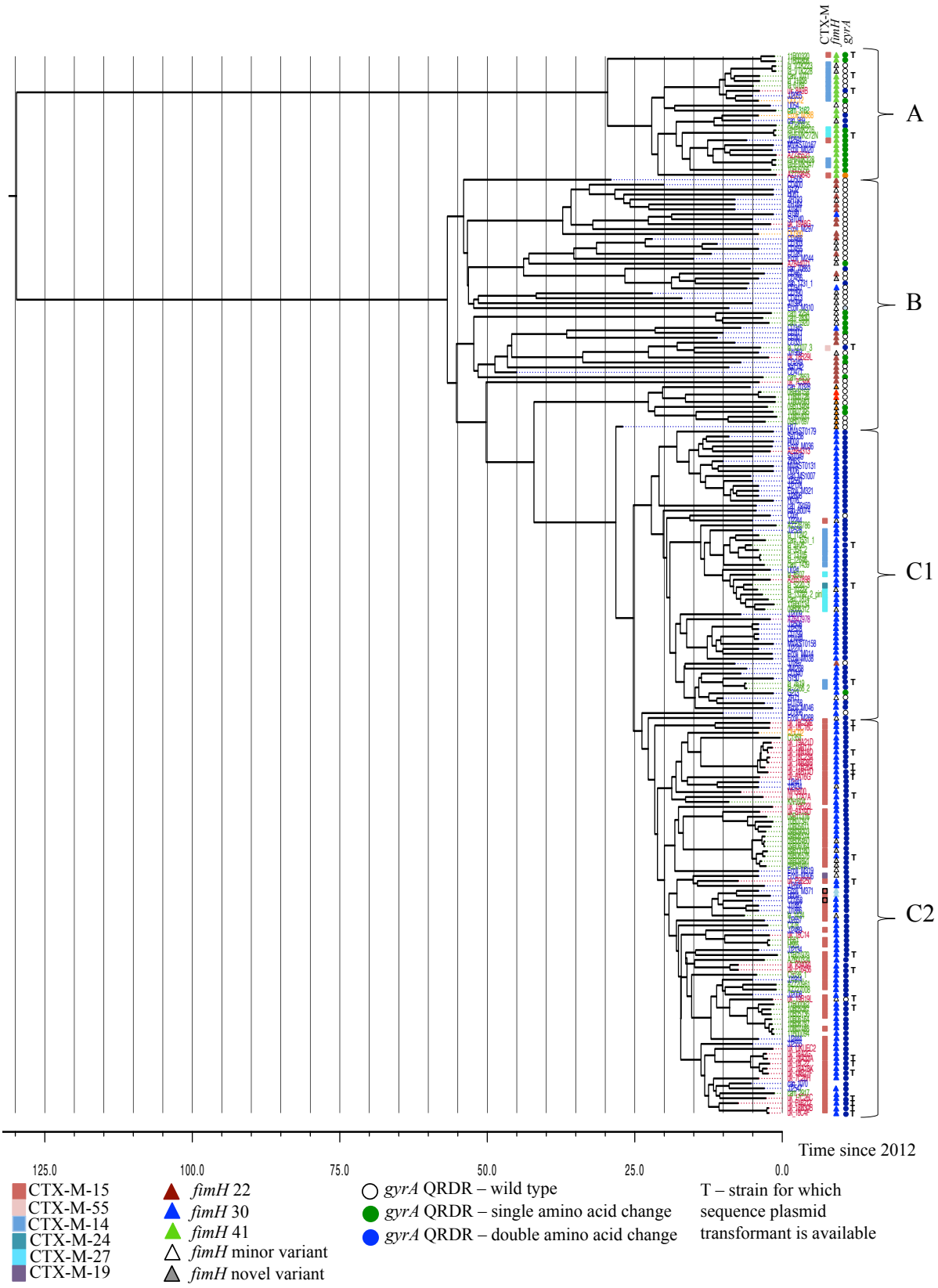


Figure 8.3. Time-scaled ST131 phylogeny based on 215 ST131 isolates. Taxon label colouring by geographic location: red = Europe; blue = North America; green = Asia; yellow = Australasia. Letters next to curly brackets represent identified clade structures, as per the Petty *et al* paper(14). CTX-M-variants with black borders represent the two occasions where only partial blast matches were identified.

The evolutionary rate within ST131 was estimated at $2.47x10^{-7}$ mutations/site or 1.01 mutations per genome/year, assuming a strict molecular clock and constant population size. This is similar to another published estimate obtained from a longitudinal in vivo sampling frame(24), and suggests that the clinical dominance of this strain is not attributable to faster core mutationary rates leading to increased adaptability.

8.3.2. CTX-M genes and flanking context of the CTX-M gene variants

Of the 215 strains, *de novo* assemblies for four isolates failed a quality control check (la_12107_3, can_70883, can_1731_01 and can_1070) using Velvet; here the number of assembled bases was substantially below the expected assembly size of 4-5.5Mb (median size: 16004 bases, with a median of only six contigs in each assembly). Using A5-Miseq, the output was improved, generating reasonable assemblies with an appropriate median size of 5,143,908bp and 269 contigs. For these four isolates de novo-based typing approaches were therefore applied using the A5-MiSeq assemblies; Velvet assemblies were used for the other 211 isolates.

One hundred and five (49%) of the ST131 isolates were $bla_{\text{CTX-M}}$ positive, with allelic variants in order of frequency as follows: CTX-M-15 (74 isolates [70%]; although in two of these cases, Ecoli_M371 and CD358, deletions [37bp (63del99) and 2bp (414del416); and 1bp (275del276) in the blast matches were present), CTX-M-14 (20 isolates [19%]), CTX-M-27 (8 isolates [8%]), and one each of CTX-M-19, CTX-M-24 and CTX-M-55. For *fimH*, the most common variant was 30 (n=123; 57%), followed by 22 (n=24; 11%) and 41 (n=21; 10%); in 23 isolates, novel *fimH* variants appeared to be present, and in one isolate no blast match could be identified (*fimH* negative). There was strong association between clade and *fimH* allele, with 21/25

(84%) isolates in clade A being *fimH*41, 23/51 (45%) in clade B being *fimH*22, and 122/139 (88%) of isolates in clade C being *fimH*30 ($p<0.001$; Fisher's Exact Test).

Only five isolates in clade C (4 in C1, 1 in C2) did not have a double mutation in the chromosomal *gyrA* gene encoding for quinolone resistance. Strikingly, the relatively restricted association of CTX-M-15 within the C2 lineage observed in the Price *et al* and Petty *et al* papers(13, 14) was also observed in this dataset, with only the sporadic presence of CTX-M-15 containing isolates in other lineages. The presence of CTX-M-14, another commonly identified clinical variant was clearly also clustered within lineages, although the numbers were much smaller and these clusters were more distributed across diverse branches of the tree. Overall, the presence of CTX-M in any lineage appears to be strongly associated with clades with a TMRCA within the last 25 years, which fits epidemiologically with the introduction and increasing use of third generation cephalosporins.

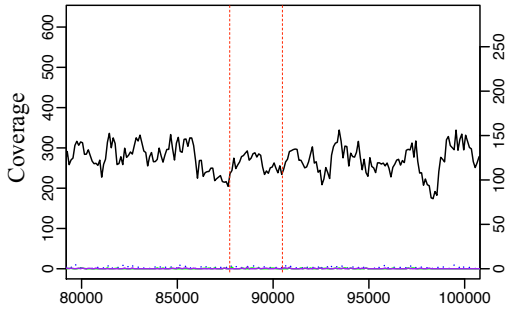
For the 74 isolates containing CTX-M-15, there were four isolates in which blast hits occurred on two different contigs (although in two isolates one of the hits was to a contig that was shorter than the length of the resistance gene itself), suggesting that these genes might be residing in duplicate in different genetic backgrounds within these isolates. Excluding the two very short contigs, there were 76 contigs in which to investigate the flanking context of the resistance gene, although the median contig length was short, at only 3,881bp (range: 1,382- 502,047bp). From the contig annotations, in eight cases these genes were clearly integrated into the host bacterial chromosome, with four unique integration events, and one stable integration event associated with evolution by descent over a period of approximately 10 years (Figure 8.4.). All eight of these events were associated with IS*Ecp1*. In 12 cases, $bla_{\text{CTX-M-15}}$

was clearly flanked by plasmid-associated sequences; in the remaining 56 cases it was not possible to define the location. IS*Ecp1* was associated with CTX-M-15 in 36/76 (47%) cases, although in several different contexts, including: (i) four separate contexts with IS*26*; (ii) two associations with phage-related elements; (iii) an association with a CAAX amino terminal protease (immunity against bacteriocins) and *pemI/K* genes (responsible for stable plasmid inheritance); (iv) two different contetxs in association with a short tract of nucleotide sequence annotated as a proQ/FinO domain (Figure 8.5.). In the other isolates it was not possible to assess the wider genetic flanking context of the $bla_{\text{CTX-M}}$ + IS*Ecp1* unit because of the short contig lengths. However, the variety of genetic contexts observed in the subset which assembled into long enough structures, already suggests that the CTX-M-15 gene has been mobilised into the ST131 C2 sub-lineage on a number of occasions, or has moved between plasmids within the sub-lineage. CTX-M-55 (a single amino acid different [Ala80Val] from CTX-M-15), present in one isolate (la_12107_3), was also identified on a short contig (2309bp), with an upstream truncated IS*Ecp1*/IS*26* complex, and a downstream ORF477 sequence (Figure 8.6.).

For the 30 CTX-M-14/14-like variants (CTX-M-19, -24, -27) three cases were likely chromosomally integrated (all CTX-M-14) in two locations, six plasmid-associated, and the rest uncertain. All cases were associated with the presence of IS*Ecp1* upstream of the gene, and in all CTX-M-27 cases this was disrupted by IS*26*, albeit in at least three different locations, suggestive of three separate mobilisation events. All genes also had varying lengths of IS*903* or IS*903*-like elements located immediately downstream of the resistance gene. Genetic contexts were broadly congruent with clustering on the host strain phylogenetic tree, with the most variability observed in the CTX-M-27/CTX-M-24 cluster containing la_10222 (Figure 8.7.).

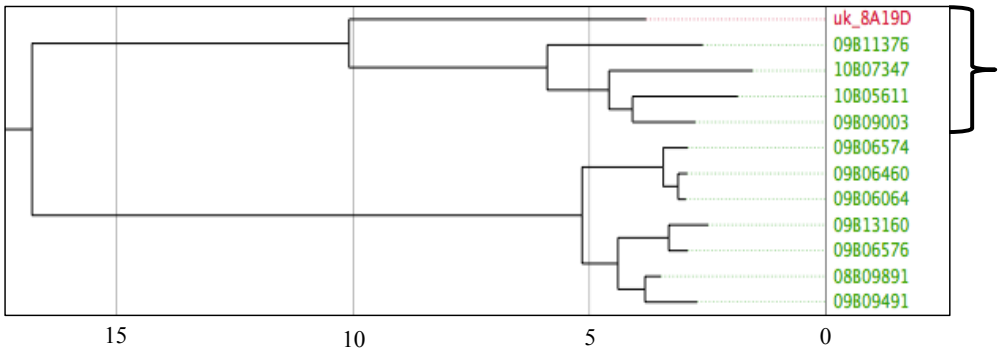
uk_8A19D

Shared by: 09B11376, 10B07347, 10B05611, 09B09003



Remapping of reads to this de novo assembly gives even coverage with no evidence of read mispairing, consistent with this structure being present (shown for 10B05611 in diagram, same for others). Position of ISEcp1+CTX-M-15 element bounded by red dashed lines

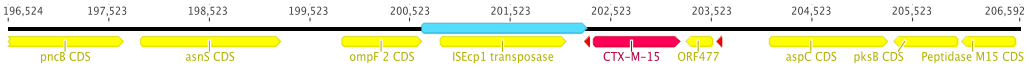
Extracted region from main ST131 phylogeny showing clustering of these isolates on the tree. The chromosomal integration event is dated to approximately 2002, and is subsequently found in ancestors cultured in the UK and in Thailand. Bracketed isolates show those with chromosomal integration of CTX-M-15



Other chromosomal integration events – "zoom-in" snapshots of relevant regions

JJ2444 (contig of 151,467bp)

KN1604 (contig of 278,281bp)



09B06064 (contig of 503,047bp)

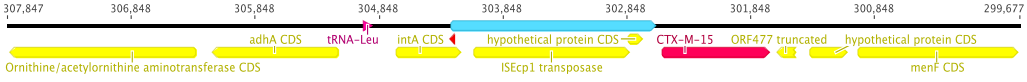
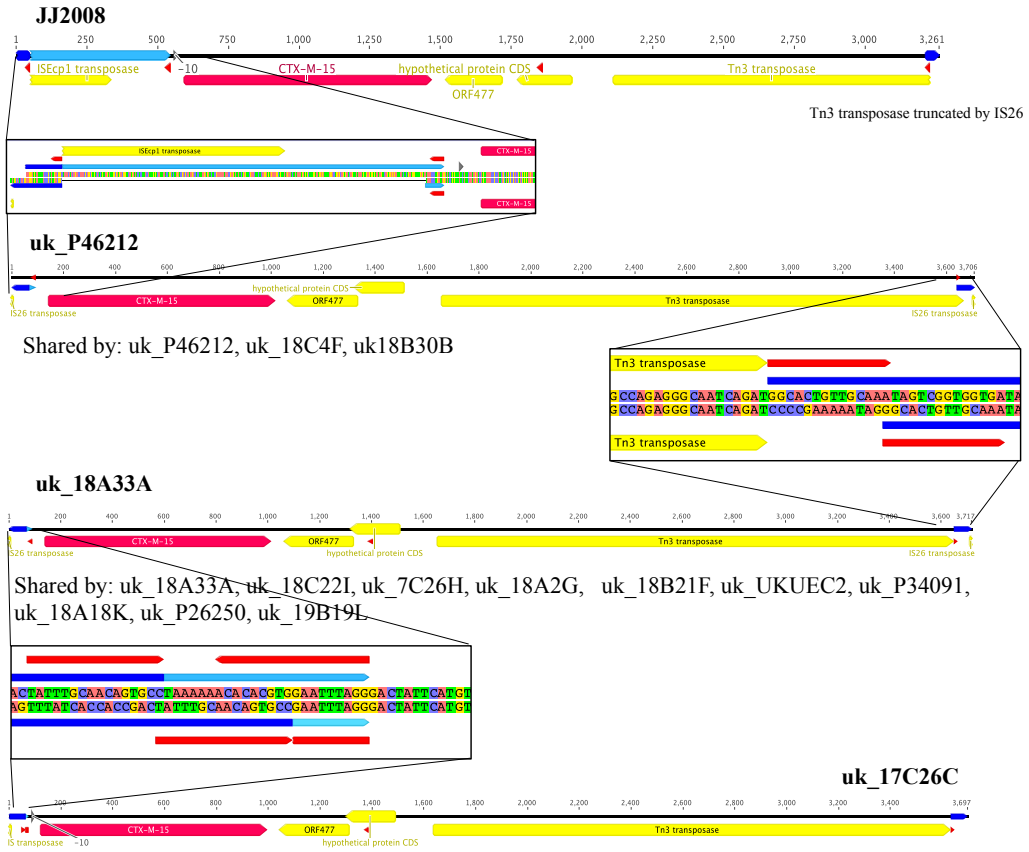


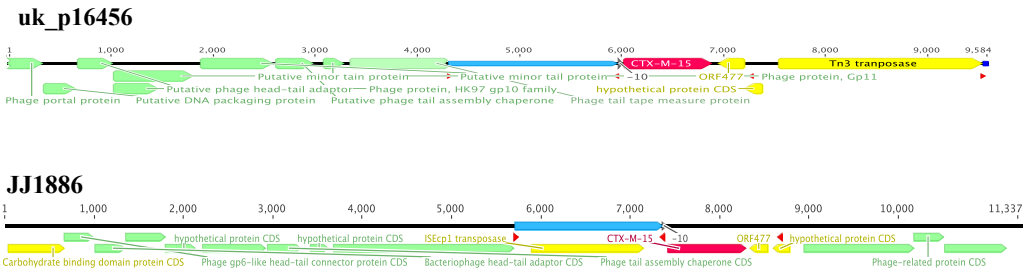
Figure 8.4. ST131 chromosomal integration events for CTX-M-15. The upper panel shows a likely stable integration event approximately 10 years ago, with evolution by descent and subsequent presence in strains isolated in Thailand and Oxford.

Figure 8.5. IS*Ecp1*-associated integration events in CTX-M-15 ST131. Inset boxes show "zoom-in" views of sequence alignments between pairs of representative sequences for each genetic context. Light blue arrows represent IS*Ecp1*-like elements; dark blue arrows IS*26*-like elements; thin red arrows inverted repeats for the corresponding IS; pale green elements are phage-associated. Other coding sequences (in yellow) are labelled. Scale varies depending on the length of the sequence assembled/the differentiating feature of interest.

IS*Ecp*1/IS*26*



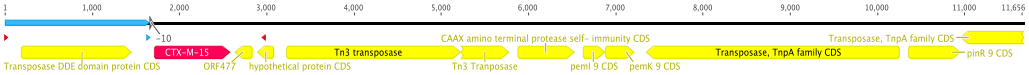
IS*Ecp*1/Phage-related elements



Shared by: JJ1887

IS*Ecp*1/CAAX aminoterminal protease motif + *pem* genes

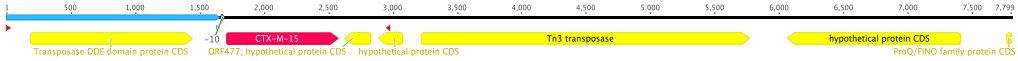
U004



Shared by: U004, uk_18B28B, uk_17B26A, uk_19A21D, uk_18B11D, uk_19B17I, uk_18B18D, C1353, JJ2489, JJ2643, uk_17A7A, uk_18C23A

IS*Ecp*1/ProQ/FINO family proteins

11B00320



JJ2591

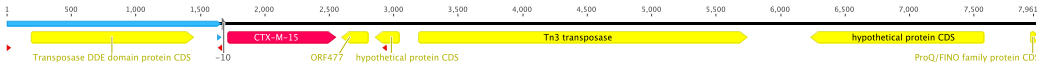
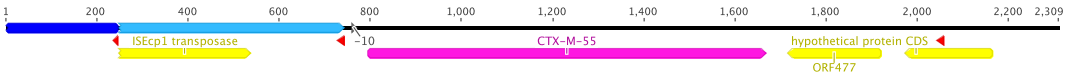


Figure 8.6. IS*Ecp1*/IS*26* associated integration in the single CTX-M-55 ST131 isolate



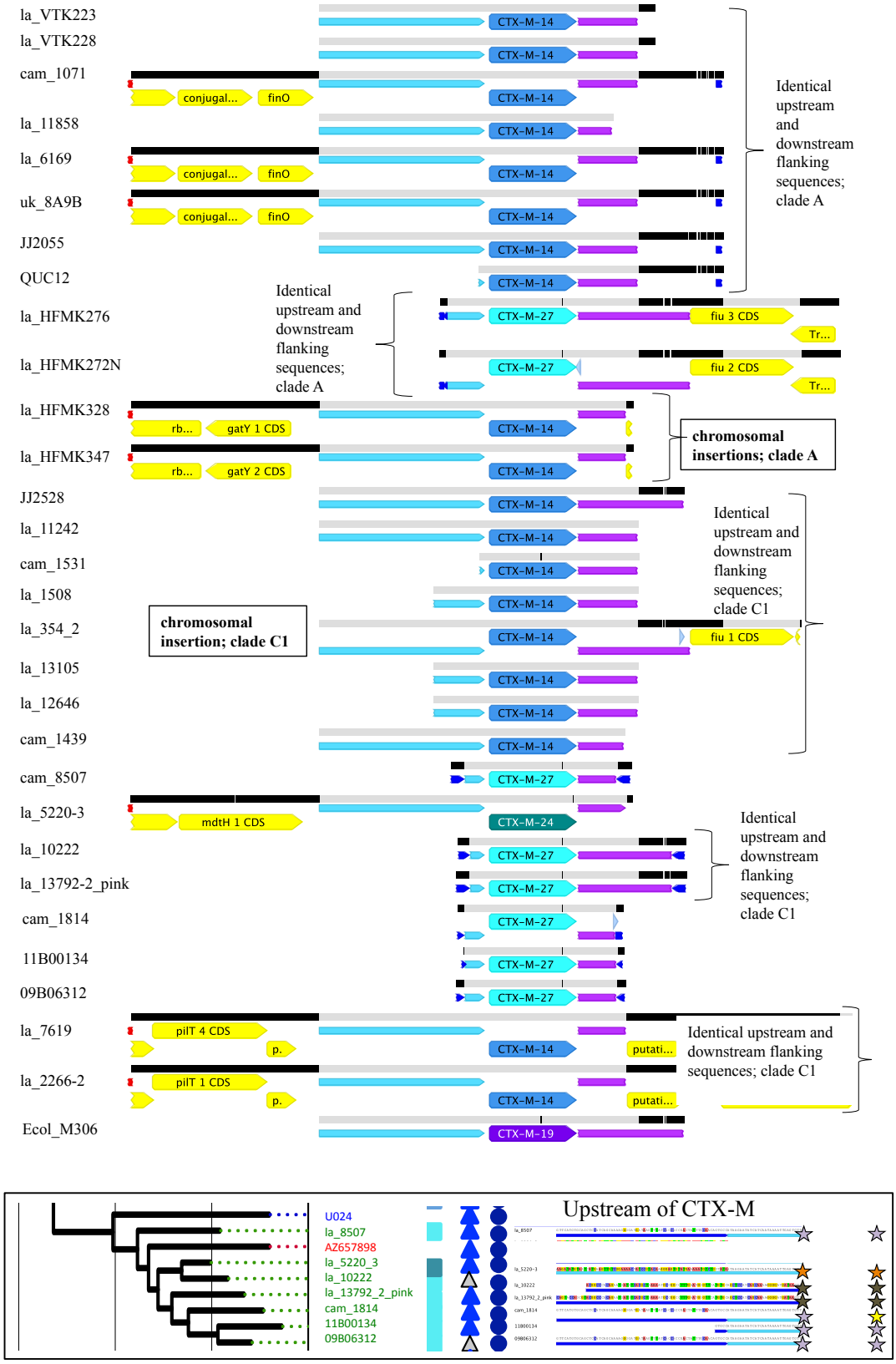


Figure 8.7. Genetic contexts for CTX-M-14 and 14-like (CTX-M-19, -24 and -27) ST-131 isolates. Isolates are ordered with respect to the host strain phylogeny (Figure 8.3.). Light blue annotations represent IS*Ecp1*-like sequences, dark blue IS*26*-like sequences, and pale purple IS*903*-like sequences. The "zoom-in" panel at the bottom demonstrates the incongruence of flanking context with the host strain phylogeny. Similar coloured stars represent similar flanking sequences.

8.3.3. Plasmid transformant analysis

Sequence data were successfully generated on 28 transformed plasmids (labelled with "T" on Figure 8.3.): (i) four in clade A, in association with CTX-M-15 (n=1), CTX-M-14 (n=2) and CTX-M-27 (n=1); (ii) one in clade B, in association with CTX-M-55; (iii) three in clade C1, in association with CTX-M-14 (n=2), CTX-M-24 (n=1); and (iii) 20 in association with CTX-M-15 in clade C2. The median plasmid assembly size was 122,786 (range: 72,449-171,919), with a median of 22 contigs in each assembly (range: 1-33). Using the longer reads (300bp) on the MiSeq platform resulted in a significantly smaller number of contigs per assembly (median 17 versus 25; ranksum p=0.003). Using mapping to assess the reliability of our plasmid constructs, almost all cases appeared to reliably reflect most of the content present in each transformed resistance plasmid (Figure 8.8.).

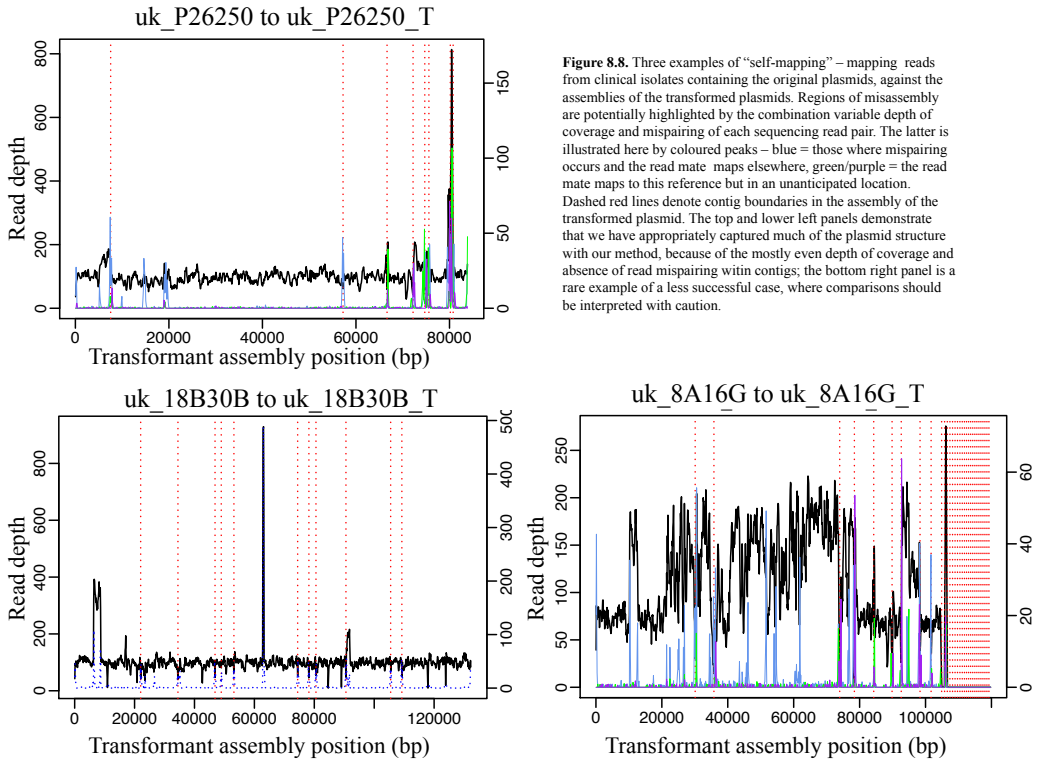


Figure 8.8. Three examples of "self-mapping" – mapping reads from clinical isolates containing the original plasmids, against the assemblies of the transformed plasmids. Regions of misassembly are potentially highlighted by the combination variable depth of coverage and mispairing of each sequencing read pair. The latter is illustrated here by coloured peaks – blue = those where mispairing occurs and the read mate maps elsewhere, green/purple = the read mate maps to this reference but in an unanticipated location. Dashed red lines denote contig boundaries in the assembly of the transformed plasmid. The top and lower left panels demonstrate that we have appropriately captured much of the plasmid structure with our method, because of the mostly even depth of coverage and absence of read mispairing witin contigs; the bottom right panel is a rare example of a less successful case, where comparisons should be interpreted with caution.

Using the transformants as references to make comparisons across the whole dataset, a number of different patterns emerged. CTX-M-55, only found in la_12107_3, appeared associated with a distinct single resistance plasmid acquisition event, because no other isolate in the dataset contained tracts of much more than 10kB with sequence homology >95%, except for isolate la_12107_3 itself (Figure 8.9.). This suggests that acquisition of either a common CTX-M-15 plasmid and evolution of the gene (CTX-M-55 differs from CTX-M-15 by a single nucleotide [C239G]) within this plasmid context, or importation of a CTX-M-55/IS element into a widely distributed plasmid in the lineage, would be implausible. (Figure 8.9.).

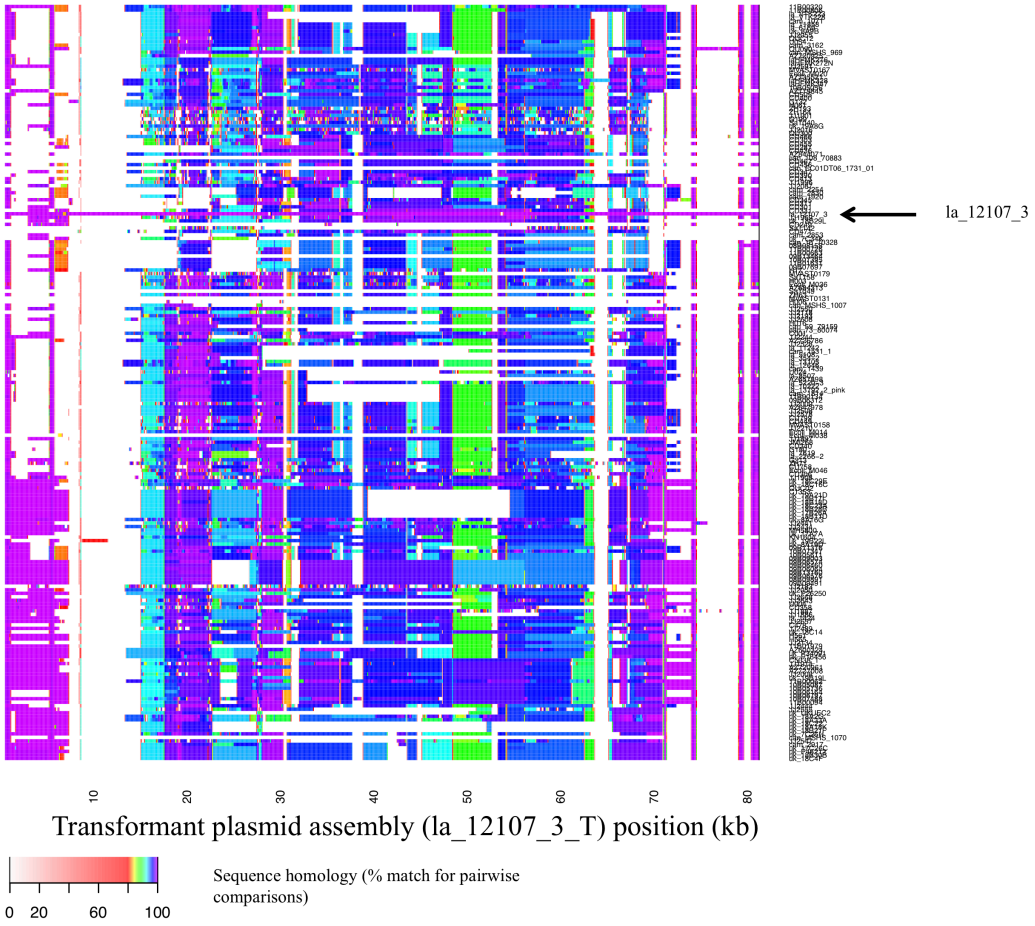


Figure 8.9. Heatmap showing ST131 % homology by BLASTn to the la_12107_3 transformant plasmid reference (x-axis). Isolates are ordered to correspond to the phylogeny depicted in Figure 8.3.

In the case of CTX-M-24, its host strain, la_5220-3, clusters with a group of CTX-M-14/27 positive isolates, and therefore this resistance gene variant could conceivably have evolved within a largely stable plasmid vector acquired by an ancestor of this group, with subsequent resistance gene/plasmid loss events for U024 and AZ65798 (Figure 8.3. and lower panel in Figure 8.7.). However, the transformant-level data generated from this cluster discounts this as a single hypothesis and reveals a number of interesting features. Firstly, for la_5220-3 (plasmid transformant assembly size: 72,449bp), it confirms that its CTX-M-24 resistance plasmid structure is distinct from other CTX-M-14-like plasmids in isolates both co-located on the host-strain phylogeny, as well as across the dataset (Figure 8.10.; data shown only for isolates containing CTX-M-14-like genes, wider dataset comparison not shown, but similar representation to Figure 8.9. above).

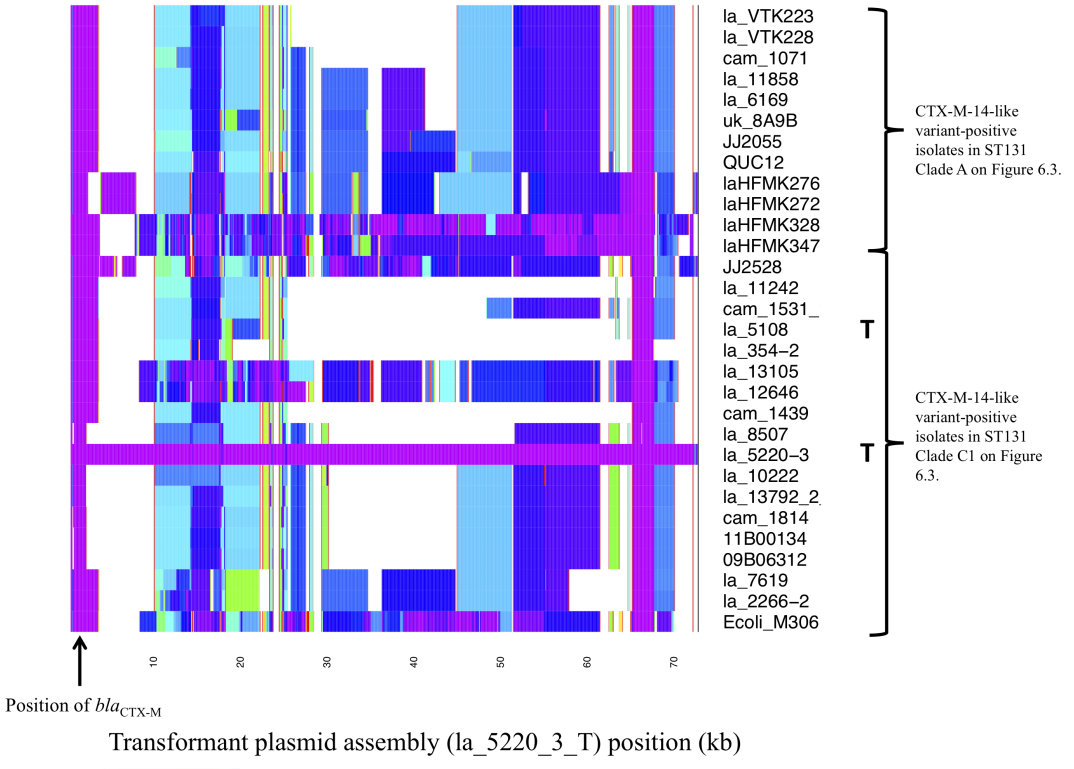


Figure 8.10. Heatmap showing ST131 % homology by blast to the la_5220_3 transformant reference (sequence represented along the x-axis). Isolates are ordered to correspond to the phylogeny depicted in Figure 8.3.; however, only those containing CTX-M-14/19/24/27 variants are included.

Historical acquisition of a single ancestral CTX-M-14 plasmid and its evolution within the cluster containing la_5108, in some cases with a number of large loss/gain events, remains a plausible hypothesis for cam_1531, la_13105, and la_12646, cam_1439, and all members of the CTX-M-27 group (Figure 8.11.). The CTX-M-negative isolates in the cluster (Figure 8.11.) also contain large stretches of sequence with ~100% homology to the la_5108 CTX-M-14 plasmid transformant. Isolates la_354_2 and la_11242 have most likely acquired different CTX-M-14 containing plasmids, because they share only 50-60% of the la_5108_T transformed plasmid sequence, but I have no transformant level information on these isolates to confirm this. Interestingly, although la_5220-3 was shown to have its own distinct CTX-M-24 plasmid, it also appears to contain tracts of sequence that are near identical to la_5108 (6.11.). Interestingly, extending the comparison to the transformed la_5108_T plasmid across the whole dataset also demonstrates that a la_5108_T plasmid-like structure(s) might exist in many of the clade A and C1 isolates (where the other CTX-M-14-like positive clusters of isolates have been found in this dataset); is almost entirely unrepresented in clade B (which is almost uniformly CTX-M negative, except for the aforementioned CTX-M-55 positive isolate, la_12107_3); and is partly represented in isolates in the clade C2 group (Fig. 6.12). Although there are limitations in this method, particularly with respect to confirming contiguity of sequence, the results suggest that plasmid population sub-structures in sub-lineages of ST131 may be related to the presence/absence of CTX-M, and which CTX-M-variant is represented.

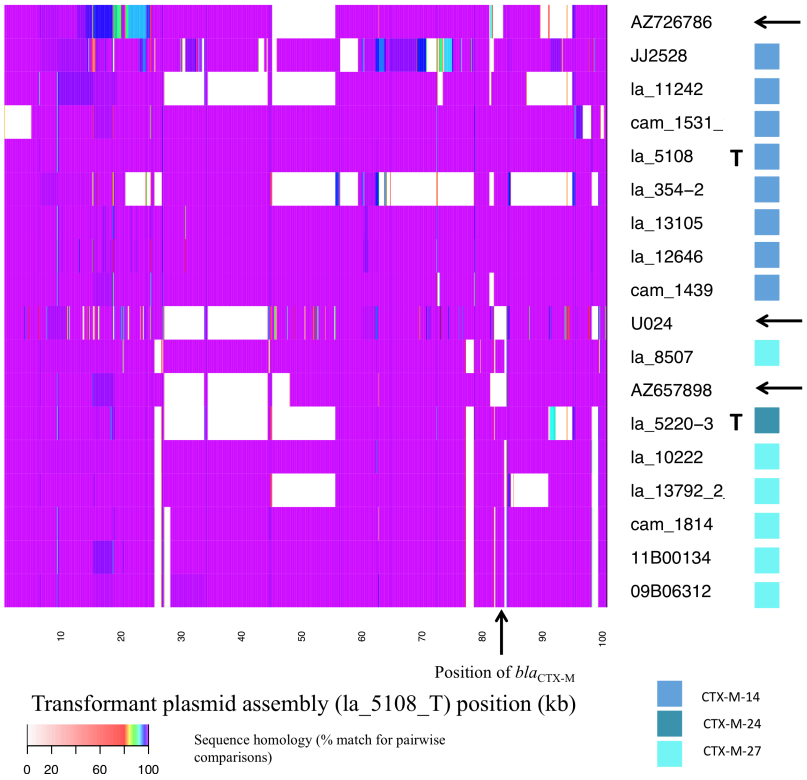


Figure 8.11. Heatmap showing ST131 % homology by blast to the la_5108 transformant reference (sequence represented along the x-axis). Isolates are ordered to correspond to the phylogeny depicted in Figure 8.3.; only isolates in the clade C1 cluster containing CTX-M-14 like variants are presented. CTX-M type is labelled with coloured squares as previous, black arrows denote CTX-M-negative isolates within the cluster.

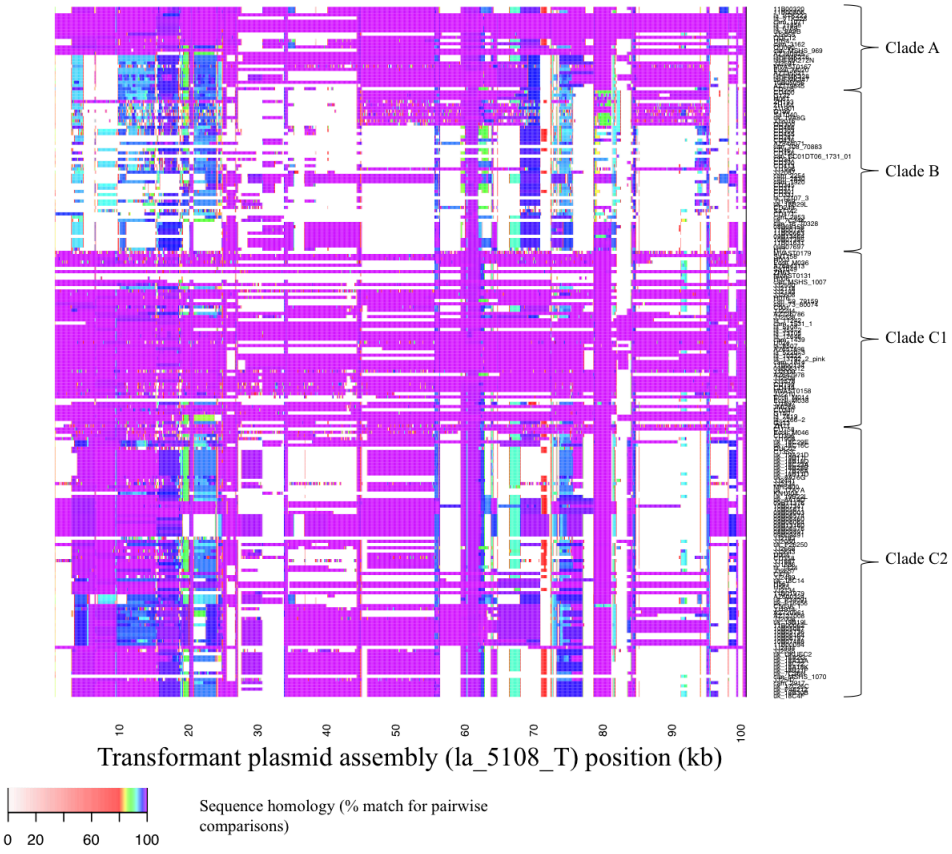


Figure 8.12. Heatmap showing ST131 % homology by blast to the la_5108 transformant reference (sequence represented along the x-axis). Isolates are ordered to correspond to the phylogeny depicted in Figure 8.3.; ST131 host-strain clades are delineated by the brackets on the right of the figure.

Given the difficulty in summarising a large number of these heatmap-based comparisons, I adopted a different approach to representing the extent of shared plasmid homology across the phylogeny, using all the transformants as a proxy measure for the CTX-M plasmid diversity seen across the sampled lineage, and including the two non-ST131 CTX-M-15 transformants as a marker for the extent of diversity one might expect to see between major *E. coli* lineages. This was done by comparing the mean percentage sequence homology determined by pairwise blastn comparisons between every transformant pair (taking each as a reference in turn), and plotting these against the time to most recent common ancestor (TMRCA) from the host strain phylogeny (Figure 8.13a.). For the non-ST131 transformants, the TMRCA was arbitrarily set at 200 years, as the most historical divergence time for ST131 was determined as being approximately 150 years ago (Figure 8.3.).

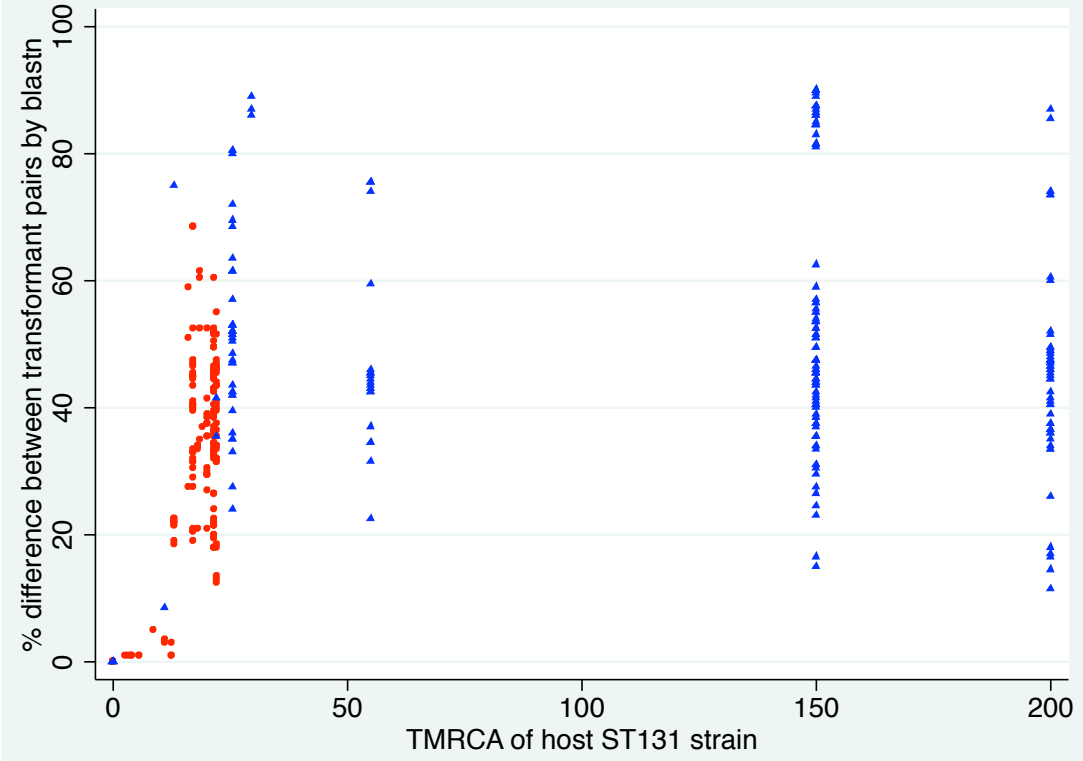


Figure 8.13a. Mean % sequence homology (from two pairwise comparisons, taking each transformant as a reference in turn) for each pair of ST131 transformants (TMRCA≤150 years) or compared with two non-ST131 transformants (TMRCA=200 years). Each point represents a comparison, red markers represent those within clade C2 (H30-Rx) and blue markers other comparison pairs.

This demonstrates that all CTX-M transformant plasmids shared at least 10% homology (% difference range: 0-90%), and that even plasmids found between lineages could be very similar (~80-85% sequence homology), suggesting that: (i) between lineage transfer of these resistance plasmids is highly likely; and (ii) plasmid similarity needs to be considered at best weak evidence for evolution by descent if sequences are not more than 80-85% similar. Finally, even within the C2 sub-lineage, relatively large amounts of sequence diversity were observed (as low as 30% homology).

A more detailed representation of comparisons with a TMRCA of ≤30 is shown in Figure 8.13b.

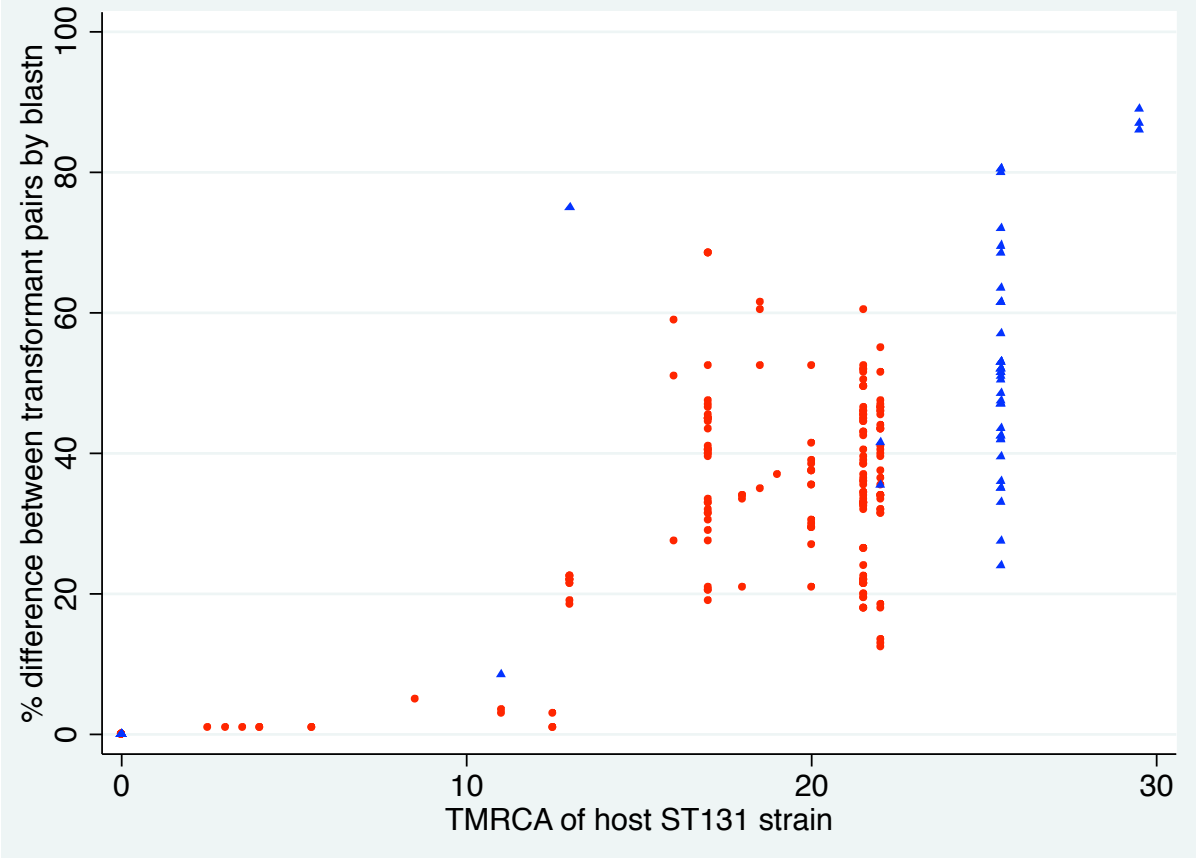


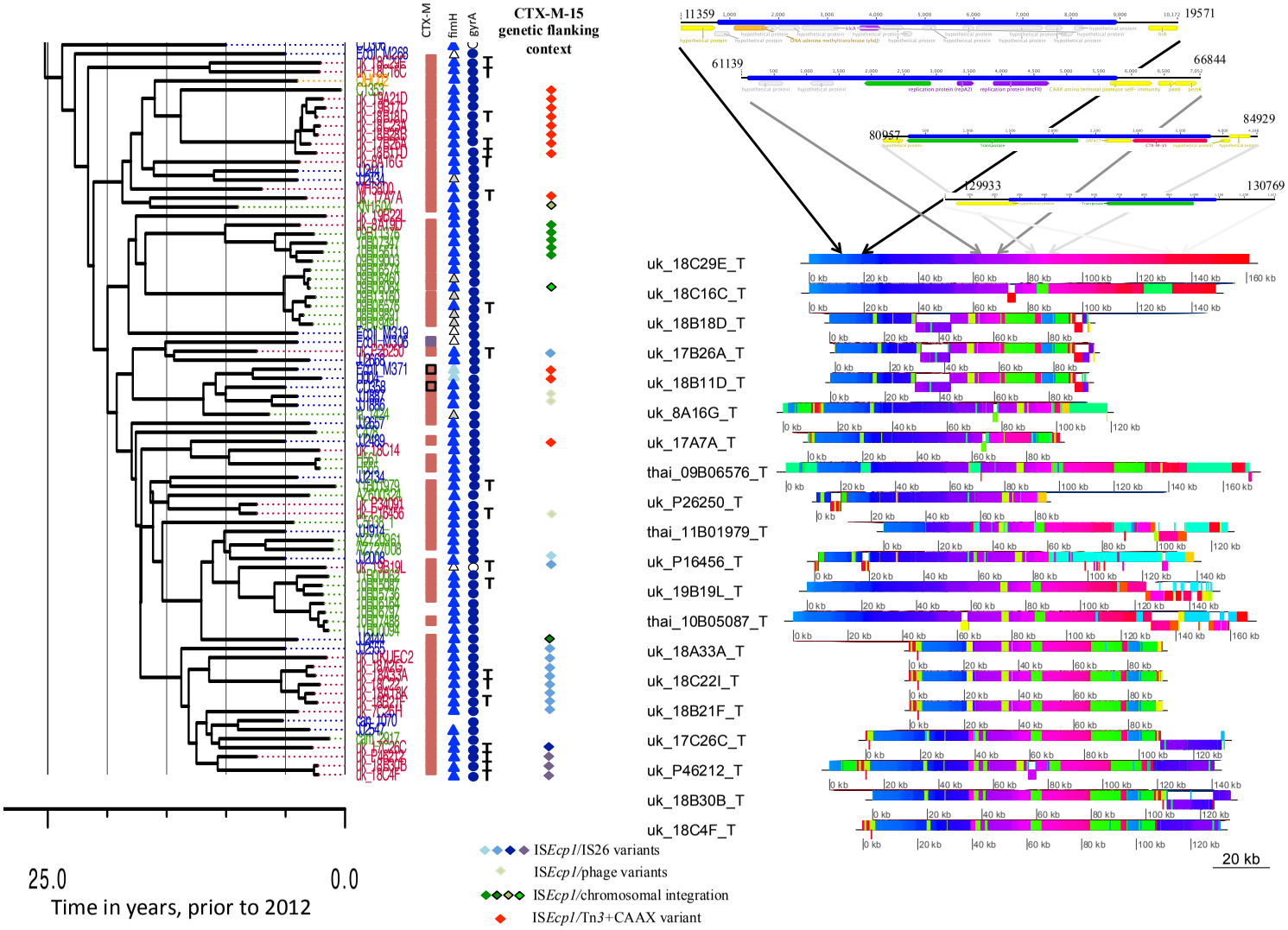
Figure 8.13b. Mean % sequence homology (from two pairwise comparisons, taking each transformant as a reference in turn) for each pair of ST131 transformants (TMRCA≤30 years). Each point represents a comparison, red markers represent those within clade C2 (H30-Rx) and blue markers other comparison pairs. Self-self "pairs" uniformly cluster at 0.

This shows that there is evidence for marked similarity between pairs of transformed plasmids with host strains having a TMRCA of up to 13 years, possibly representing acquisition and then evolution within a host, and after this time-point sequence divergence becomes more marked.

I subsequently focused specifically on clade C2 (H30-Rx), to investigate whether a single CTX-M-15 containing plasmid was acquired at the time of the emergence of the clade, perhaps being responsible for its success. CDHit was used to cluster similar coding sequences within all the CTX-M-15 transformed plasmids in this clade. CTX-M-15 was found in all clusters, acting as a proxy control for the adequacy of the method. In addition, across the whole cluster, containing 20 transformants, only 15 additional shared sequences were found, including those encoding for: six hypothetical proteins, two transposases, a glucose-1-phosphatase-like-enzyme, a dimethyl adenosine transferase DNA methyltransferase (yhdJ), an antirestriction protein (klcA), a DNA transport protein (traD), a tRNA-specific endonuclease (vapC), and antitoxin (vapB) and a replication protein. Comparisons of host strain, gene, genetic flanking region and plasmid transformant context for the C2 clade is summarised in Figure 8.14. Clearly some information is missing, since I have not transformed all the plasmids across the sub-lineage and have been unable to define the genetic flanking regions of many strains because of limitations resulting of d*e novo* assemblies. The diversity in plasmid transformants in terms of length and sequence homology is represented in the right-hand panel (all the contigs making up the transformants have been first arbitrarily ordered with respect to an independent, fully sequence CTX-M-15 plasmid, pEK499 [RefSeq: NC_013122.1]), demonstrating that major genetic rearrangement and gain/loss would have had to occur for there to have

been a single CTX-M-15 plasmid acquisition event and evolution of that structure within the C2 clade. Although incomplete, the overlay of the data relating to the genetic flanking context supports the notion that there is further sub-structure within the C2 clade, with possible IS*Ecp1* + Tn*3*/CAAX associated CTX-M-15 acquisition and transfer between plasmid populations in isolates with TMRCA clustering at the top of the Figure 8.14, and IS*Ecp1*/IS*26* associated transposition in isolates at the bottom of the figure. It is however difficult to confirm this given the limitations of data.

Figure 8.14. Host strain, CTX-M-15 genetic flanking context and plasmid transformant comparisons for the C2 sub-lineage. Panel on the left is a "zoom-in" depiction of the C2 host-strain phylogeny from Figure 8.3, with summary of genetic flanking contexts as colored diamond shapes to the right. Right panel represents progressiveMauve/GenoplotR alignment of ordered plasmid transformant contigs with coloured blocks representing sequence homology between plasmids. Shared sequences across all 20 transformants are represented at the top of the panel.



8.4. DISCUSSION

This study has confirmed that the ST131 lineage has a distinct population sub-structure, consisting of four sub-lineages, A, B, C1 and C2, as defined recently in 2014(14), and that this population sub-structure persists across a large isolate collection that represents sampling from four major geographical regions, namely North America (including Canada and the US), Europe, Australasia, and South-East Asia, and over four-and-a-half decades. The strong association of clades C1 and C2 with quinolone resistance, mediated by dual mutations in the quinolone-resistance determining region of *gyrA*, and of clade C2 with CTX-M are also evident.

I have calculated a molecular clock for the ST131 lineage at ~1 mutation/genome/year, which would be consistent with what has been previously published for non-enteropathogenic *E. coli* by investigating longitudinal evolution within a clustered set of infections(24). The time-scaled ST131 phylogeny dates the emergence of the C1 and C2 sub-lineages to around the time that clinical use of fluoroquinolones and third generation cephalosporins such as ceftriaxone became widespread (ciprofloxacin was licenced in 1988, and ceftriaxone was marketed globally by Roche in the early 1980s, coming off patent in 1999). This provides additional evidence that the emergence of these sub-lineages could be attributed to the presence of antibiotic selection pressures. These data have also shown for the first time that the presence of CTX-M-14 and CTX-M-14 like variants is also apparently associated with distinct host sub-lineages, across geographic locations. These have not become as predominant as CTX-M-15, but also exist in sub-lineages with TMRCAs estimated as occurring around 1997. Interestingly, clade B isolates almost invariably lack CTX-M genes.

A number of different genetic contexts for CTX-M genes was observed across the dataset, although these were broadly-speaking all consistent with previously described common associations with IS*Ecp1*, IS*26* and IS*903*(13, 25). The strong association with IS*Ecp1* may have explained the success of the gene in this lineage, as IS*Ecp1* is able to recognise a range of different repeats for insertion, operates via a one-ended transposition process, and enhances CTX-M gene expression. The fine-scale variability I observed in these flanking regions is however novel, and suggests that on-going transposition in the C2 sub-lineage, with loss and gain of short stretches of repeat sequences has occurred, either within the sub-lineage itself, or as a result of repeated introduction events from external lineages, either within the ST131 group, or from the wider species. The *de novo* assemblies were a major limitation of short-read sequencing data meaning I could only investigate a limited region around the resistance genes of interest. Nevertheless, for CTX-M-15, comparisons of the host strain phylogeny, flanking sequence around the CTX-M-15 gene, and transformant-derived plasmid sequences, the data would support the hypothesis of at least two acquisitions of the gene. The first would be in association with a CAAX amino terminal protease/Tn*3* unit, and the second in association with an IS*Ecp1*/IS26 structure. There are also potentially a handful of introductions by phages.

Chromosomal integration appears to have been mediated by one-ended IS*Ecp1* transposition in all CTX-M-15 cases, and was alarmingly shown on one of five occasions to have been likely stably inherited over ten years, and across geographic locations. The clinical implications of the stable chromosomal presence of CTX-M-15 in a fluoroquinolone-resistant lineage are substantial, particularly when that lineage is

C2-ST131 and represents a sub-group of the most significant current disease-causing lineages(1), and is itself associated with an increased risk of severe sepsis(13).

From the transformant-derived plasmid sequences themselves, and the proxy BLAST-based comparisons of the whole dataset against these plasmid references, it is evident that substantial plasmid diversity exists both within the lineage, and amongst plasmids containing the same CTX-M variant, although there is the suggestion that some of these plasmids can be acquired and stably inherited over at least a decade. Comparing all isolates in the dataset with all plasmid references also suggests that the different sub-lineages have plasmid population sub-structuring (shown in Figure 8.12), although I have not found an optimal method to summarise and represent these comparisons.

The limitations of this study partly reflect the failure to clearly assemble chromosomal and plasmid structures in all isolates using short-read Illumina data. Although I am now exploring the use of longer-read sequences to better define the plasmid content of drug-resistant Gram-negative bacilli through a recent collaboration with colleagues at Icahn School of Medicine at Mt Sinai in New York, the cost of long-read sequencing a dataset of this size remains prohibitive at this stage. Even were it possible to completely reconstruct all the genetic sequences involved, there are currently also a number of difficulties in alignment, phylogenetic and comparative genomic approaches that make comparisons between highly fluid, frequently re-assorting structures where evolutionary events may be represented not just by mutation but by variable insertions, deletions and genetic rearrangements, extremely difficult. Nevertheless, this remains an exciting challenge for the future.

CHAPTER 8 REFERENCES

1. **Rogers BA, Sidjabat HE, Paterson DL.** 2011. Escherichia coli O25b-ST131: a pandemic, multiresistant, community-associated strain. The Journal of antimicrobial chemotherapy **66:**1-14.
2. **Woodford N, Turton JF, Livermore DM.** 2011. Multiresistant Gram-negative bacteria: the role of high-risk clones in the dissemination of antibiotic resistance. FEMS microbiology reviews **35:**736-755.
3. **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 beta-lactamase CTX-M-15. Emerging infectious diseases **14:**195-200.
4. **Nicolas-Chanoine MH, Blanco J, Leflon-Guibout V, Demarty R, Alonso MP, Canica MM, Park YJ, Lavigne JP, Pitout J, Johnson JR.** 2008. Intercontinental emergence of Escherichia coli clone O25:H4-ST131 producing CTX-M-15. The Journal of antimicrobial chemotherapy **61:**273-281.
5. **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 : an official publication of the Infectious Diseases Society of America **51:**286-294.
6. **Johnson JR, Tchesnokova V, Johnston B, Clabots C, Roberts PL, Billig M, Riddell K, Rogers P, Qin X, Butler-Wu S, Price LB, Aziz M, Nicolas-Chanoine MH, Debroy C, Robicsek A, Hansen G, Urban C, Platell J,**

Trott DJ, Zhanel G, Weissman SJ, Cookson BT, Fang FC, Limaye AP, Scholes D, Chattopadhyay S, Hooper DC, Sokurenko EV. 2013. Abrupt emergence of a single dominant multidrug-resistant strain of Escherichia coli. The Journal of infectious diseases **207:**919-928.

7. **Weissman SJ, Johnson JR, Tchesnokova V, Billig M, Dykhuizen D, Riddell K, Rogers P, Qin X, Butler-Wu S, Cookson BT, Fang FC, Scholes D, Chattopadhyay S, Sokurenko E.** 2012. High-resolution two-locus clonal typing of extraintestinal pathogenic Escherichia coli. Applied and environmental microbiology **78:**1353-1360.
8. **Christiansen N, Nielsen L, Jakobsen L, Stegger M, Hansen LH, Frimodt-Moller N.** 2011. Fluoroquinolone resistance mechanisms in urinary tract pathogenic Escherichia coli isolated during rapidly increasing fluoroquinolone consumption in a low-use country. Microbial drug resistance **17:**395-406.
9. **Cullen IM, Manecksha RP, McCullagh E, Ahmad S, O'Kelly F, Flynn RJ, McDermott T, Murphy P, Grainger R, Fennell JP, Thornhill JA.** 2012. The changing pattern of antimicrobial resistance within 42,033 Escherichia coli isolates from nosocomial, community and urology patient-specific urinary tract infections, Dublin, 1999-2009. BJU international **109:**1198-1206.
10. **Canton R, Gonzalez-Alba JM, Galan JC.** 2012. CTX-M Enzymes: Origin and Diffusion. Frontiers in microbiology **3:**110.
11. **Novais A, Pires J, Ferreira H, Costa L, Montenegro C, Vuotto C, Donelli G, Coque TM, Peixe L.** 2012. Characterization of globally spread Escherichia coli ST131 isolates (1991 to 2010). Antimicrobial agents and chemotherapy **56:**3973-3976.

12. **Naseer U, Sundsfjord A.** 2011. The CTX-M conundrum: dissemination of plasmids and Escherichia coli clones. Microbial drug resistance **17:**83-97.

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-beta-lactamase-producing Escherichia coli ST131 is driven by a single highly pathogenic subclone, H30-Rx. mBio **4:**e00377-00313.

14. **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-Bano 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. Proceedings of the National Academy of Sciences of the United States of America **111:**5694-5699.

15. **Jean SS, Hsueh PR.** 2011. High burden of antimicrobial resistance in Asia. International journal of antimicrobial agents **37:**291-295.

16. **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.** 2010. Complete genome sequence of the wild-type commensal Escherichia coli strain SE15, belonging to phylogenetic group B2. Journal of bacteriology **192:**1165-1166.

17. **Wirth T, Falush D, Lan R, Colles F, Mensa P, Wieler LH, Karch H, Reeves PR, Maiden MC, Ochman H, Achtman M.** 2006. Sex and virulence

in Escherichia coli: an evolutionary perspective. Molecular microbiology **60:**1136-1151.

18. **Didelot X, Meric G, Falush D, Darling AE.** 2012. Impact of homologous and non-homologous recombination in the genomic evolution of Escherichia coli. BMC genomics **13:**256.

19. **Didelot X, Falush D.** 2007. Inference of bacterial microevolution using multilocus sequence data. Genetics **175:**1251-1266.

20. **Minin VN, Bloomquist EW, Suchard MA.** 2008. Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. Molecular biology and evolution **25:**1459-1471.

21. **Drummond AJ, Ho SY, Phillips MJ, Rambaut A.** 2006. Relaxed phylogenetics and dating with confidence. PLoS biology **4:**e88.

22. **Baele G, Lemey P, Bedford T, Rambaut A, Suchard MA, Alekseyenko AV.** 2012. Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. Molecular biology and evolution **29:**2157-2167.

23. **Li W, Godzik A.** 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics **22:**1658-1659.

24. **Reeves PR, Liu B, Zhou Z, Li D, Guo D, Ren Y, Clabots C, Lan R, Johnson JR, Wang L.** 2011. Rates of mutation and host transmission for an Escherichia coli clone over 3 years. PloS one **6:**e26907.

25. **Poirel L, Naas T, Nordmann P.** 2008. Genetic support of extended-spectrum beta-lactamases. Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases **14 Suppl 1:**75-81.

CHAPTER 9: CONCLUSIONS AND FUTURE WORK

This thesis has contributed to the growing evidence that the evolution and transmission of antimicrobial resistance in *Escherichia coli* and *Klebsiella pneumoniae* has to be considered holistically in the context of the various potential genetic units involved, namely: (i) the resistance genes themselves; (ii) transposable elements surrounding these resistance genes; (iii) the resistance plasmids on which these elements reside; (iv) the bacterial strains hosting all of these genetic components. The structural diversity each of these genetic units in these two species is substantial, and detailed analysis of this in large datasets pushes the limits of what can currently be achieved using high throughput short-read based methods and analytical approaches.

The initial evaluation of the read processing methods used for the analysis of the data in chapter 4 has expanded on previous assessments of the robustness of our group's reference based mapping and nucleotide variant calling algorithm(1, 2). Benchmarking of bioinformatic pipelines used for the analysis of pathogen whole genome sequence data is not a typical feature of most published analyses, but is important in establishing error rates and identifying whether there are any obvious sources of bias. With genetically diverse organisms such as *E. coli* and *K. pneumoniae*, it is also important to realise that a significant proportion of the genetic data will be excluded from an analysis of core genomes depending on the dataset being analysed and the reference used. The unanalysed accessory component could however provide valuable additional insights into evolutionary pathways and

relationships between organisms – however, new methods need to be developed that will be able to include this in analyses.

The assessment of the robustness of de novo assemblies builds on previously published comparisons of *de novo* assemblers(3, 4), and is of particular value in that I was in a position to analyse sequence replicates of the same reference encompassing: (i) sequencing of the same DNA extract; (ii) sequencing of the same frozen isolate stock extracted on a number of occasions; (iii) sequencing the same reference across sequencing runs. Although I have not made a comprehensive assessment of all the various different parameterisations that are possible with the different assembly methods, opting instead to use the settings recommended by the authors/listed as the default (as many users will do), my comparison highlights that there is significant variability between assemblers in the degree and nature of error introduced, which may be important, depending on what questions are being asked of the data. In identifying genes being present/absent, for example, around 2% of genes will be completely missed by all assembly methods, which seems like a small number, but nevertheless represents ~100 coding sequences in these species. I have not investigated whether there is any systematic bias in the types of genes that are missed, but this could be of relevance. In terms of SNV-level variation, the lowest number of miscalled bases per 100kB of sequence across the analysis was 1, representing 53 erroneously called bases in the assembly. Again, a systematic analysis of whether these errors could be predicted and excluded was beyond the scope of this thesis, but this highlights the caution needed in using de novo assemblies for assessments of SNV-level variation for interpreting both phylogenetic and transmission-based

comparisons, yet these methods have been utilised in studies that are published in high-quality journals(5-7).

The analysis of the robustness of the approach by which I sequenced transformants in a known host *E. coli* background and assembled unmapped reads, is also novel, although again not totally comprehensive. This method was developed as a proxy in the absence of having access to long-read sequencing data at the time the analysis was being undertaken. Again this method evaluation highlights the limitations of short-read data in obtaining complete structural information for episomal structures, particularly in the context of repeat sequences that might be present in more than one copy, or present in both the structure of interest and the host background strain used for transformation.

Work undertaken in this thesis has facilitated a on-going collaboration with colleagues who run a PacBio sequencing facility, and this has given me the exciting opportunity to be involved in current and future work using this method to better define the chromosomal and episomal structures of significant disease-causing lineages and broad-spectrum antimicrobial resistance plasmids in Enterobacteriaceae. I am currently involved in projects using PacBio to create closed chromosomal and plasmid assemblies for: (i) several other CTX-M-positive disease-causing global lineages of *E. coli*, namely ST10, ST38, ST648 and ST405; (ii) an investigation of the emergence of KPC carbapenemase resistance in H30 ST131 E. coli, and (iii) the spread of KPC-containing plasmids, including the pKPC_UVA01 and pKPC_UVA02 plasmids described in this work, but across multiple species of Enterobacteriaceae. As more fully closed reference sequences are published, it will hopefully become

possible to implement some of our comparative methods and use the databank of nearly 2000 *E. coli* and *K. pneumoniae* sequences that I have been involved in collecting and sequencing for wider comparisons.

At the gene level, work in Chapter 5 has demonstrated that it is feasible to use whole genome sequence data for the identification of antimicrobial resistance genes, an approach which was shown in a proof-of-principle analysis to have value in accurately predicting antimicrobial susceptibility phenotype in a group of clinical *E. coli* and *K. pneumoniae* isolates with a diverse range of genotypic mechanisms underlying their susceptibility profiles(8). A similar approach has been shown to be feasible in a much larger study of resistance prediction in *Staphylococcus aureus* undertaken by one of my colleagues(9), and a modified approach based on using mapped data for tuberculosis (Walker, TM, unpublished data). Both of these approaches show that this method could be applied in a routine context to clinical samples, in the event that sequencing turnaround and analysis times could be streamlined. In order to validate this approach for Enterobacteriacaeae, I am currently involved in a project to validate our prediction approach on a much larger dataset of around 750 clinical isolates, including a set of 150 isolates with less commonly observed resistance mechanisms, obtained from our colleagues at the Antimicrobial Resistance Monitoring Reference Laboratory. One of my colleagues has developed a method to extract and sequence *E. coli* DNA directly from blood culture bottles(10), and the aspiration will be to implement a real-time pilot head-to-head with our diagnostic laboratory by April 2015.

With respect to transmission and the genetic population structure of NDM- and KPC-positive strains and plasmids, the work presented in Chapters 6 and 7 has identified a number of novel features. For NDM-1 in *K. pneumoniae,* we observed the almost completely stable propagation of single ST15 clone with a set of four plasmids totalling 700kB in size over a period of nearly a year, causing a series of case clusters in a region of high NDM prevalence. This is the first time an extended outbreak of NDM-1 *K. pneumoniae* cases has been investigated with high resolution, and in the context of adopting a wide sampling approach including other community and hospital-associated strains of several species. We identified a number of other host strain clusters for *K. pneumoniae*, *E. coli* and *Enterobacter cloacae* (data not shown here for the latter two species), suggesting additional community and hospital-based transmission, some in association with bla_{NDM}, but on different plasmid vectors (Stoesser N et al. - "Dynamics of multiple drug-resistant *Enterobacter cloacae* outbreaks in a neonatal unit in Nepal: Insights using wider sampling frames and next generation sequencing"; in submission to JAC, Sept 2014.) Despite adopting such a wide sampling approach, we were still unable to identify the source of the outbreak, although inference-based transmission analyses supported a number of direct and indirect transmission events, the latter conceivably occurring either through transmission between asymptomatic carriers or environmental contamination.

In the case of KPC-*K. pneumoniae* in a US hospital, we identified the widespread dissemination of a promiscuous KPC-containing plasmid (pKPC_UVA01) amongst diverse *K. pneumoniae* lineages, and the within-institution transmission predominance of a single lineage, ST941, in association with pKPC_UVA01 and a second novel plasmid (pKPC_UVA02). This contrasts with the current prevailing view that KPC is

predominantly host-restricted to a single lineage, ST258, and may represent one of the earliest, well-characterised examples of the "escape" of KPC from its historical association with ST258 and closely related members of this clonal complex. I am currently involved in a project to characterise the KPC gene, Tn*4401* transposon, plasmid and host-strain transmission within a collection of *K. pneumoniae* and a number of other Enterobacteriaceae (~400 isolates obtained from this institution over a 5 year period). Part of this analysis has involved the use of PacBio sequencing, which much more clearly resolves the contexts of bla_{KPC} and the episomal structures involved. Very recently published and impressive work from the NIH in Maryland has used a similar approach(11), but their isolate collection was limited, and the dynamics of within-institution transmission could not be well-characterised with such a small number of diagnosed cases.

The work in Chapter 8 represents the most difficult and least complete analysis in this thesis, and its novelty was unfortunately somewhat superseded by the similar, albeit smaller analyses that were published earlier this year(7, 12). Nevertheless, it confirms the population sub-structure identified in these studies in a wider sampling frame, as well as the striking association of fluoroquinolone resistance with sub-lineages C1 and C2, and of CTX-M-15 with C2. In addition, it includes a larger number of CTX-M-14-family positive genes, which are also mostly clustered within sub-lineages on the host-strain phylogeny. Unlike the two other studies, I have been able to date the emergence of the sub-lineages, and interestingly C1 and C2 appear to have emerged soon after the use of fluoroquinolones and third generation cephalosporins became widespread. I have also been able to estimate an evolutionary rate for the ST131 lineage, which confirms other smaller longitudinal in vivo estimates(13), and suggests

that hyper-mutability is not a feature of the global clinical "success" of this strain. I have also investigated the largest number of ST131 CTX-M-positive plasmids to date, and although the data is not conclusive, it suggests that acquisition of CTX-M-15 into C2 may have occurred on at least two different occasions, with subsequent within-lineage evolution, as opposed to a single acquisition event approximately 25 years ago. I have also demonstrated evidence of plasmid population sub-structure within the lineage, although further work in this area remains to be undertaken. Current studies are underway to obtain fully resolved resistance plasmid structures for some of the strains and to make comparisons with other non-ST131 lineages.

In conclusion, this thesis has developed and applied methods using whole genome sequence data in order to identify a number of novel aspects of the molecular epidemiology of broad-spectrum beta-lactam resistance in *Escherichia* coli and *Klebisella pneumoniae*. Although a large number of questions remain unresolved, sequencing technologies, computing infrastructures and analytical methods continue to evolve rapidly, and hold exciting promise for further defining the complicated molecular epidemiology of resistance in these species as they themselves change and adapt in response to the variable selection pressures imposed by the global environmental and clinical niches they inhabit.

CHAPTER 9 REFERENCES

1. **Eyre DW, Cule ML, Wilson DJ, Griffiths D, Vaughan A, O'Connor L, Ip CL, Golubchik T, Batty EM, Finney JM, Wyllie DH, Didelot X, Piazza P, Bowden R, Dingle KE, Harding RM, Crook DW, Wilcox MH, Peto TE, Walker AS.** 2013. Diverse sources of C. difficile infection identified on whole-genome sequencing. The New England journal of medicine **369:**1195-1205.
2. **Walker TM, Ip CL, Harrell RH, Evans JT, Kapatai G, Dedicoat MJ, Eyre DW, Wilson DJ, Hawkey PM, Crook DW, Parkhill J, Harris D, Walker AS, Bowden R, Monk P, Smith EG, Peto TE.** 2013. Whole-genome sequencing to delineate Mycobacterium tuberculosis outbreaks: a retrospective observational study. The Lancet infectious diseases **13:**137-146.
3. **Magoc T, Pabinger S, Canzar S, Liu X, Su Q, Puiu D, Tallon LJ, Salzberg SL.** 2013. GAGE-B: an evaluation of genome assemblers for bacterial organisms. Bioinformatics **29:**1718-1725.
4. **Salzberg SL, Phillippy AM, Zimin A, Puiu D, Magoc T, Koren S, Treangen TJ, Schatz MC, Delcher AL, Roberts M, Marcais G, Pop M, Yorke JA.** 2012. GAGE: A critical evaluation of genome assemblies and assembly algorithms. Genome research **22:**557-567.
5. **Wright MS, Perez F, Brinkac L, Jacobs MR, Kaye K, Cober E, van Duin D, Marshall SH, Hujer AM, Rudin SD, Hujer KM, Bonomo RA, Adams MD.** 2014. Population Structure of KPC-Producing Klebsiella pneumoniae Isolates from Midwestern U.S. Hospitals. Antimicrobial agents and chemotherapy **58:**4961-4965.

6. **Snitkin ES, Zelazny AM, Thomas PJ, Stock F, Group NCSP, Henderson DK, Palmore TN, Segre JA.** 2012. Tracking a hospital outbreak of carbapenem-resistant Klebsiella pneumoniae with whole-genome sequencing. Science translational medicine **4:**148ra116.

7. **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-Bano 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. Proceedings of the National Academy of Sciences of the United States of America **111:**5694-5699.

8. **Stoesser N, Batty EM, Eyre DW, Morgan M, Wyllie DH, Del Ojo Elias C, Johnson JR, Walker AS, Peto TE, Crook DW.** 2013. Predicting antimicrobial susceptibilities for Escherichia coli and Klebsiella pneumoniae isolates using whole genomic sequence data. The Journal of antimicrobial chemotherapy **68:**2234-2244.

9. **Gordon NC, Price JR, Cole K, Everitt R, Morgan M, Finney J, Kearns AM, Pichon B, Young B, Wilson DJ, Llewelyn MJ, Paul J, Peto TE, Crook DW, Walker AS, Golubchik T.** 2014. Prediction of Staphylococcus aureus antimicrobial resistance by whole-genome sequencing. Journal of clinical microbiology **52:**1182-1191.

10. **Anson LWP, L.J.; Votintseva, A.; Stoesser N.; Crook, D.W. .** 2014. Establishing the potential for whole genome sequencing of Escherichia coli directly from positive blood culture media, ECCMID, vol. P0535, Barcelona, Spain.

11. **Conlan S, Thomas PJ, Deming C, Park M, Lau AF, Dekker JP, Snitkin ES, Clark TA, Luong K, Song Y, Tsai YC, Boitano M, Dayal J, Brooks SY, Schmidt B, Young AC, Thomas JW, Bouffard GG, Blakesley RW, Program NCS, Mullikin JC, Korlach J, Henderson DK, Frank KM, Palmore TN, Segre JA.** 2014. Single-molecule sequencing to track plasmid diversity of hospital-associated carbapenemase-producing Enterobacteriaceae. Science translational medicine **6:**254ra126.

12. **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-beta-lactamase-producing Escherichia coli ST131 is driven by a single highly pathogenic subclone, H30-Rx. mBio **4:**e00377-00313.

13. **Reeves PR, Liu B, Zhou Z, Li D, Guo D, Ren Y, Clabots C, Lan R, Johnson JR, Wang L.** 2011. Rates of mutation and host transmission for an Escherichia coli clone over 3 years. PloS one **6:**e26907.

Appendix 1.

Table A1.1. β-Lactam resistance mechanisms included in the study resistance locus database (general references: 1-5)

Enzyme family	Likely source	Major sub-groups	Variants	Molecular class	Bush-Jacoby functional classification	Reference
bla_{A}	*Shewanella oneidensis*			D		6
bla_{ACC}	*Klebsiella pneumoniae*		ACC-1-4	C	1	7
bla_{ACI}	*Acidaminococcus fermentans*		ACI-1	A	2be	8, 9
bla_{ACT}	*Enterobacter cloacae*		ACT-1-7, ACT-9-10, ACT-12, ACT-14-16	C	1	10, 11
bla_{ADC}	*Acinetobacter baumannii*		ADC-1-7	C	1	12
bla_{AER}	*Aeromonas hydrophila*		AER-1	A	2c	13, 14
bla_{AST}	*Nocardia asteroides*		AST-1	A	2b	9, 15
bla_{B}	*Chryseobacterium meningosepticum*		B2-14	B		9, 16
bla_{BES}	*Serratia marcescens*		BES-1	A	2be	9, 17
bla_{BIC}	*Pseudomonas fluorescens*		BIC-1	A	2e	9, 18
bla_{BIL}	*Citrobacter freundii*		BIL-1 (CMY-2)	A	2ber	9, 19
bla_{BRO}	*Moraxella catarrhalis*		BRO-1	A	2a	20
bla_{CAM}	*Campylobacter* spp.		CAM-1 (bla_{OXA-61})	D		21
bla_{CARB}	*Pseudomonas aeruginosa, Vibrio cholera, Acinetobacter* spp.	CARB-1-like CARB-2-like CARB-5-like	CARB-1 (PSE-4), 4, 6, 7, 9, 11, 12 CARB-2 (PSE-1), 3 CARB-5 (RTG-2), 8, 10 (RTG-4)	A A	2c, 2ce	11, 14, 22-26
bla_{CBL}	*Bacteroides uniformis*		CBL-A	A	2b	27
bla_{CEPA}	*Bacteroides fragilis*		bla_{CEPA}, bla_{CEPA}-14, 29, 44, 49, 85	A	2e	11, 28
bla_{CFE}	*Citrobacter freundii*		CFE-1	C	1	9, 29

bla$_{CGA}$	*Chryseobacterium gleum*		CGA-1	A	2b	30
bla$_{CGB}$	*Chryseobacterium gleum*		CGB-1	B1	3a	31
bla$_{CKO}$	*Citrobacter koseri*		CKO-1	A	2be	32
bla$_{CIA}$	*Chryseobacterium indologenes*		CIA-1-4	A	2be	33
bla$_{CME}$	*Chryseobacterium meningosepticum*		CME-1, 2	A	2be	34, 35
bla$_{CMG}$	*Enterobacter cloacae*		CMG-1	C		36
bla$_{CMH}$	*Enterobacter cloacae*		CMH-1	C		37
bla$_{CMY}$	*Pseudomonas aeruginosa* *Citrobacter freundii*	CMY-1-like CMY-2-like	CMY-1, 8-11, 19 CMY-2-7, 12-19, 20-45, 47-49, 53-60, 64-68, 70-81, 83, 84, 87	C	1, 1e	11
bla$_{CSP}$	*Capnocytophaga sputigena*		CSP-1	A	2be	38
bla$_{CTX-M}$ (*bla*$_{UOE}$, *bla*$_{KLUC}$, *bla*$_{KLUG}$)	*Kluyvera* spp.	CTX-M-1-like CTX-M-2-like CTX-M-8-like CTX-M-9-like	CTX-M-1, 3, 10, 11, 12, 15, 22, 23, 28, 29, 30, 32, 33, 34, 36, 37, 42, 52, 53, 54, 55/57, 58, 60, 61, 62, 66, 68, 69, 71, 72, 79, 80, 82, 88, 96, 101, 114, 116, 117, 132, 133 CTX-M-2, 4, 5, 6, 7, 20, 31, 35, 43, 44 (TOHO-1), 56, 59, 74, 75, 92, 95, 97, 124, 131 CTX-M-8, 40, 63 CTX-M-9, 13, 14/18, 16, 17, 19, 21, 24, 27, 38, 45 (TOHO-2, 46, 47, 48, 49, 50, 51, 65, 67, 81, 83, 84, 85, 86, 87, 90, 93, 98, 99, 104, 105, 111, 112, 113, 121, 122, 126	A	2be	11, 39

		CTX-M-25-like Hybrids	CTX-M-25, 26, 39, 41, 89, 91, 94 CTX-M-64, CTX-M-123, CTX-M-78			
bla$_{DES}$	*Desulfovibrio desulfuricans*		DES-1	A	2be	40
bla$_{DHA}$	*Morganella morganii*		DHA-1, 2, 3, 5, 6, 7	C	1e	41, 42
bla$_{DIM}$	*Pseudomonas stutzeri*		DIM-1	B1	3a	43
bla$_{EBR}$	*Empedobacter brevis*		EBR-1	B1	3a	44
bla$_{ERP}$	*Erwinia persicina*		ERP-1	A	2be	45
bla$_{FAR}$	*Nocardia farcinica*		FAR-1	A	2br	46
bla$_{FIM}$	*Pseudomonas aeruginosa*		FIM-1	B1	3a	47
bla$_{FONA}$	*Serratia fonticola*		FONA-1-6	A	2be	48, 49
bla$_{FOX}$	*Pseudomonas aeruginosa*		FOX-1-9	C	1	11, 50
bla$_{FUS}$	*Fusobacterium nucleatum*		FUS-1 (OXA-58)	D	2d	51
bla$_{GES}$	*Klebsiella pneumoniae*		GES-1-22	A	2e, 2be, 2f	49, 52
bla$_{GIM}$	*Pseudomonas aeruginosa*		GIM-1	B1	3a	53
bla$_{GOB}$	*Chryseobacterium meningosepticum*		GOB-1-3, 6-18	B3	3a	11, 54
bla$_{HERA}$	*Escherichia hermannii*		HERA-1-8	A	2a	55
bla$_{IBC}$	(*Enterobacter cloacae, Pseudomonas aeruginosa*)		IBC-1 (GES-7), IBC-2 (GES-8)	A	2be	56
bla$_{IMI}$	*Enterobacter cloacae*		IMI-2, 3 IMI-H, IMI-S	A	2f	57
bla$_{IMP}$	*Serratia marcescens*		IMP-1-16, 18-23, 25, 26, 29, 30, 33-35, 37, 38, 40-42	B1	3a	58
bla$_{IND}$	*Chryseobacterium indologenes*		IND-1, 2, 2a-c, 3-7	B1	3a	11, 59

bla$_{JOHN}$	*Flavobacterium johnsoniae*		JOHN-1	B1	3a	60
bla$_{KHM}$	*Citrobacter freundii*		KHM-1	B1	3a	61
bla$_{KPC}$	*Klebsiella pneumoniae*		KPC-2-13	A	2f	11, 62
bla$_{L}$ (*bla*$_{S}$)	*Stenotrophomonas maltophilia*		L-1, 2	B	3	63
bla$_{LAP}$			LAP-1, 2	A	2b	64
bla$_{LAT}$	*Citrobacter freundii*		LAT-1	C	1	42
bla$_{LEN}$			LEN-1-26	A	2a	65
bla$_{LUT}$			LUT-1	A	2be	66
bla$_{M}$			M-1	A	2a	67
bla$_{MAL}$	*Levinea malonatica (Citrobacter koseri)*		MAL-1, 2			42, 68
bla$_{MIR}$	*Enterobacter cloacae*		MIR-1-6	C	1	63
bla$_{MOR}$	*Morganella morganii*		MOR, MOR-2	C	1	69
bla$_{MOX}$	*Pseudomonas aeruginosa*	MOX-1-like MOX-5-like	MOX-1-4 MOX-5-7	C	1	70
bla$_{MUS}$	*Myroides odoratimimus*		MUS-1	B1	3a	71
bla$_{NDM}$	Possibly *Acinetobacter baumannii*		NDM-1-7	B1	3a	72
bla$_{NMC}$			NMC-A	A	2f	63
bla$_{NPS}$	*Pseudomonas aeruginosa*		NPS-1	A	2c	73
bla$_{OCH}$	*Ochrobactrum anthropi*		OCH-2-8	C	1	74
bla$_{OHIO}$ (*bla*$_{SHV}$)			OHIO-1	A	2b	75
bla$_{OKP}$		OKP-A OKP-B	OKP-A-1-16 OKP-B-1-20	A	2b	65
bla$_{ORN}$	*Raoultella ornitholytica*		ORN-1b	A	2be	76
bla$_{OXA}$		Carbapenemases OXA-23-like OXA-40-like OXA-48-like	 OXA-23 (ARI-1), 27, 49, 73, 133, 146, 165-171, 225, 239 OXA-40, 24-26, 72, 139, 143, 160, 182, 231	 D D D	 2df 2df 2df	77, 78

			OXA-48, 162, 163, 181			
		OXA-51-like	OXA-51, 64-71, 76-80, 82-84, 86-93, 95, 98, 99, 104, 106-113, 115, 128, 130-132, 138, 144, 148-150, 172-180, 194-197, 200-203, 216, 217, 219,	D	2df	
		OXA-58-like	223, 241, 242, 248-250			
		Others	OXA-58, 97, 164			
		Non-carbapenemases	OXA-55, OXA-50b-e, OXA-60a-d			
		OXA-1-like	OXA-1/30, 4, 31, 33, 47, 224	D	2d	
		OXA-2-like	OXA-2, 3, 15, 21, 32, 34, 46, 53, 118, 119, 141, 161, 210	D	2de	
		OXA-10-like	OXA-10, 5, 7, 11, 13, 14, 16, 17, 19, 28, 35, 56, 74, 101, 129, 142, 145, 147,	D	2d, 2de	
		Other	183, 211, 212, 240 OXA-20, 37; OXA-29; OXA-9, OXA-18, 45; OXA-12 (AsbB1), OXA-22: OXA-114a-g, 243; OXA-36, OXA-54, OXA-62, OXA-198; OXA-214, 215; OXA-213, OXA-75, 116, 117, 94; OXA-134a, 235, 236, 237; OXA-96, 228-230; OXA-61, OXA-85			
bla$_{OXY}$	*Klebsiella oxytoca*	OXY-1 OXY-2 OXY-3	1-7 1-13 1	A	2be	79

		OXY-4 OXY-5 OXY-6	1 1, 2 1-4			
bla$_{P}$	*Bacillus licheniformis*		P-1	A	2a	80
bla$_{PER}$	*Pseudomonas aeruginosa*		PER-1-6	A	2be	81
bla$_{PLA}$	*Raoultella planticola*		PLA-2a	A	2be	76
bla$_{PME}$	*Pseudomonas aeruginosa*		PME-1	A	2be	82
bla$_{POM}$	*Pseudomonas otitidis*		POM-1	B3	3a	83
bla$_{PSE}$	*Pseudomonas aeruginosa*		PSE-1 (CARB-2)	A	2c	11, 84
bla$_{RAHN}$	*Rahnella aquatilis*		RAHN-1, 2	A	2be	85
bla$_{ROB}$	*Actinobacillus pleuropneumoniae*		ROB-1	A	2a	86
bla$_{SCO}$	*Escherichia coli*		SCO-1	A	2c	87
bla$_{SED}$	*Citrobacter sedlakii*		SED-1	A	2b	88
bla$_{SFB}$	*Shewanella frigidimarina*		SFB-1	B1	3a	89
bla$_{SFC}$	*Serratia fonticola*		SFC-1	A	2f	90
bla$_{SFH}$	*Serratia fonticola*		SFH-1	B2	3b	11
bla$_{SFO}$	*Serratia fonticola*		SFO-1	A	2be	91
bla$_{SHV}$	*Klebsiella pneumoniae*		SHV-1-3, 5-8, 11, 12, 14, 18, 24-42, 44-51, 53, 55-57, 59-67, 69-72, 77, 84-86, 89, 92-97, 101-105, 108, 109, 112, 121, 122, 128-135	A	2b, 2be, 2br	11
bla$_{SIM}$	*Acinetobacter baumannii*		SIM-1	B1	3a	92
bla$_{SLB}$	*Shewanella livingstonensis*		SLB-1	B1	3a	89
bla$_{SMB}$	*Serratia marcescens*		SMB-1	B3	3a	93
bla$_{SME}$			SME-1-3	A	2f	63
bla$_{SPU}$	*Capnocytophaga sputigena*		SPU-1	A	2be	94

bla$_{SRT}$	*Serratia marcescens*		SRT-1	C	1	95
bla$_{SST}$	*Serratia marcescens*		SST-1	C	1	95
bla$_{TEM}$			TEM-1-13, 15-22, 24-26B, 28-40, 42-44, 47-49, 52-55, 7, 63/64, 67-68, 70-72, 75-91, 93-99, 101, 104-134, 136-139, 141-152, 154-160, 162-164, 166-169, 171, 176-178, 183, 185-195, 197-198, 201	A	2b, 2be, 2br, 2ber,	11
bla$_{TER}$	*Raoultella terrigena*		TER-1	A	2be	96
bla$_{TLA}$	*Escherichia coli*		TLA-1	A	2be	97
bla$_{TMB}$	*Achromobacter xylosoxidans*		TMB-1, TMB-2	B1	3a	98
bla$_{TRU}$	*Aeromonas enteropelogenes*		TRU-1	C	1	99
bla$_{TUS}$	*Myroides odoratus*		TUS-1	B (B1)	3a	71
bla$_{VEB}$			VEB-1-7	A	2be	11
bla$_{VHH}$	*Vibrio harveyi*		VHH-1	A	2a	100
bla$_{VHW}$	*Vibrio harveyi*		VHW-1	A	2a	100
bla$_{VIM}$			VIM-1-20, 22-34	B1	3a	11
bla$_{ZEG}$			ZEG-1	C	1	9
cfiA (ccrA)	*Bacteroides fragilis*		*cfiA*1-16	B1	3a	101
cfxA	*Bacteroides vulgatus*		*cfxA*, *cfxA*3-5	A	2be	102
cphA	*Aeromonas hydrophila*		*cphA*2, *cphA*4-8	B2	3b	103, 104
hugA				A	2be	105
mecA	*Staphylococcus* spp.					106
penA	*Burkholderia* spp.			A		107
Promoter mechanisms						
bla$_{TEM}$ promoter		P3, P4, P5, Pa/b, Pc/d				2
ampC promoter						3

References for Table A1.1.

1. Jacoby G, Bush K. Beta-lactamase classification and amino acid sequences for TEM, SHV and OXA extended-spectrum and inhibitor-resistant enzymes. Available at: http://www.lahey.org/Studies/. Last accessed: 27th December 2012.
2. Lartigue MF, Leflon-Guibout V, Poirel L, Nordmann P, Nicolas-Chanoine MH. Promoters P3, Pa/Pb, P4, and P5 upstream from *bla*(TEM) genes and their relationship to beta-lactam resistance. Antimicrob Agents Chemother. 2002 Dec;46(12):4035-7.
3. Mulvey MR, Bryce E, Boyd DA, Ofner-Agostini M, Land AM, Simor AE, Paton S. Molecular characterization of cefoxitin-resistant *Escherichia coli* from Canadian hospitals. Antimicrob Agents Chemother. 2005 Jan;49(1):358-65.
4. Caroff N, Espaze E, Bérard I, Richet H, Reynaud A. Mutations in the *ampC* promoter of *Escherichia* coli isolates resistant to oxyiminocephalosporins without extended spectrum beta-lactamase production. FEMS Microbiol Lett. 1999 Apr 15;173(2):459-65.
5. Caroff N, Espaze E, Gautreau D, Richet H, Reynaud A. Analysis of the effects of -42 and -32 *ampC* promoter mutations in clinical isolates of *Escherichia coli* hyperproducing *ampC*. J Antimicrob Chemother. 2000 Jun;45(6):783-8.
6. Daraselia N, Dernovoy D, Tian Y, Borodovsky M, Tatusov R, Tatusova T OMICS. Reannotation of *Shewanella oneidensis* genome. 2003 Summer;7(2):171-5.
7. Bauernfeind A, Schneider I, Jungwirth R, Sahly H, Ullmann U. A novel type of AmpC beta-lactamase, ACC-1, produced by a *Klebsiella pneumoniae* strain causing nosocomial pneumonia. Antimicrob Agents Chemother. 1999 Aug;43(8):1924-31.
8. Galán JC, Reig M, Navas A, Baquero F, Blázquez J. ACI-1 from *Acidaminococcus fermentans*: characterization of the first beta-lactamase in Anaerobic cocci. Antimicrob Agents Chemother. 2000 Nov;44(11):3144-9.
9. van Hoek AH, Mevius D, Guerra B, Mullany P, Roberts AP, Aarts HJ. Acquired antibiotic resistance genes: an overview. Front Microbiol. 2011;2:203. doi: 10.3389/fmicb.2011.00203.
10. Bradford PA, Urban C, Mariano N, Projan SJ, Rahal JJ, Bush K. Imipenem resistance in *Klebsiella pneumoniae* is associated with the combination of ACT-1, a plasmid-mediated AmpC beta-lactamase, and the foss of an outer membrane protein. Antimicrob Agents Chemother. 1997 Mar;41(3):563-9.
11. Bush K, Jacoby GA. Updated functional classification of beta-lactamases. Antimicrob Agents Chemother. 2010 Mar;54(3):969-76.
12. Hujer KM, Hamza NS, Hujer AM, Perez F, Helfand MS, Bethel CR, Thomson JM, Anderson VE, Barlow M, Rice LB, Tenover FC, Bonomo RA. Identification of a new allelic variant of the *Acinetobacter baumannii* cephalosporinase, ADC-7 beta-lactamase: defining a unique family of class C enzymes. Antimicrob Agents Chemother. 2005 Jul;49(7):2941-8.
13. Hedges RW, Medeiros AA, Cohenford M, Jacoby GA. Genetic and biochemical properties of AER-1, a novel carbenicillin-hydrolyzing beta-lactamase from *Aeromonas hydrophila*. Antimicrob Agents Chemother. 1985 Apr;27(4):479-84.
14. Sanschagrin F, Bejaoui N, Levesque RC. Structure of CARB-4 and AER-1 carbenicillin-hydrolyzing beta-lactamases. Antimicrob Agents Chemother. 1998 Aug;42(8):1966-72.
15. Poirel L, Laurent F, Naas T, Labia R, Boiron P, Nordmann P. Molecular and biochemical analysis of AST-1, a class A beta-lactamase from *Nocardia asteroides sensu stricto*. Antimicrob Agents Chemother. 2001 Mar;45(3):878-82.
16. Woodford N, Palepou MF, Babini GS, Holmes B, Livermore DM. Carbapenemases of *Chryseobacterium (Flavobacterium) meningosepticum*: distribution of *blaB* and characterization of a novel metallo-beta-lactamase gene, blaB3, in the type strain, NCTC 10016. Antimicrob Agents Chemother. 2000 Jun;44(6):1448-52.
17. Bonnet R, Sampaio JL, Chanal C, Sirot D, De Champs C, Viallard JL, Labia R, Sirot J. A novel class A extended-spectrum beta-lactamase (BES-1) in *Serratia marcescens* isolated in Brazil. Antimicrob Agents Chemother. 2000 Nov;44(11):3061-8.
18. Girlich D, Poirel L, Nordmann P. Novel ambler class A carbapenem-hydrolyzing beta-lactamase from a *Pseudomonas fluorescens* isolate from the Seine River, Paris, France. Antimicrob Agents Chemother. 2010 Jan;54(1):328-32.
19. Payne DJ, Woodford N, Amyes SG. Characterization of the plasmid mediated beta-lactamase BIL-1. J Antimicrob Chemother. 1992 Aug;30(2):119-27.
20. Bootsma HJ, van Dijk H, Verhoef J, Fleer A, Mooi FR. Molecular characterization of the BRO beta-lactamase of *Moraxella (Branhamella) catarrhalis*. Antimicrob Agents Chemother. 1996 Apr;40(4):966-72.
21. Alfredson DA, Korolik V. Isolation and expression of a novel molecular class D beta-lactamase, OXA-61, from *Campylobacter jejuni*. Antimicrob Agents Chemother. 2005 Jun;49(6):2515-8.

22. Potron A, Poirel L, Croizé J, Chanteperdrix V, Nordmann P. Genetic and biochemical characterization of the first extended-spectrum CARB-type beta-lactamase, RTG-4, from *Acinetobacter baumannii*. Antimicrob Agents Chemother. 2009 Jul;53(7):3010-6. doi: 10.1128/AAC.01164-08.
23. Melano R, Petroni A, Garutti A, Saka HA, Mange L, Pasterán F, Rapoport M, Rossi A, Galas M. New carbenicillin-hydrolyzing beta-lactamase (CARB-7) from *Vibrio cholerae* non-O1, non-O139 strains encoded by the VCR region of the *V. cholerae* genome. Antimicrob Agents Chemother. 2002 Jul;46(7):2162-8.
24. Choury D, Szajnert MF, Joly-Guillou ML, Azibi K, Delpech M, Paul G. Nucleotide sequence of the *bla(RTG-2)* (CARB-5) gene and phylogeny of a new group of carbenicillinases. Antimicrob Agents Chemother. 2000 Apr;44(4):1070-4.
25. Philippon AM, Paul GC, Thabaut AP, Jacoby GA. Properties of a novel carbenicillin-hydrolyzing beta-lactamase (CARB-4) specified by an IncP-2 plasmid from *Pseudomonas aeruginosa*. Antimicrob Agents Chemother. 1986 Mar;29(3):519-20.
26. Lachapelle J, Dufresne J, Levesque RC. Characterization of the *blaCARB-3* gene encoding the carbenicillinase-3 beta-lactamase of *Pseudomonas aeruginosa*. Gene. 1991 Jun 15;102(1):7-12.
27. Smith CJ, Bennett TK, Parker AC. Molecular and genetic analysis of the *Bacteroides uniformis* cephalosporinase gene, *cblA*, encoding the species-specific beta-lactamase. Antimicrob Agents Chemother. 1994 Aug;38(8):1711-5.
28. Rogers MB, Parker AC, Smith CJ. Cloning and characterization of the endogenous cephalosporinase gene, *cepA*, from *Bacteroides fragilis* reveals a new subgroup of Ambler class A beta-lactamases. Antimicrob Agents Chemother. 1993 Nov;37(11):2391-400.
29. Nakano R, Okamoto R, Nakano Y, Kaneko K, Okitsu N, Hosaka Y, Inoue M. 2004. CFE-1, a novel plasmid-encoded AmpC β-lactamase with an *ampR* gene originating from *Citrobacter freundii*. Antimicrob. Agents Chemother. 48:1151-1158.
30. Bellais S, Naas T, Nordmann P. Molecular and biochemical characterization of Ambler class A extended-spectrum beta-lactamase CGA-1 from *Chryseobacterium gleum*. Antimicrob Agents Chemother. 2002 Apr;46(4):966-70.
31. Bellais S, Naas T, Nordmann P. Genetic and biochemical characterization of CGB-1, an Ambler class B carbapenem-hydrolyzing beta-lactamase from *Chryseobacterium gleum*. Antimicrob Agents Chemother. 2002 Sep;46(9):2791-6.
32. Petrella S, Renard M, Ziental-Gelus N, Clermont D, Jarlier V, Sougakoff W. Characterization of the chromosomal class A beta-lactamase CKO from *Citrobacter koseri*. FEMS Microbiol Lett. 2006 Jan;254(2):285-92.
33. Matsumoto T, Nagata M, Ishimine N, Kawasaki K, Yamauchi K, Hidaka E, Kasuga E, Horiuchi K, Oana K, Kawakami Y, Honda T. Characterization of CIA-1, an Ambler class A extended-spectrum β-lactamase from *Chryseobacterium indologenes*. Antimicrob Agents Chemother. 2012 Jan;56(1):588-90.
34. Rossolini GM, Franceschini N, Lauretti L, Caravelli B, Riccio ML, Galleni M, Frère JM, Amicosante G. Cloning of a *Chryseobacterium (Flavobacterium) meningosepticum* chromosomal gene (*blaA(CME)*) encoding an extended-spectrum class A beta-lactamase related to the *Bacteroides cephalosporinases* and the VEB-1 and PER beta-lactamases. Antimicrob Agents Chemother. 1999 Sep;43(9):2193-9.
35. Bellais S, Poirel L, Naas T, Girlich D, Nordmann P. Genetic-biochemical analysis and distribution of the Ambler class A beta-lactamase CME-2, responsible for extended-spectrum cephalosporin resistance in *Chryseobacterium (Flavobacterium) meningosepticum*. Antimicrob Agents Chemother. 2000 Jan;44(1):1-9.
36. Obert CA, Goldstone CM, Gordon DM, Riley MA. Novel beta-lactamase isolated from wild Australian enteric isolates. Unpublished – direct submission to NCBI nucleotide database (GenBank AY265892.1)
37. Yu WL, Lee MF, Chuang YC. A novel plasmid-borne *ampC* gene (*blaCMH-1*) in an *Enterobacter cloacae* isolate from southern Taiwan. Unpublished – direct submission to NCBI nucleotide database (GenBank JQ673557.1)
38. Guillon H, Eb F, Mammeri H. Characterization of CSP-1, a novel extended-spectrum beta-lactamase produced by a clinical isolate of *Capnocytophaga sputigena*. Antimicrob Agents Chemother. 2010 May;54(5):2231-4
39. Bonnet R. Growing group of extended-spectrum beta-lactamases: the CTX-M enzymes. Antimicrob Agents Chemother. 2004 Jan;48(1):1-14.

40. Morin AS, Poirel L, Mory F, Labia R, Nordmann P. Biochemical-genetic analysis and distribution of DES-1, an Ambler class A extended-spectrum beta-lactamase from *Desulfovibrio desulfuricans*. Antimicrob Agents Chemother. 2002 Oct;46(10):3215-22.
41. Barnaud G, Arlet G, Verdet C, Gaillot O, Lagrange PH, Philippon A. *Salmonella enteritidis*: AmpC plasmid-mediated inducible beta-lactamase (DHA-1) with an *ampR* gene from *Morganella morganii*. Antimicrob Agents Chemother. 1998 Sep;42(9):2352-8.
42. Jacoby GA. AmpC beta-lactamases. Clin Microbiol Rev. 2009 Jan;22(1):161-82.
43. Poirel L, Rodríguez-Martínez JM, Al Naiemi N, Debets-Ossenkopp YJ, Nordmann P. Characterization of DIM-1, an integron-encoded metallo-beta-lactamase from a *Pseudomonas stutzeri* clinical isolate in the Netherlands. Antimicrob Agents Chemother. 2010 Jun;54(6):2420-4.
44. Bellais S, Girlich D, Karim A, Nordmann P. EBR-1, a novel Ambler subclass B1 beta-lactamase from *Empedobacter brevis*. Antimicrob Agents Chemother. 2002 Oct;46(10):3223-7.
45. Vimont S, Poirel L, Naas T, Nordmann P. Identification of a chromosome-borne expanded-spectrum class a beta-lactamase from *Erwinia persicina*. Antimicrob Agents Chemother. 2002 N6v;46(11):3401-5.
46. Laurent F, Poirel L, Naas T, Chaibi EB, Labia R, Boiron P, Nordmann P. Biochemical-genetic analysis and distribution of FAR-1, a class A beta-lactamase from *Nocardia farcinica*. Antimicrob Agents Chemother. 1999 Jul;43(7):1644-50.
47. Pollini S, Maradei S, Pecile P, Olivo G, Luzzaro F, Docquier JD, Rossolini GM. FIM-1, a New Acquired Metallo-β-Lactamase from a *Pseudomonas aeruginosa* Clinical Isolate from Italy. Antimicrob Agents Chemother. 2013 Jan;57(1):410-6.
48. Péduzzi J, Farzaneh S, Reynaud A, Barthélémy M, Labia R. Characterization and amino acid sequence analysis of a new oxyimino cephalosporin-hydrolyzing class A beta-lactamase from *Serratia fonticola* CUV. Biochim Biophys Acta. 1997 Aug 15;1341(1):58-70.
49. Naas T, Poirel L, Nordmann P. Minor extended-spectrum beta-lactamases. Clin Microbiol Infect. 2008 Jan;14 Suppl 1:42-52.
50. Gonzalez Leiza M, Perez-Diaz JC, Ayala J, Casellas JM, Martinez-Beltran J, Bush K, Baquero F. Gene sequence and biochemical characterization of FOX-1 from *Klebsiella pneumoniae*, a new AmpC-type plasmid-mediated beta-lactamase with two molecular variants. Antimicrob Agents Chemother. 1994 Sep;38(9):2150-7.
51. Voha C, Docquier JD, Rossolini GM, Fosse T. Genetic and biochemical characterization of FUS-1 (OXA-85), a narrow-spectrum class D beta-lactamase from *Fusobacterium nucleatum* subsp *polymorphum*.. Antimicrob Agents Chemother. 2006 Aug;50(8):2673-9.
52. Poirel L, Le Thomas I, Naas T, Karim A, Nordmann P. Biochemical sequence analyses of GES-1, a novel class A extended-spectrum beta-lactamase, and the class 1 integron In52 from *Klebsiella pneumoniae*. Antimicrob Agents Chemother. 2000 Mar;44(3):622-32.
53. Castanheira M, Toleman MA, Jones RN, Schmidt FJ, Walsh TR. Molecular characterization of a beta-lactamase gene, blaGIM-1, encoding a new subclass of metallo-beta-lactamase. Antimicrob Agents Chemother. 2004 Dec;48(12):4654-61.
54. Bellais S, Aubert D, Naas T, Nordmann P. Molecular and biochemical heterogeneity of class B carbapenem-hydrolyzing beta-lactamases *in Chryseobacterium meningosepticum*. Antimicrob Agents Chemother. 2000 Jul;44(7):1878-86.
55. Beauchef-Havard A, Arlet G, Gautier V, Labia R, Grimont P, Philippon A. Molecular and biochemical characterization of a novel class A beta-lactamase (HER-1) from *Escherichia hermannii*. Antimicrob Agents Chemother. 2003 Aug;47(8):2669-73.
56. Vourli S, Giakkoupi P, Miriagou V, Tzelepi E, Vatopoulos AC, Tzouvelekis LS. Novel GES/IBC extended-spectrum beta-lactamase variants with carbapenemase activity in clinical enterobacteria. FEMS Microbiol Lett. 2004 May 15;234(2):209-13.
57. Rasmussen BA, Bush K, Keeney D, Yang Y, Hare R, O'Gara C, Medeiros AA. Characterization of IMI-1 beta-lactamase, a class A carbapenem-hydrolyzing enzyme from *Enterobacter cloacae*. Antimicrob Agents Chemother. 1996 Sep;40(9):2080-6.
58. Osano E, Arakawa Y, Wacharotayankun R, Ohta M, Horii T, Ito H, Yoshimura F, Kato N. Molecular characterization of an enterobacterial metallo beta-lactamase found in a clinical isolate of *Serratia marcescens* that shows imipenem resistance. Antimicrob Agents Chemother. 1994 Jan;38(1):71-8.
59. Bellais S, Poirel L, Leotard S, Naas T, Nordmann P. Genetic diversity of carbapenem-hydrolyzing metallo-beta-lactamases from *Chryseobacterium (Flavobacterium) indologenes*. Antimicrob Agents Chemother. 2000 Nov;44(11):3028-34.

60. Naas T, Bellais S, Nordmann P. Molecular and biochemical characterization of a carbapenem-hydrolysing beta-lactamase from *Flavobacterium johnsoniae*. J Antimicrob Chemother. 2003 Feb;51(2):267-73.
61. Sekiguchi J, Morita K, Kitao T, Watanabe N, Okazaki M, Miyoshi-Akiyama T, Kanamori M, Kirikae T. KHM-1, a novel plasmid-mediated metallo-beta-lactamase from a *Citrobacter freundii* clinical isolate. Antimicrob Agents Chemother. 2008 Nov;52(11):4194-7.
62. Yigit H, Queenan AM, Anderson GJ, Domenech-Sanchez A, Biddle JW, Steward CD, Alberti S, Bush K, Tenover FC. Novel carbapenem-hydrolyzing beta-lactamase, KPC-1, from a carbapenem-resistant strain of *Klebsiella pneumoniae*.
63. Bush K, Jacoby GA, Medeiros AA. A functional classification scheme for beta-lactamases and its correlation with molecular structure. Antimicrob Agents Chemother. 1995 Jun;39(6):1211-33.
64. Poirel L, Cattoir V, Soares A, Soussy CJ, Nordmann P. Novel Ambler class A beta-lactamase LAP-1 and its association with the plasmid-mediated quinolone resistance determinant QnrS1. Antimicrob Agents Chemother. 2007 Feb;51(2):631-7.
65. Haeggman S, Löfdahl S, Paauw A, Verhoef J, Brisse S. Diversity and evolution of the class A chromosomal beta-lactamase gene in *Klebsiella pneumoniae*. Antimicrob Agents Chemother. 2004 Jul;48(7):2400-8.
66. Doublet B, Robin F, Casin I, Fabre L, Le Fleche A, Bonnet R, Weill FX. Molecular and biochemical characterization of the natural chromosome-encoded class A beta-lactamase from *Pseudomonas luteola*. Antimicrob Agents Chemother. 2010 Jan;54(1):45-51.
67. Uyaguari MI Jr, Fichot EB, Scott GI, Norman RS. Characterization and quantitation of a novel beta-lactamase gene within a wastewater treatment facility and surrounding coastal ecosystem. Unpublished – direct submission to NCBI nucleotide database (GenBank HQ605913.1)
68. Jacoby GA. Beta-lactamase nomenclature. Antimicrob Agents Chemother. 2006 Apr;50(4):1123-9.
69. Barnaud G, Arlet G, Danglot C, Philippon A. Cloning and sequencing of the gene encoding the AmpC beta-lactamase of *Morganella morganii*. FEMS Microbiol Lett. 1997 Mar 1;148(1):15-20.
70. Horii T, Arakawa Y, Ohta M, Ichiyama S, Wacharotayankun R, Kato N. Plasmid-mediated AmpC-type beta-lactamase isolated from *Klebsiella pneumoniae* confers resistance to broad-spectrum beta-lactams, including moxalactam. Antimicrob Agents Chemother. 1993 May;37(5):984-90.
71. Mammeri H, Bellais S, Nordmann P. Chromosome-encoded beta-lactamases TUS-1 and MUS-1 from *Myroides odoratus* and *Myroides odoratimimus* (formerly *Flavobacterium odoratum*), new members of the lineage of molecular subclass B1 metalloenzymes. Antimicrob Agents Chemother. 2002 Nov;46(11):3561-7.
72. Karthikeyan K, Thirunarayan MA, Krishnan P. Coexistence of blaOXA-23 with blaNDM-1 and armA in clinical isolates of *Acinetobacter baumannii* from India. J Antimicrob Chemother. 2010 Oct;65(10):2253-4.
73. Livermore DM, Jones CS. Characterization of NPS-1, a novel plasmid-mediated beta-lactamase, from two *Pseudomonas aeruginosa* isolates. Antimicrob Agents Chemother. 1986 Jan;29(1):99-103
74. Nadjar D, Labia R, Cerceau C, Bizet C, Philippon A, Arlet G. Molecular characterization of chromosomal class C beta-lactamase and its regulatory gene in *Ochrobactrum anthropi*. Antimicrob Agents Chemother. 2001 Aug;45(8):2324-30.
75. Shlaes DM, Currie-McCumber C, Hull A, Behlau I, Kron M. OHIO-1 beta-lactamase is part of the SHV-1 family. Antimicrob Agents Chemother. 1990 Aug;34(8):1570-6.
76. Walckenaer E, Poirel L, Leflon-Guibout V, Nordmann P, Nicolas-Chanoine MH. Genetic and biochemical characterization of the chromosomal class A beta-lactamases of *Raoultella* (formerly *Klebsiella*) *planticola* and *Raoultella ornithinolytica*. Antimicrob Agents Chemother. 2004Jan;48(1):305-12.
77. Walther-Rasmussen J, Høiby N. OXA-type carbapenemases. J Antimicrob Chemother. 2006 Mar;57(3):373-83.
78. Naas T, Nordmann P. OXA-type beta-lactamases. Curr Pharm Des. 1999 Nov;5(11):865-79.
79. Fournier B, Roy PH, Lagrange PH, Philippon A. Chromosomal beta-lactamase genes of *Klebsiella oxytoca* are divided into two main groups, blaOXY-1 and blaOXY-2. Antimicrob Agents Chemother. 1996 Feb;40(2):454-9.
80. Neugebauer K, Sprengel R, Schaller H. Penicillinase from *Bacillus licheniformis*: nucleotide sequence of the gene and implications for the biosynthesis of a secretory protein in a Gram-positive bacterium. Nucleic Acids Res. 1981 Jun 11;9(11):2577-88.
81. Nordmann P, Naas T. Sequence analysis of PER-1 extended-spectrum beta-lactamase from *Pseudomonas aeruginosa* and comparison with class A beta-lactamases. Antimicrob Agents Chemother. 1994 Jan;38(1):104-14.

82. Tian GB, Adams-Haduch JM, Bogdanovich T, Wang HN, Doi Y. PME-1, an extended-spectrum β-lactamase identified in *Pseudomonas aeruginosa*. Antimicrob Agents Chemother. 2011 Jun;55(6):2710-3.
83. Thaller MC, Borgianni L, Di Lallo G, Chong Y, Lee K, Dajcs J, Stroman D, Rossolini GM.. Metallo-beta-lactamase production by *Pseudomonas otitidis*: a species-related trait. Antimicrob Agents Chemother. 2011 Jan;55(1):118-23.
84. Huovinen P, Jacoby GA. Sequence of the PSE-1 beta-lactamase gene. Antimicrob Agents Chemother. 1991 Nov;35(11):2428-30.
85. Ruimy R, Meziane-Cherif D, Momcilovic S, Arlet G, Andremont A, Courvalin P. RAHN-2, a chromosomal extended-spectrum class A beta-lactamase from *Rahnella aquatilis*. J Antimicrob Chemother. 2010 Aug;65(8):1619-23.
86. Juteau JM, Levesque RC. Sequence analysis and evolutionary perspectives of ROB-1 beta-lactamase. Antimicrob Agents Chemother. 1990 Jul;34(7):1354-9.
87. Papagiannitsis CC, Loli A, Tzouvelekis LS, Tzelepi E, Arlet G, Miriagou V. SCO-1, a novel plasmid-mediated class A beta-lactamase with carbenicillinase characteristics from *Escherichia coli*. Antimicrob Agents Chemother. 2007 Jun;51(6):2185-8.
88. Petrella S, Clermont D, Casin I, Jarlier V, Sougakoff W. Novel class A beta-lactamase Sed-1 from *Citrobacter sedlakii*: genetic diversity of beta-lactamases within the *Citrobacter* genus. Antimicrob Agents Chemother. 2001 Aug;45(8):2287-98.
89. Poirel L, Héritier C, Nordmann P. Genetic and biochemical characterization of the chromosome-encoded class B beta-lactamases from *Shewanella livingstonensis* (SLB-1) and *Shewanella frigidimarina* (SFB-1). J Antimicrob Chemother. 2005 May;55(5):680-5. Epub 2005 Mar 16.
90. Henriques I, Moura A, Alves A, Saavedra MJ, Correia A. Molecular characterization of a carbapenem-hydrolyzing class A beta-lactamase, SFC-1, from *Serratia fonticola* UTAD54. Antimicrob Agents Chemother. 2004 Jun;48(6):2321-4.
91. Matsumoto Y, Inoue M.Characterization of SFO-1, a plasmid-mediated inducible class A beta-lactamase from *Enterobacter cloacae*. Antimicrob Agents Chemother. 1999 Feb;43(2):307-13.
92. Lee K, Yum JH, Yong D, Lee HM, Kim HD, Docquier JD, Rossolini GM, Chong Y. Novel acquired metallo-beta-lactamase gene, bla(SIM-1), in a class 1 integron from *Acinetobacter baumannii* clinical isolates from Korea. Antimicrob Agents Chemother. 2005 Nov;49(11):4485-91.
93. Wachino J, Yoshida H, Yamane K, Suzuki S, Matsui M, Yamagishi T, Tsutsui A, Konda T, Shibayama K, Arakawa Y. SMB-1, a novel subclass B3 metallo-beta-lactamase, associated with ISCR1 and a class 1 integron, from a carbapenem-resistant *Serratia marcescens* clinical isolate. Antimicrob Agents Chemother. 2011 Nov;55(11):5143-9.
94. Ehrmann E, Handal T, Giraud-Morin C, Bonnaure-Mallet M, Fosse T. Prevalence of beta-lactamase genes in *Capnocytophaga* spp., and characterization of Spu-1, a new ESBL in *Capnocytophaga sputigena*. Unpublished – direct submission to NCBI nucleotide database (GenBank Q919044.1)
95. Matsumura N, Minami S, Mitsuhashi S. Sequences of homologous beta-lactamases from clinical isolates of *Serratia marcescens* with different substrate specificities. Antimicrob Agents Chemother. 1998 Jan;42(1):176-9.
96. Walckenaer E, Delmas J, Leflon-Guibout V, Bonnet R, Nicolas-Chanoine MH. Genetic, Biochemical Characterization and Mutagenesis of the Chromosomal Class A Beta-Lactamase of *Raoultella* (formerly *Klebsiella*) *terrigena*. Unpublished – direct submission to NCBI nucleotide database (GenBankFJ263091.1)
97. Silva J, Aguilar C, Ayala G, Estrada MA, Garza-Ramos U, Lara-Lemus R, Ledezma L . TLA-1: a new plasmid-mediated extended-spectrum beta-lactamase from *Escherichia coli*. Antimicrob Agents Chemother. 2000 Apr;44(4):997-1003.
98. El Salabi A, Borra PS, Toleman MA, Samuelsen Ø, Walsh TR. Genetic and biochemical characterization of a novel metallo-β-lactamase, TMB-1, from an *Achromobacter xylosoxidans* strain isolated in Tripoli, Libya. Antimicrob Agents Chemother. 2012 May;56(5):2241-5.
99. De Luca F, Giraud-Morin C, Rossolini GM, Docquier JD, Fosse T. Genetic and biochemical characterization of TRU-1, the endogenous class C beta-lactamase from *Aeromonas enteropelogenes*. Antimicrob Agents Chemother. 2010 Apr;54(4):1547-54.
100. Teo JW, Suwanto A, Poh CL. Novel beta-lactamase genes from two environmental isolates of *Vibrio harveyi*. Antimicrob Agents Chemother. 2000 May;44(5):1309-14.
101. Thompson JS, Malamy MH. Sequencing the gene for an imipenem-cefoxitin-hydrolyzing enzyme (CfiA) from *Bacteroides fragilis* TAL2480 reveals strong similarity between CfiA and *Bacillus cereus* beta-lactamase II. J Bacteriol. 1990 May;172(5):2584-93.

102. Parker AC, Smith CJ. Genetic and biochemical analysis of a novel Ambler class A beta-lactamase responsible for cefoxitin resistance in *Bacteroides* species. Antimicrob Agents Chemother. 1993 May;37(5):1028-36.
103. Massidda O, Rossolini GM, Satta G. The *Aeromonas hydrophila cphA* gene: molecular heterogeneity among class B metallo-beta-lactamases. J Bacteriol. 1991 Aug;173(15):4611-7.
104. Queenan AM, Bush K. Carbapenemases: the versatile beta-lactamases. Clin Microbiol Rev. 2007 Jul;20(3):440-58.
105. Liassine N, Madec S, Ninet B, Metral C, Fouchereau-Peron M, Labia R, Auckenthaler R. Postneurosurgical meningitis due to *Proteus penneri* with selection of a ceftriaxone-resistant isolate: analysis of chromosomal class A beta-lactamase HugA and its LysR-type regulatory protein HugR. Antimicrob Agents Chemother. 2002 Jan;46(1):216-9.
106. Ubukata K, Nonoguchi R, Matsuhashi M, Konno M. Expression and inducibility in *Staphylococcus aureus* of the mecA gene, which encodes a methicillin-resistant *S. aureus*-specific penicillin-binding protein. J Bacteriol. 1989 May;171(5):2882-5.
107. Rholl DA, Papp-Wallace KM, Tomaras AP, Vasil ML, Bonomo RA, Schweizer HP. Molecular Investigations of PenA-mediated β-lactam Resistance in *Burkholderia pseudomallei*. Front Microbiol. 2011;2:139.

Table A1.2. Fluoroquinolone resistance mechanisms included in the study resistance locus database.

Enzyme family	Major sub-groups	Variants	Major phenotypic resistance	Reference
norA			Low-level hydrophilic fluoroquinolone resistance	1
qnr	A	1-8	Low-level fluoroquinolone resistance	2, 3
	B	1-15, 17-25, 27-38, 40-53, 56-59, 62	Low-level fluoroquinolone resistance	2, 3
	C		Low-level fluoroquinolone resistance	2, 3
	D		Low-level fluoroquinolone resistance	2, 3
	S	1-6	Low-level fluoroquinolone resistance	2, 3
qepA		*qepA, qepA*2	fluoroquinolone resistance	2, 3
oqxAB			fluoroquinolone resistance	2, 3
gyrA	Quinolone-resistance determining region: amino acids 67-106		Cumulative increase in resistance to fluoroquinolones through stepwise mutations	2, 4, 5
gyrB	Quinolone-resistance determining region: amino acids 426-464		Cumulative increase in resistance to fluoroquinolones through stepwise mutations	2, 4
parC	Quinolone-resistance determining region: amino acids 47-133		Cumulative increase in resistance to fluoroquinolones through stepwise mutations	2, 4
parE	Quinolone-resistance determining region: amino acids 420-458		Cumulative increase in resistance to fluoroquinolones through stepwise mutations	2, 4

References for Table A1.2.

1. Yoshida H, Bogaki M, Nakamura S, Ubukata K, Konno M. Nucleotide sequence and characterization of the *Staphylococcus aureus norA* gene, which confers resistance to quinolones. J Bacteriol. 1990 Dec;172(12):6942-9.
2. Jacoby GA. Mechanisms of resistance to quinolones. Clin Infect Dis. 2005 Jul 15;41 Suppl 2:S120-6.
3. Ruiz J, Pons MJ, Gomes C. Transferable mechanisms of quinolone resistance. Int J Antimicrob Agents. 2012 Sep;40(3):196-203.
4. Hopkins KL, Davies RH, Threlfall EJ. Mechanisms of quinolone resistance in *Escherichia coli* and *Salmonella*: recent developments. Int J Antimicrob Agents. 2005 May;25(5):358-73.
5. Le TM, Baker S, Le TP, Le TP, Cao TT, Tran TT, Nguyen VM, Campbell JI, Lam MY, Nguyen TH, Nguyen VV, Farrar J, Schultsz C. High prevalence of plasmid-mediated quinolone resistance determinants in commensal members of the Enterobacteriaceae in Ho Chi Minh City, Vietnam. J Med Microbiol. 2009 Dec;58(Pt 12):1585-92.

Table A1.3. Aminoglycoside resistance mechanisms included in the study resistance locus database.

Enzyme family	Major sub-groups	Variants (alternative nomenclature)	Major phenotypic resistance relevant to human clinical medicine	Reference
Aminoglycoside N-acetyltransferases	AAC(3)-I	a (*aacC1*)-e	Gentamicin	1-3
	AAC(3)-II	a (*aacC3*, *aacC5*, *aacC2*, aac(3)-Va) -e	Gentamicin, tobramycin	2, 3
	AAC(3)-III	a (*aacC3*)-c (*ant(2'')-Ib*)	Gentamicin, tobramycin	3, 4
	AAC(3)-IV	a	Gentamicin, tobramycin	1, 3, 4
	AAC(3)-VI	a	Gentamicin	3
	AAC(3)-VII	a (*aacC7*)	Gentamicin	4
	AAC(3)-VIII	a (*aacC8*)	Neomycin	5
	AAC(3)-IX	a (*aacC9*)	Neomycin	5
	AAC(3)-X	a	Amikacin	2
	AAC(2')-I	a-e	Gentamicin, tobramycin	2, 4
	AAC(6')-I	A (*aacA1*)-ai, numerous Ib (*aacA4*) sub-variants including: 3, 4, 7, 8, 9, 11, 29a, 29b, 30, 31, 32, 33, 43, Suzhou, Hangzhou	Amikacin, tobramycin, sometimes gentamicin (e.g. AAC(6')-Ib_{11})	2, 4, 6
	AAC(6')-Ib-cr		Amikacin, tobramycin, low-level fluoroquinolone resistance	2, 6
	AAC(6')-II	a-c	Gentamicin, tobramycin	2, 4
	Fusion variants	aac(6')-aph(2''), ant(3'')-Ii-aac(6')-IId, aac(6')-30/aac(6')-Ib', aac(3)-Ib/aac(6')-Ib'', aacA-aphD	Variable, including amikacin , gentamicin, tobramycin, streptomycin	2
Aminoglycoside O-nucleotidyltransferases	ANT(2'')-I	a (*aadB*)	Gentamicin, tobramycin	2
	ANT(3'')-I	a (*aadA, aadA1, aad(3'')(9)*); *aadA* 2-17, 1a, 1b, 2b, 21, 23-	Streptomycin, (spectinomycin)	2

		25		
	ANT(4)-I	a C (*ant(4')-Ia*, *aadD2*, *aadD*, *ant(4',4'')-I*)	Amikacin, tobramycin	2
	ANT(4)-II	a, b	Amikacin, tobramycin	2
	ANT(6)-I	a (*ant6, aadE*), b, also *aadK*, *aad(6)* variants	Streptomycin	2
	ANT(9)-I	a (*aad(9), spc*), b (also called *aad(9), spc*)	(Spectinomycin)	2
	Fusion variants	aadA6/aadA10, ant(3'')-Ih-aac(6')-Iid	Variable	2
Aminoglycoside O-phosphotransferases	APH(2'')-I	e	Gentamicin	2
	APH(2'')-II	a	Gentamicin	2
	APH(2'')-III	a	Gentamicin	2
	APH(2'')-IV	a	Gentamicin	2
	APH(3')-I	a (*aphA-1*), b (*aphA-like*), c (*apha1-IAB, apha7*)	Neomycin	2, 3
	APH(3')-II	IIa (*aphA-2*), IIb, IIc	Neomycin	3
	APH(3')-III	IIIa	Neomycin	3
	APH(3')-IV	Iva (*aphA4*)	Neomycin	3
	APH(3')-V	a (*aphA-5a*), b (*aphA-5b)*, c (*aphA-5c)*	Neomycin	3
	APH(3')-VI	a (*aphA6*), b	Neomycin	3
	APH(3')-VII	a (*aphA7*)	Neomycin	3
	APH(3'')-I	a (*aphE, aphD2*), b (*strA*, *orfH*), c	Streptomycin	2, 3
	APH(4)-I	a (*hph)*, b (*hyg)*	(Hygromcyin)	2, 3
	APH(6)-I	a (*aphD*/*strA*), b (*sph*), c (*str*), d (*strB*, *orfI*)	Streptomycin	2, 3
	APH(7'')-I	a	(Hygromycin)	2
	APH(9)-I	a, b	(Spectinomycin)	2
rRNA methylases	*apmA*		All aminoglycosides	7
	npmA		All aminoglycosides	7

	armA		All aminoglycosides	7
	rmt	A, B, C, D, D2	All aminoglycosides	7

References for Table A1.3.

1. Card R, Zhang J, Das P, Cook C, Woodford N, Anjum MF. Evaluation of an expanded microarray for detecting antibiotic resistance genes in a broad range of gram-negative bacterial pathogens. Antimicrob Agents Chemother. 2013 Jan;57(1):458-65.
2. Ramirez MS, Tolmasky ME. Aminoglycoside modifying enzymes. Drug Resist Updat. 2010 Dec;13(6):151-71.
3. Shaw KJ, Rather PN, Hare RS, Miller GH. Molecular genetics of aminoglycoside resistance genes and familial relationships of the aminoglycoside-modifying enzymes. Microbiol Rev. 1993 Mar;57(1):138-63.
4. Vakulenko SB, Mobashery S. Versatility of aminoglycosides and prospects for their future. Clin Microbiol Rev. 2003 Jul;16(3):430-50.
5. Salauze D, Perez-Gonzalez JA, Piepersberg W, Davies J. Characterisation of aminoglycoside acetyltransferase-encoding genes of neomycin-producing *Micromonospora chalcea* and *Streptomyces fradiae*. Gene. 1991 May 15;101(1):143-8.
6. Robicsek A, Jacoby GA, Hooper DC. The worldwide emergence of plasmid-mediated quinolone resistance. Lancet Infect Dis. 2006 Oct;6(10):629-40.
7. Doi Y, Arakawa Y. 16S ribosomal RNA methylation: emerging resistance mechanism against aminoglycosides. Clin Infect Dis. 2007 Jul 1;45(1):88-94.

Appendix 2.

Table A2.1. List of sizes, sequencing and assembly details for plasmid transformants in Chapter 8.

Transformant	Sequencing platform	Assembler	CTX-M variant	ST131 clade	Contigs in assembly	Assembled genome size (bp)
uk_18B21F_T	HiSeq 2x150PE	standard A5	15	C	22	92421
uk_18C22I_T	HiSeq 2x150PE	standard A5	15	C	21	92448
uk_18A33A_T	HiSeq 2x150PE	standard A5	15	C	24	93292
uk_18B18D_T	HiSeq 2x150PE	standard A5	15	C	25	95507
uk_18B11D_T	HiSeq 2x150PE	standard A5	15	C	25	95819
uk_17B26A_T	HiSeq 2x150PE	standard A5	15	C	25	96420
uk_18C4F_T	HiSeq 2x150PE	standard A5	15	C	25	130907
uk_17C26C_T	HiSeq 2x150PE	standard A5	15	C	27	131222
uk_18B30B_T	HiSeq 2x150PE	standard A5	15	C	30	131474
uk_19B19L_T	HiSeq 2x150PE	standard A5	15	C	33	148844
la_5220-3_T	MiSeq 2x300PE	A5-MiSeq	24	B	2	72449
la_12107_3_T	MiSeq 2x300PE	A5-MiSeq	55	B	6	80841
P26250_T	MiSeq 2x300PE	A5-MiSeq	15	C	9	83980
uk_17A7A_T	MiSeq 2x300PE	A5-MiSeq	15	C	7	91988
11B00320_T	MiSeq 2x300PE	A5-MiSeq	15	A	4	92477
la_7619_T	MiSeq 2x300PE	A5-MiSeq	14	B	1	93517
la_5108_T	MiSeq 2x300PE	A5-MiSeq	14	B	18	100260
uk_8A16G_T	MiSeq 2x300PE	A5-MiSeq	15	C	32	119837
11B01979_T	MiSeq 2x300PE	A5-MiSeq	15	C	22	125735
HFMK272NLF_T	MiSeq 2x300PE	A5-MiSeq	27	A	15	138934
cam_1071_T	MiSeq 2x300PE	A5-MiSeq	14	A	27	139269
P46212_T	MiSeq 2x300PE	A5-MiSeq	15	C	14	140660

uk_8A9B_T	MiSeq 2x300PE	A5-MiSeq	14	A	20	145408
P16456_T	MiSeq 2x300PE	A5-MiSeq	15	C	21	147554
uk_1816C_T	MiSeq 2x300PE	A5-MiSeq	15	C	16	149236
uk_18C29E_T	MiSeq 2x300PE	A5-MiSeq	15	C	27	161487
10B05087_T	MiSeq 2x300PE	A5-MiSeq	15	C	19	166080
09B06576_T	MiSeq 2x300PE	A5-MiSeq	15	C	23	171919