Investigating the increasing incidence of Gram-negative bloodstream infections in England using linked epidemiology, genomic and clinical datasets

Samuel I. W. Lipworth

Lady Margaret Hall, University of Oxford

Nuffield Department of Medicine

A thesis submitted for the degree of Doctor of Philosophy

Trinity Term 2022

Word Count 37,817



Table of Contents

Abstract 8

0 Preface 9

0.1 Acknowledgments 9

0.2 Funding 10

0.3 Ethics Statements 10

0.4 Index of Figures and Tables 11

0.5 Papers arising directly from this work 14

Peer reviewed 14

Preprints currently under review 14

0.6 Other papers arising in the course of this work 15

Peer reviewed 15

Preprints currently under review 17

0.7 Declaration of contribution to thesis 19

0.7.1 Chapters 2, 3 and 4 19

0.7.2 Chapter 5 20

0.8 Abbreviations 20

1. Introduction 23

1.2 Direction of this thesis 24

1.3 A brief history of epidemiology 25

1.4 A brief history of whole genome sequencing 26

1.5 Molecular epidemiology 30

1.6 Identification of bacteria in the routine microbiology laboratory 32

1.7 Sequence based taxonomy/relatedness 32

1.8 Taxonomy/History of *E. coli* and *Klebsiella* spp. 33

1.8.1 *E. coli* 34

1.8.2 *Klebsiella* spp. 35

1.9 Epidemiology of *E. coli*/*Klebsiella* spp. BSIs 36

1.10 Thesis outline 37

2. Molecular epidemiology of *E. coli* and *K. pneumoniae* bloodstream infections in Oxfordshire 2008-2018 39

2.1 Introduction 39

2.2 Methods 41

2.2.1 Study setting, laboratory procedures and DNA extraction 41

2.2.2 Epidemiological linkage 42

2.2.3 Genomic analysis 43

2.2.4 Statistical analysis 45

2.3 Results 46

2.3.1 *Escherichia coli* bloodstream infections are stably dominated by major lineages across community and healthcare-associated settings, but with diverse sporadic lineages accounting for almost half of cases, and evidence of sub-lineage replacement in major STs 46

2.3.2 Klebsiella spp. bloodstream infections are caused by a diverse representation of sub-species with significant intra-species diversity, with the exception of *K. pneumoniae* ST490, causing a transient, clonal, local outbreak 56

2.3.4 No evidence of intra- or inter-regional clustering of major *Escherichia coli* lineages causing BSI in either community- or healthcare-associated settings 60

2.3.5 Urinary and hepatobiliary sources of *E. coli* and *Klebsiella* spp. BSI predominate, with major *E. coli* STs over-represented in urinary-associated *E. coli* BSIs 65

2.3.6 Multi-drug resistant *E. coli* BSIs are increasing, likely driven by the association of AMR genes within lineages, and ceftriaxone resistance is driven by the expansion of resistant ST131 and ST73 sub-lineages 68

2.3.7 Drug-susceptible *Klebsiella* spp. BSIs are increasing faster than drug-resistant strains, with ceftriaxone-resistance largely healthcare-associated 69

2.3.8 Virulence factors are strongly structured by ST amongst *E. coli* but not *Klebisella* spp. BSI isolates, with no evidence in *Klebsiella* spp. of a difference in virulence gene carriage between community vs hospital acquired BSIs 72

2.3.9 Plasmid replicon and plasmidome analyses reflect diversity consistent with a highly mobile accessory genome within species 76

2.4 Discussion 82

2.5 Conclusions 86

3 Seroepidemiological study of *E. coli*/*Klebsiella* spp. BSIs in Oxfordshire UK 87

3.1 Introduction 87

3.2 Methods 89

3.2.1 Study description 89

3.2.2 Statistics 90

3.3.3 Bioinformatics 91

3.3.4 Antibiotic resistance 91

3.4 Results 91

3.4 Discussion 101

3.5 Conclusions 103

4 Bacterial genomic features associated with patient outcome in *E. coli* bloodstream infections 105

4.1 Introduction 105

4.2 Methods 106

4.2.1 Study description 106

4.2.2 Linkage to clinical data 107

4.2.2 Sequencing and bioinformatics 107

4.3.3 Genome wide association study 108

4.2.4 Patient factors model 109

4.2.5 Software used for visualisation 109

4.3 Results 109

4.3.1 Characteristics of 3394 patients with *E. coli* BSI 109

4.3.2 Core patient factors model 113

4.3.3 Association of population structure with patient outcome 116

4.3.4 Curated association study 120

4.3.5 Agnostic genome wide association study 120

4.3.6 Combined host/pathogen model 122

4.4 Discussion 127

5. The Mobilome associated with Gram-negative bloodstream infections: A large-scale observational hybrid sequencing based study 129

5.1 Introduction 129

5.2 Methods 131

5.2.1 Methods development 131
5.2.2 Isolate selection 132
5.2.3 Sequencing 133
5.2.4 Bioinformatics 133
5.2.5 Plasmid clustering 135
5.2.6 Comparison with existing plasmid sequencing data 138
5.2.7 Statistical analysis 138
5.2.8 Data visualisation 140
5.3 Results 140
5.3.1 Most BSI isolates carry a large (>100,000bp) plasmid from a small number of common plasmid groups 142
5.3.2 Plasmid groups are structured by host phylogeny but there is evidence of intra- and inter-species transfer events 142
5.3.3 Common plasmid groups share genes with each other; gene sharing with chromosomes is also frequent 146
5.3.4 Plasmids associated with ARG carriage are often highly similar to those with no such genes 147
5.3.5 Hybrid assembly reveals complex nested diversity associated with key AMR genes, significant chromosomal integration of ARGs and presence of multiple copies in different contexts 149
5.3.6 Comparison with wider plasmid datasets highlights undersampled plasmid diversity, more widespread interspecies and inter-niche plasmid sharing, and the potential for carbapenemase dissemination amongst “high-risk” plasmid groups 156
5.3.7 Factors predictive of plasmid group “success” 157

5.3.8 Machine learning allows risk stratification of plasmids 160

5.4 Discussion 164

6. Conclusions and future work 168

Bibliography 176

Abstract

The incidence of bloodstream infections (BSIs) caused by Gram-negative bacteria has increased over the past decade with substantial associated morbidity and mortality. Furthermore the association of these bacteria with plasmids and other mobile genetic elements which facilitate rapid dissemination of genes causing antibiotic resistant phenotypes is a cause for global concern. Whilst infection control interventions and antimicrobial stewardship initiatives have had some success in controlling the incidence of other key pathogens causing bloodstream infections, to date there has been little progress to reverse the trends observed for *E. coli* and *Klebsiella* spp in particular.

In this thesis I combine large-scale whole genome sequencing and electronic healthcare record datasets to try to better understand factors driving the increasing incidence of Gram-negative bloodstream infections with a particular focus on *E. coli* and *Klebsiella* spp. In *E. coli*, I observe a clonal population structure with four dominant sequence types (ST 131/95/73 and 69) comprising 48% of all isolates. In contrast, most *Klebsiella* spp. isolates belonged to sporadic STs and the most common ST (ST490) only comprised 5% of isolates. I find evidence of a relative increase in the carriage of antimicrobial resistance genes in *E. coli* whereas the converse was true for *Klebsiella* spp.. I find that 10-valent O-antigen targeted vaccines would be predicted to provide high levels of coverage against both *E. coli* (2347/3278, 72%) and *K. pneumoniae* (541/549, 99%) BSIs. I conduct a genome wide association study to explore genomic factors associated with patient mortality discovering significant associations for three genes (*papGII* OR 0.94, 95%CI 0.91-0.96, $p=5.22x10^{-7,}$ *hin* OR 1.23, 95%CI 1.15-1.33, $p=3.17x10^{-6}$ and a hypothetical protein OR1.34, 95%CI 1.23-1.46, $p=5.22x10^{-7}$). Finally I utilise recent advances in long read sequencing to characterise the plasmidome associated with these pathogens and the key antimicrobial resistance/virulence genes which they carry. The key finding from

this is the similarity between plasmids carrying antimicrobial resistance genes and those with no such genes leading me to introduce the concept of “high risk” plasmid groups.

A common recurring theme is the importance and challenge of appropriate study design in molecular epidemiology. Whilst high costs have previously necessitated isolate selection by e.g. antimicrobial phenotype, the work I have conducted in this thesis demonstrates the limitations and dangers of extrapolating from such an approach. I conclude that the linkage of large scale sequencing and clinical datasets affords opportunities to better understand the epidemiology of pathogens, test hypothesis about their transmission and ecology, investigate the likely efficacy of possible intervention strategies and discover new insights into their biology. Global collaboration to develop secure and harmonised methodologies for sharing granular metadata in addition to the vast quantities of sequencing data already available would significantly accelerate future such work, to the benefit of patients across the world.

0 Preface

0.1 Acknowledgments

I am hugely grateful for the amazing support, resources, encouragement and mentorship provided to me by my fantastic four supervisors: Derrick Crook, Tim Peto, Sarah Walker and Nicole Stoesser. They have all contributed hugely to this work and provided scrutiny, encouragement and mentorship in their own unique way which has made this period of my life so rewarding and productive. The work described in this thesis would not have been possible without their foresight and vision as well as the huge amount of methodological development which they have overseen over the years. Lots of people both in the Oxford University Hospitals Clinical Microbiology laboratory and the Modernising Medical Microbiology research laboratory have contributed to the collection, storage, processing and sequencing of isolates over the years and I am very grateful for

these efforts without which the work in this thesis would never have been possible. The amazing colleagues that I have had the privilege of working with in Oxford and further afield have taught me so much and challenged my thinking in a way that I hope will make me a better scientist and clinician. I would also like to pay tribute the patience of my fiancée Helen Lilley who has had to listen to me rant about much of the enclosed work for the past three years and is now able to give a much more convincing summary of my work than I of hers. My parents were instrumental in providing the early opportunities in my life and inspiring me to go into medicine and become interested in research and I am very grateful for their continuous support. Finally I would like to pay tribute to the patients whose data is analysed in this thesis, many of whom had serious illness or died as a result of their infections; their contribution willingness to participate in research for the greater good is inspiring and encapsulates what our profession is ultimately all about.

0.2 Funding

This work was supported by a Medical Research Council Clinical Research Training Fellowship. Additional funding was provided by the National Institute for Health Research Health Protection Research Unit (NIHR HPRU) in Healthcare Associated Infections and Antimicrobial Resistance at the University of Oxford in partnership with Public Health England (PHE) (NIHR200915). Computation used the Oxford Biomedical Research Computing (BMRC) facility, a joint development between the Wellcome Centre for Human Genetics and the Big Data Institute supported by Health Data Research UK and the NIHR Oxford Biomedical Research Centre.

0.3 Ethics Statements

The Infections in Oxfordshire Research Database (IORD; https://oxfordbrc.nihr.ac.uk/research-themes-overview/antimicrobial-resistance-and-modernising-microbiology/infections-in-oxfordshire-research-database-iord/) has generic Research Ethics Committee, Health Research Authority and Confidentiality Advisory Group approvals (19/SC/0403, 19/CAG/0144) which facilitate the pseudo-anonymised linkage of routinely collected NHS electronic healthcare record data from the Oxford University Hospitals NHS Foundation Trust Clinical Systems Data Warehouse and research data (e.g. sequencing data) from the Antimicrobial Resistance and Modernising Microbiology Theme of the Oxford NIHR Biomedical Research Centre, Oxford. IORD links records by a specific, random, number ensuring that no patient-identifiable information is shared with researchers using this resource. We sequenced bacterial isolates from bloodstream infections (BSIs) that are routinely stored by the John Radcliffe Hospital Microbiology laboratory (Oxford, UK). In the UK, bacterial isolates (such as those sequenced in this study) routinely cultured from human clinical samples do not require ethical approval for analysis under the provisions of the Human Tissue Act as they do not contain any material considered to be human tissue.

0.4 Index of Figures and Tables

All figures have been additionally deposited in full resolution at: https://figshare.com/projects/Sam_Lipworth_DPhil/138853.

Figure 1 - Population dynamics of *E. coli* (A, B) and Klebsiella spp. (C, D) STs over time. 47

Figure 2 - Gene presence/absence heatmap showing AMR gene presence/absence against the core genome phylogeny for 3461 Oxfordshire *E. coli*..... 48

Figure 3 - Time-scaled phylogenies for A) ST131 B) ST95 C) ST73 and D) ST69. 52

Figure 4 – Minimal evidence of transmission within nursing homes for *E. coli*..... 55
Figure 5 - Gene presence/absence heatmap showing AMR gene presence/absence against the core genome phylogeny for *Klebsiella* spp..... 59
Figure 6 – Genomic relatedness of community/healthcare onset *E. coli* isolates and those from the same/different countries..... 61
Figure 7 - Manhattan plots of a pangenome wide association study of the association of genes with community/healthcare associated onset in *E. coli*. 64
Figure 8 - Proportions of presumed infectious foci for CA (left panel) and HA (central panel) *E. coli* BSI. Source attribution by year for *E. coli* (right panel)..... 67
Figure 9 - Changes in the incidence of clinically relevant groups of antimicrobial resistance genes considering total effects of resistance to individual antibiotics, regardless of mechanism, over the study period. 70
Figure 10 - Timescaled phylogeny of *Klebsiella pneumoniae* ST490 with a heatmap of AMR genes..... 71
Figure 11 - Virulence elements are structured by ST/Phylogroup in *E. coli*. 73
Figure 12 - Phylogenetic tree of *Klebsiella* spp annotated with species (inner ring) and virulence score (outer ring). 74
Figure 13 - Distribution of virulence factors and antimicrobial resistance classes across community acquired (left) and healthcare associated (right) *Klebsiella* spp. isolates..... 75
Figure 14 Plasmid replicon types structured by species but with substantial apparent intra-species diversity 77
Figure 15 – No evidence of intra-species structure by replicon type..... 78
Figure 16 - Networks of genes/plasmids/insertion sequences commonly co-occurring in *E. coli.* 80
Figure 17 - Networks of genes commonly co-occurring in *Klebsiella* spp. 81
Figure 18 Limited O-antigen diversity in Oxfordshire *E. coli* and *K. pneumoniae* BSIs with a broadly stable population structure over time..... 93
Figure 19 – Distribution of AMR phenotypes by *E. coli* O-antigen serotype..... 95

Figure 20 - Distribution of AMR phenotypes by K. pneumoniae O-antigen serotype 98
Figure 21 – Sequencing/bioinformatic workflow 108
Figure 22 – Phylogenetic tree of isolates included in the study 117
Figure 23 - ProteInfer prediction of the possible function of the hypothetical protein found to be significantly associated with increased 30-day mortality in the Pyseer GWAS. 122
Figure 24 – Details of isolates included/excluded from the chapter 5 study..... 133
Figure 25 – Assembly methods used 134
Figure 26 - Optimisation of threshold and method using to assign plasmids to plasmid groups. 137
Figure 27 – Distribution of plasmids by species and ARG carriage 141
Figure 28 - Phylogenetic distribution of the most common (n≥10 members) plasmid groups (n=33 groups) and the gene content of these 144
Figure 29 – Plasmidomes are constrained by species but not ARG content, year of isolation or infection source (portal of entry)..... 145
Figure 30 - Overlap between plasmid and chromosomal pangenomes. 147
Figure 31 – Antibiotic resistance gene (ARG) carrying plasmids are not a distinct entity 148
Figure 32 – Similarity of *bla*$_{CTX\text{-}M\text{-}15}$ gene flanking regions 152
Figure 33 - Core genome phylogeny and gene presence/absence heatmap (for genes in the AMRFinder/PlasmidFinder databases) for my group 3 plasmids. 153
Figure 34 - Core genome phylogeny and gene presence/absence heatmap (for genes in the AMRFinder/PlasmidFinder databases) for my group 2 plasmids. 154
Figure 35 - Nested genetic complexity associated with bla$_{CTX\text{-}M\text{-}15}$ mobilisation. 155
Figure 36 - Plasmids carrying carbapenemase genes are highly similar to plasmids without these genes found in Oxfordshire BSIs. 157

0.5 Papers arising directly from this work

Peer reviewed

Lipworth, S., Pickford, H., Sanderson, N., Chau, K.K., Kavanagh, J., Barker, L., Vaughan, A., Swann, J., Andersson, M., Jeffery, K. and Morgan, M., 2020. Optimized use of Oxford Nanopore flowcells for hybrid assemblies. *Microbial genomics*, *6*(11).

Lipworth S, Vihta KD, Chau K, Barker L, George S, Kavanagh J, Davies T, Vaughan A, Andersson M, Jeffery K, Oakley S. Ten-year longitudinal molecular epidemiology study of Escherichia coli and Klebsiella species bloodstream infections in Oxfordshire, UK. Genome medicine. 2021 Dec;13(1):1-3.

Lipworth, S., Vihta, K.D., Chau, K.K., Kavanagh, J., Davies, T., George, S., Barker, L., Vaughan, A., Andersson, M., Jeffery, K. and Oakley, S., 2021. Ten years of population-level genomic Escherichia coli and Klebsiella pneumoniae serotype surveillance informs vaccine development for invasive infections. *Clinical Infectious Diseases*, *73*(12), pp.2276-2282.

Matlock, W*., Lipworth, S*., Constantinides, B., Peto, T.E., Walker, A.S., Crook, D., Hopkins, S., Shaw, L.P. and Stoesser, N., 2021. Flanker: a tool for comparative genomics of gene flanking regions. *Microbial genomics*, *7*(9).

*Joint first author

Preprints currently under review

Lipworth, S*., Matlock, W.*, Shaw, L., Rodger, G., Chau, K., Barker, L., George, S., Kavanagh, J., Davies, T., Vaughan, A. and Andersson, M., 2022. The mobilome

associated with Gram-negative bloodstream infections: A large-scale observational hybrid sequencing based study. *MedRxiv preprint*.

*Joint first author

Lipworth, S., Vihta, K.D., Davies, T., Wright, S., Tabirao, M., Chau, K., Vaughan, A., Kavanagh, J., Barker, L., George, S. and Segal, S., 2021. Molecular epidemiology and antimicrobial resistance phenotype of paediatric bloodstream infections caused by Gram-negative bacteria in Oxfordshire, UK. *medRxiv*,preprint pp.2021-06.

0.6 Other papers arising in the course of this work

Peer reviewed

Lipworth, S., Hough, N., Weston, N., Muller-Pebody, B., Phin, N., Myers, R., Chapman, S., Flight, W., Alexander, E., Smith, E.G. and Robinson, E., 2021. Epidemiology of Mycobacterium abscessus in England: an observational study. *The Lancet Microbe*, *2*(10), pp.e498-e507.

Lipworth, S., Peto, T.E., Crook, D.W., Walker, A.S. and Stoesser, N., 2021. Antimicrobial resistance genes and clonal success in Escherichia coli isolates causing bloodstream infection. *The Lancet Microbe*, *2*(10), p.e492.

Kizny Gordon, A., Phan, H.T.T., Lipworth, S.I., Cheong, E., Gottlieb, T., George, S., Peto, T.E.A., Mathers, A.J., Walker, A.S., Crook, D.W. and Stoesser, N., 2020. Genomic dynamics of species and mobile genetic elements in a prolonged bla IMP-4-associated

carbapenemase outbreak in an Australian hospital. *Journal of Antimicrobial Chemotherapy*, *75*(4), pp.873-882.

Mashar, M., Kilgour, J., Nanapragasam, H. and Lipworth, S., 2020. Academic medicine: the continuing challenges. *The Clinical Teacher*, *17*(1), pp.81-85.

Dagens A, Sigfrid L, Cai E, Lipworth S, Cheng V, Harris E, Bannister P, Rigby I, Horby P. Scope, quality, and inclusivity of clinical guidelines produced early in the covid-19 pandemic: rapid review. bmj. 2020 May 26;369.

Lipworth B, Chan R, Lipworth S, Kuo CR. Weathering the cytokine storm in susceptible patients with severe SARS-CoV-2 infection. The Journal of Allergy and Clinical Immunology: In Practice. 2020 Jun 1;8(6):1798-801.

Constantinides B, Chau KK, Quan TP, Rodger G, Andersson MI, Jeffery K, Lipworth S, Gweon HS, Peniket A, Pike G, Millo J. Genomic surveillance of Escherichia coli and Klebsiella spp. in hospital sink drains and patients. Microbial genomics. 2020 Jul;6(7).

Folegatti PM, Ewer KJ, Aley PK, Angus B, Becker S, Belij-Rammerstorfer S, Bellamy D, Bibi S, Bittaye M, Clutterbuck EA, Dold C. Safety and immunogenicity of the ChAdOx1 nCoV-19 vaccine against SARS-CoV-2: a preliminary report of a phase 1/2, single-blind, randomised controlled trial. The Lancet. 2020 Aug 15;396(10249):467-78.

Lipworth B, Kuo CR, Lipworth S, Chan R. Inhaled corticosteroids and COVID-19. American Journal of Respiratory and Critical Care Medicine. 2020 Sep 15;202(6):899-900.

Ramasamy MN, Minassian AM, Ewer KJ, Flaxman AL, Folegatti PM, Owens DR, Voysey M, Aley PK, Angus B, Babbage G, Belij-Rammerstorfer S. Safety and immunogenicity of ChAdOx1 nCoV-19 vaccine administered in a prime-boost regimen in young and old adults (COV002): a single-blind, randomised, controlled, phase 2/3 trial. The Lancet. 2020 Dec 19;396(10267):1979-93.

Voysey M, Clemens SA, Madhi SA, Weckx LY, Folegatti PM, Aley PK, Angus B, Baillie VL, Barnabas SL, Bhorat QE, Bibi S. Safety and efficacy of the ChAdOx1 nCoV-19 vaccine (AZD1222) against SARS-CoV-2: an interim analysis of four randomised controlled trials in Brazil, South Africa, and the UK. The Lancet. 2021 Jan 9;397(10269):99-111.

Barrett JR, Belij-Rammerstorfer S, Dold C, Ewer KJ, Folegatti PM, Gilbride C, Halkerston R, Hill J, Jenkin D, Stockdale L, Verheul MK. Phase 1/2 trial of SARS-CoV-2 vaccine ChAdOx1 nCoV-19 with a booster dose induces multifunctional antibody responses. Nature medicine. 2021 Feb;27(2):279-88.

Ewer KJ, Barrett JR, Belij-Rammerstorfer S, Sharpe H, Makinson R, Morter R, Flaxman A, Wright D, Bellamy D, Bittaye M, Dold C. T cell and antibody responses induced by a single dose of ChAdOx1 nCoV-19 (AZD1222) vaccine in a phase 1/2 clinical trial. Nature medicine. 2021 Feb;27(2):270-8.

Voysey M, Clemens SA, Madhi SA, Weckx LY, Folegatti PM, Aley PK, Angus B, Baillie VL, Barnabas SL, Bhorat QE, Bibi S. Single-dose administration and the influence of the timing of the booster dose on immunogenicity and efficacy of ChAdOx1 nCoV-19 (AZD1222) vaccine: a pooled analysis of four randomised trials. The Lancet. 2021 Mar 6;397(10277):881-91.

Preprints currently under review

Matlock, W.*, Lipworth, S.*, Chau, K.K., AbuOun, M., Barker, L., Kavanagh, J., Andersson, M., Oakley, S., Morgan, M., Crook, D.W. and Read, D.S., 2022. Plasmid overlap and evolution between Enterobacterales isolates from bloodstream infections and non-human compartments. *bioRxiv*.

*Joint first author

Davies, T.J., Swan, J., Sheppard, A.E., Pickford, H., Lipworth, S., AbuOun, M., Ellington, M., Fowler, P.W., Hopkins, S., Hopkins, K.L., Crook, D.W., Peto T.E.A., Anjum M.F., Walker A.S., ans Stoesser, N. 2021. Discordance between different bioinformatic methods for identifying resistance genes from short-read genomic data, with a focus on Escherichia coli. *bioRxiv*.

Stoesser, N., George, R., Aiken, Z., Phan, H.T., Lipworth, S., WYLLIE, D.H., Quan, P., Mathers, A.J., de Maio, N., Seale, A.C. and Eyre, D.W., 2021. Genomic epidemiology and longitudinal sampling of ward wastewater environments and patients reveals complexity of the transmission dynamics of blaKPC-carbapenemase-producing Enterobacterales in a hospital setting. *medRxiv*.

Sanderson, N., Kapel, N., Rodger, G., Webster, H., Lipworth, S., Peto, T., Crook, D. and Stoesser, N., 2022. Comparison of R9. 4.1/Kit10 and R10/Kit12 Oxford Nanopore flowcells and chemistries in bacterial genome reconstruction. *bioRxiv*.

Webb E, Michelen M, Rigby I, Dagens A, Dahmash D, Cheng V, Joseph R, Lipworth S, Harriss E, Cai E, Nartowski R. An evaluation of global Chikungunya clinical management guidelines-a systematic review. medRxiv. 2022 Jan 1.

Chau KK, Matlock W, Constantinides B, Lipworth S, Newbold L, Tipper H, Goodall T, Brett H, Hughes J, Crook DW, Eyre DW. Evaluation of metagenomic, 16S rRNA gene and ultra-plexed PCR-based sequencing approaches for profiling antimicrobial resistance gene and bacterial taxonomic composition of polymicrobial samples. bioRxiv. 2022 Jan 1.

Webb, Eika, Ishmeala Rigby, Melina Michelen, Andrew Dagens, Vincent Cheng, Amanda Rojek, Dania Dahmash, *et al*. 2022. “Availability, Scope, and Quality of Monkeypox Clinical Management Guidelines Globally: A Systematic Review.” *medRxiv*. https://doi.org/10.1101/2022.06.09.22276219.

0.7 Declaration of contribution to thesis

I, Samuel Lipworth, designed and conducted all of the analyses presented in this thesis with appropriate support from my supervisors and colleagues. I hereby declare and gratefully acknowledge the assistance I have received from others in relation to this work below.

0.7.1 Chapters 2, 3 and 4

All laboratory work including microbiological culture, DNA extraction and sequencing was performed by members of the Modernising Medical Microbiology (MMM) group over the past 10 years. In particular, I acknowledge the assistance of Hayleah Pickford, Kevin Chau, Leanne Barker, Sophie George, James Kavanagh, Alison Vaughan, Tim Davies and Dai Griffiths. The use of these isolates was only possible with the cooperation and support of the OUH clinical microbiology laboratory and, in particular, I am grateful for the support of Monique Andersson, Katie Jeffery, Sarah Oakley and Marcus Morgan as well as the many biomedical scientists who have been involved in routine clinical work over the years. I acknowledge the work of the team at the Wellcome Trust Centre for Human Genetics who undertook library preparation and sequencing on our behalf. I am grateful to Karina-Doris Vihta who provided me with her R code to help clean the data as well as providing general statistical advice.

0.7.2 Chapter 5

Gillian Rodger performed extraction and sequencing for 24 of the isolates. Sophie George taught me how to perform ONT library prep and Hayleah Pickford how to perform DNA extraction using the genomic tip method. Anthony Brown performed DNA extraction on the EasyMag machine. All library preparation and sequencing except that performed by Gillian Rodger as above was performed by me. The MMM informatics team including Nick Sanderson and Jez Swann provided invaluable infrastructure support which made data acquisition much easier. Will Matlock provided me with the method for plasmid clustering, extensive helpful discussion and reviewed several versions of the eventual manuscript as well as working with me to produce the Flanker programme on which this work is partly based. Will and I jointly conceived the idea for flanker and I developed the initial code which was subsequently revised and improved by Will with further collaborative iterative improvements involving both of us. Karina-Doris Vihta generously reviewed the statistical methods and provided advice around the machine learning methods as well as sharing with me some example code she had used for an unrelated project.

0.8 Abbreviations

aHR – Adjusted Hazard Ratio

AIC – Akaike Information Criterion

AMR – Antimicrobial Resistance

aOR – Adjusted Odds Ratio

API – Analytical Profile Index

ARG – antimicrobial resistance gene

AUC – Area Under the (Receiver Operating Characteristic) Curve

BP – Base Pairs

C. difficile – Clostridioides difficile

CA- Community Associated

US CDC – United States Centers for Disease Control and Prevention

CI – Confidence Interval

COPD – Chronic Obstructive Pulmonary Disease

CPE – Carbapenemase Producing Enterobacteriaceae

DAPC – Discriminant analysis of principal components

DNA – Deoxyribonucleic acid

E. coli – Escherichia coli

ED – Evolutionary Distinctiveness

ESBL – Extended Spectrum Beta-lactamase

GC – Guanine-Cytosine content

GNBSI – Gram-negative bloodstream infection, BSI – bloodstream infection

GWAS – Genome wide association study

HA – Healthcare Associated

IORD – Infections in Oxfordshire Research Database

IRRy – Incidence Rate Ratio per Year

IS – Insertion Sequence

K. pneumoniae – Klebsiella pneumoniae

LPS - Lipopolysaccharide

MALDI-TOF – Matrix-assisted laser desorption/ionization-time of flight

MDR – Multidrug Resistant

MDS – Multidimensional Scaling

MGE – Mobile Genetic Element

MLST – Multilocus Sequence Type

MRSA – Methicillin Resistant Staphylococcus Aureus

MSSA – Methicillin Sensitive Staphylococcus Aureus

NMI – Normalised Mutual Information

ONT – Oxford Nanopore Technologies

OUH – Oxford University Hospitals NHS Foundation Trust

PCR – Polymerase Chain Reaction

PGWAS – Pangenome-wide Association Study

PTU – Plasmid Taxonomic Unit

SNP – Single Nucleotide Polymorphism

ST – Sequence Type

UPEC – Uropathogenic *Escherichia coli*

UTI – Urinary Tract Infection

1. Introduction

In 1884, Hans Christian Gram was searching for a way to identify bacteria which he had isolated in post-mortem specimens from the lungs of patients who had had pneumonia. He devised a method that selectively stained these ovoid shaped bacteria (now known to be *Streptococcus pneumoniae*) a dark blue/violet colour but left the surrounding human cells clear. His breakthrough moment came when he realised that adding ethanol to typhoid bacilli on which he had applied his stain caused them to become discoloured[1]. Remarkably, this nearly 140-year old technique continues to underpin diagnostic microbiology and clinical infectious diseases today. It was later discovered that this difference in staining characteristics was caused by the fact that the peptidoglycan layer of the cell wall is thinner in Gram-negative bacteria which allows the stain to be decolourised on the addition of alcohol.

Writing in the New England Journal of Medicine in 1959, David Rogers noted that "The hope that antimicrobial drugs would abolish infections as a cause of death has not been adequately realised"[2]. In a study of infectious causes of death in his New York hospital, Rogers found substantial declines in pneumococcal/streptococcal infections between the pre-antibiotic period (1938-1940) and the post antimicrobial period (1957-1958). In contrast there was an "impressive" increase in the number of life-threatening enterobacterial infections. Since the early days of antibiotics and the widespread availability of penicillin in the early 1940s, finding agents with good activity against Gram-negative bacteria has been a constant challenge. In recent years therapeutic options have been further limited by the global dissemination of multi-drug resistant (MDR, sometimes defined as carrying resistance to three or more antibiotic classes[3]) clones (e.g. ST131 in *E. coli*[4], clonal group 258 in *Klebsiella pneumoniae*[5]) and plasmids which facilitate the

rapid dissemination of antimicrobial resistance genes (ARGs) within and between species[6].

History has in some ways repeated itself in recent years. During the 1990s and early part of this millennium, Gram-positive infections caused the majority of bloodstream infections in the UK[7,8]. In particular, Methicillin-resistant *Staphylococcus aureus* (MRSA)[9] and *Clostridium difficile* (*C. difficile*)[10,11] were associated with some very high-profile outbreaks. Several measures including improved hygiene in hospitals, increased focus on antimicrobial stewardship and greater awareness of intravascular line related infections eventually had a dramatic effect to reduce their incidence (although there has been a sustained increase in methicillin susceptible *Staphylococcus aureus* (MSSA) bloodstream infections (BSIs) more recently)[12]. In contrast the incidence of Gram-negative infections has continued to grow over the past ten years to the point where they are now the predominant cause of BSIs in the UK, and many other settings[12]. Furthermore, *E. coli* and *Klebsiella pneumoniae* are estimated globally to be the 1st and 3rd leading causes of fatal antimicrobial resistant infection[13], and carbapenem-resistant/extended spectrum beta-lactamase producing (ESBL) *Enterobacteriaceae* are regarded as critical priority pathogens for the development of new antibiotics by the World Health Organisation[14]. Whilst there have been a range of government targets designed to attenuate this[15], to date these have had limited impact.

1.2 Direction of this thesis

Whilst there are a very large number of potential species encompassed within the title of this thesis, I have chosen to focus on *E. coli* and *Klebsiella* spp. as these are by far the most common causes of Gram-negative bloodstream infections in the UK. To try to understand factors influencing their increasing incidence I will utilise a technique that has become known as “molecular epidemiology” which uses data from whole genome

sequencing to enrich the much older discipline of epidemiology to try to better understand patterns of disease caused by pathogens. In the next section I will briefly review the relevant history behind these combined disciplines and technologies.

1.3 A brief history of epidemiology

Epidemiology is “the study of how often diseases occur in different groups of people and why”[16]. It is worth reflecting on the origins and history of epidemiology, as the basic principles and methods, whilst latterly enhanced by new technologies such as whole genome sequencing, still underpin modern practice. The history of epidemiology can be traced back as far as Hippocrates (c. 400BC), but one of the most famous applications was John Snow’s efforts to map cholera cases in London in 1854[17]. By plotting cases on a map, Snow realised that patients with cholera had common exposure to a water pump on Broad Street. Subsequently the pump handle was removed and the outbreak ended. In this landmark early example, Snow had a hypothesis (that water was involved in the transmission of cholera) and through meticulous investigation, gathered the data to support this and presented it in a way that depicted a clear and indisputable story.

The field of epidemiology enjoyed multiple subsequent successes during the 20th century which saved countless lives and transformed our understanding of disease. In one such study, Doll and Hill asked a group of London hospitals to notify them about lung cancer cases[18]. They interviewed these cases, as well as age and sex matched controls from the same hospitals, to find out about their exposure to various environmental exposures that might be associated with lung cancer. Through this investigation they demonstrated that in both men and woman, the proportion of lung cancer patients who smoked was significantly higher than that of non-smokers and that there appeared to be a dose-response relationship. Peto and colleagues subsequently demonstrated that smoking

cessation, even after many years of prior exposure, was an effective intervention to reduce the risk of lung cancer[19]. This work eventually led to the relatively strict tobacco controls in place in the UK today. Similarly a detailed study of a cohort of residents from Framingham, Massachusetts has underpinned the way we think about cardiovascular risk today and has been crucial to informing primary and secondary prevention strategies for ischaemic heart disease[20].

In the field of infectious disease, one of the best examples of the application of epidemiology to identify an infectious agent and subsequently its common exposures and mechanism of transmission is human immunodeficiency virus (HIV). In 1981, the United States Centres for Disease Control and Prevention (US CDC) reported 5 cases of *Pneumocystis carinii* (now known as *Pneumocystis jirovecii*) pneumonia amongst a group of otherwise apparently well young men in Los Angeles[21]. Subsequently the US CDC published a case definition describing a "disease at least moderately predictive of a defect in cell-mediated immunity, occurring in a person with no known cause for diminished resistance to that disease" with an accompanying list of associated opportunitistic infections/malignancies[22]. Subsequent efforts to compare exposures between those with and without this clinical syndrome demonstrated that it was (at least in this early period) predominantly spread amongst men who had sex with men, leading to the realisation that this was a blood-borne, sexually transmitted pathogen later characterised as HIV.

1.4 A brief history of whole genome sequencing

In February 1953, Francis Crick pronounced to a small pub on Bene't Street in Cambridge that he and James Watson had "discovered the secret of life". In fact, in work which relied heavily on contributions from Rosalind Franklin and others, they had described for the first time the structure of DNA molecules[23]. The real biological secrets encoded in these

structures would however remain hidden for decades because there was as yet no way to read these molecules quickly and at scale. This was a problem later to be at least partially solved by DNA sequencing, a technology that has evolved with staggering pace over the past 30 years. The first *E. coli* genome was published in 1997 following laboratory processes which took many years to complete[24]. There are now tens or even hundreds of thousands of *E. coli* genomes available to freely download from public repositories due to huge reductions in cost and increases in throughput which have been achieved as sequencing technology has evolved. I will now briefly review this technological evolution and the basic theory behind the sequencing technologies used in this thesis.

One of the first significant breakthroughs in sequencing was the development of so-called "Sanger sequencing" in 1977[25]. In this technique, primers and DNA polymerase are first used to extend the strand adjacent to a region of interest. A ribonucleotide with an oxygen atom removed (called a dideoxynucleotide) is then used to terminate the expanding chain of nucleotides and the terminal nucleotide is then labelled with a fluorescent probe (though originally Sanger sequencing was performed with radioactive dideoxynucleoside triphosphate (ddNTP) in four separate sequencing reactions). The resulting molecules are separated according to weight by electrophoresis and a base call can be made for every position by reading the colour of fluorescence at each 1 base interval. Such "first generation" sequencing techniques are capable of producing reads of around one kilobase. After sequencing many overlapping fragments in what is called "shotgun sequencing", one can assemble these *in silico* to produce longer stretches of contiguous sequence (referred to as contigs) and eventually complete genomes.

"Second generation" techniques, such as Solexa/Illumina sequencing, involved the ligation of "adapters" to target DNA regions. These are added to a flowcell (a glass slide)

containing complementary oligonucleotides which bind the adapters and a PCR reaction occurs to clonally expand the template in a process called “bridge amplification”. Sequencing occurs by synthesis with the reversible addition of fluorescent nucleotides which allow the base at each position to be identified. An advantage of this approach is that a secondary step can be introduced to sequence the complementary strand leading to paired-end reads which are helpful in improving the accuracy of resulting base calls when mapping the reads to a reference. This type of sequencing is now dominated by Illumina platforms (and hereafter referred to as “Illumina” or “short-read sequencing”) and typically produces read lengths of around 150-250bp, though platforms such as the Roche 454 were capable of producing longer reads of around 1000bp. Illumina sequencing is still very widely used due to its relatively cheap cost, potential for very high throughput and high-per-read accuracy.

A “third generation” of sequencing devices (also known as “long-read sequencers”) has emerged over the past decade which have several significant advantages. One of the most widely used is the Oxford Nanopore Technologies system (hereafter referred to as “ONT/Nanopore sequencing”). ONT devices use a completely different approach where DNA strands are passed through protein nanopores causing a change to the electrical current proportional to the length of the nucleic acid allowing its identity to be assessed by a connected chip. This approach produces reads which are often thousands, or even hundreds of thousands of bases long. These reads can be analysed in real-time which has the potential to be very advantageous for some applications (e.g. clinical metagenomics). The technology is very accessible since it has a low setup cost for a small sequencing unit (~1000 USD) and can be extremely portable, allowing sequencing to take place in remote environments with minimal laboratory equipment or in local laboratories near to the point of patient care[26]. There are other long-read sequencing approaches, most notably PacBio sequencing, which uses ligated hairpin adapters

followed by measurement of light emitted as nucleotides are incorporated using a polymerase in millions of tiny wells in a flowcell. However, the infrastructural costs of PacBio platforms are significant and these sequencers are typically only found in major sequencing hubs.

Chapters 2,3 and 4 highlight the utility of Illumina short-read approaches which enabled our research group to sequence a large collection of *E. coli* and *Klebsiella* spp. isolates from BSIs relatively quickly and cheaply. The resulting assemblies are very accurate at the nucleotide level, permitting informative epidemiological inferences from marker genes (e.g. the housekeeping genes used in multi-locus sequence type (MLST) schemes and antibiotic resistance/virulence genes). Chapter 5 however demonstrates the limitations of this method, namely the inability to create complete circularised assemblies using short reads alone, which has led to the huge recent interest in third-generation technologies.

The Australian bioinformatician Torsten Seemann (whose tools I rely on heavily for much of the work in this thesis) has coined a famous saying that "It is impossible to resolve repeats of length S unless you have reads longer than S."[27] This is the fundamental truism that explains why assemblies produced using short read data alone are generally highly fragmented compared to those produced with ONT long-read data which are often complete (i.e. the chromosome and all plasmids are circularised). For many applications (e.g. those in chapters 1/2/3) highly fragmented assemblies provide sufficient resolution; however they are of very limited use for accurately examining the genetic context of specific genes and the mechanisms by which they mobilise[28]. The most obvious example is the investigation of whether important genes (e.g. those encoding antimicrobial resistance) are located on plasmids or chromosomes. If for example there is an outbreak of carbapenem-resistant (e.g. *bla*$_{\text{OXA-48}}$-containing) *E. coli* and *Klebsiella pneumoniae*

within an intensive care unit, it might be useful to understand whether this is a single outbreak of a plasmid that is spreading between species or whether there are two (or more) clones which might have different relevant epidemiological exposures. These two scenarios might require different intervention strategies and can only be reliably distinguished from each other using long-read sequencing.

1.5 Molecular epidemiology

In very simple terms an epidemiologist is somebody that “counts (cases), divides (by an appropriate denominator) and compares (between groups and/or over time)[29]. In order to count cases, the epidemiologist needs a case definition. Typically this might include a clinical syndrome and a laboratory diagnosis. For example the US CDC has a viral haemorrhagic fever (VHF) case definition which has clinical (fever accompanied by one or more of a list of symptoms e.g. headache/myalgia/vomiting/rash etc), epidemiological (e.g. travel to an endemic area or contact with bodily fluids of an infected person) and laboratory criteria (e.g. detection of VHF antigens by ELISA/immunohistochemistry or of virus by RT-PCR)[30]. A diagnosis is suspected based on clinical and epidemiological criteria but is only confirmed when the laboratory criteria are met.

As blood is normally a sterile site, isolation of bacteria from blood cultures is mostly abnormal and a sign of infection, although positive results can also be caused by contamination at the point of sampling/processing. Further identification of pathogens using e.g. Gram staining/MALDI-TOF can give some indication as to how the isolate ended up in the bloodstream. Nevertheless, even for a single species, the disease caused may be rather heterogenous with very different relevant epidemiological exposures. There may be too much noise to identify relevant exposures where a case is defined at the species-level alone. The promise of molecular epidemiology is to utilise genomics to

further refine the laboratory case definition (by stratifying cases according to genomic characteristics) in the hope that this might increase the power of studies to identify common exposures.

Let us consider a hypothetical example case-control study involving two clades of *E. coli* where it later turns out that long-term urinary catheter use is particularly associated with clade A. We interview 20 patients with *E. coli* BSI and 20 age and sex matched controls. 8/20 (40%) *E. coli* BSIs have a long-term catheter compared to 6/20 (30%) controls. If however we have a molecular test (e.g. PCR, or genomics) available to further delineate isolates to clade A and B, we might discover that 8/10 (80%) clade A isolates come from patients with a long-term catheter compared to 0/10 (0%) clade B. One might then generate a hypothesis that clade A has somehow adapted to better colonise long-term urinary catheters and/or cause infection, which might lead to molecular insights into pathogenesis as well as an opportunity to prevent a proportion of BSIs.

Molecular epidemiology can also help to estimate the probability that isolates from several patients belong to a common transmission chain/point-source outbreak. A famous example of this occurred when cardiothoracic surgeons noticed that their patients where getting much higher than expected rates of infective endocarditis due to Mycobacterium chimaera after bypass surgery[31]. When public health authorities later performed a sequencing based investigation, they discovered that isolates collected from patients across Europe were highly genomically similar. Further investigation revealed that the patients had all had exposure to a particular manufacturer's heater cooler unit machines during surgery and that isolates taken from the manufacturer's production site were near identical to those collected from patients. This unveiled a point-source outbreak of isolates causing severe disease which would likely have taken much longer to understand had

sequencing not been available, with major subsequent implications for patients and clinical practice.

1.6 Identification of bacteria in the routine microbiology laboratory

Prior to ~2013, following incubation of blood in BD BACTEC bottles, the routine clinical laboratory at OUH used chromogenic agar and analytical profile index (API) tests to identify the organisms I focus on in this thesis. The former relies on expression of enzymes specific to the target species which activate a chromogenic substrate, turning colonies a characteristic colour. A similar principle is used by API tests where the activities of enzymes characteristic to particular species (e.g. those involved in fermentation of carbohydrate/catabolism of amino acids/proteins) are tested. Matrix-assisted laser desorption/ionization time of flight spectrometry (MALDI-TOF) on the other hand identifies bacteria based on their mass to charge ratio measured in a time of flight spectrometer. Since ~2013, this has been the primary method for initial identification of bacterial species in our laboratory as it is much faster, cheaper and more accurate than API tests.

1.7 Sequence based taxonomy/relatedness

As implied above, sequenced strains can be further subdivided using genetic features. One of the most popular ways of doing this is to use multilocus sequence typing (MLST), which compares the allelic profiles of housekeeping genes to a reference database and based on this profile assigns isolates to a sequence type (ST). *E. coli* can also be classified into phylogroups using an *in silico* PCR method targeting three loci which uses

the schema derived by Clermont *et al*[32]. Whilst both phylotyping and MLST can be conducted using targeted PCR followed by amplicon sequencing, it is now more convenient and faster (especially for unknown species and when one desires to derive multiple characteristics in parallel) to do this *in silico* using whole genome sequencing data. These methods are convenient and easily human-understandable but are also relatively low resolution.

WGS affords much higher resolution into genetic relatedness than targeted genetic typing schemes because it enables comparisons using much more of the genome. Several methods have recently been published to cluster strains based on nucleotide identity from WGS data[33,34]. Furthermore, by aligning reads to a reference and looking for single nucleotide polymorphisms (SNPs), one can compare strains at the highest possible level of resolution. By relating SNP distances between isolates to the rate of evolution for a given species (known as the molecular clock), we can begin to make inferences about the probability of transmission between two individuals or niches[35].

1.8 Taxonomy/History of *E. coli* and *Klebsiella* spp.

E. coli and *Klebsiella* spp. belong to the family *Enterobacteriaceae* from the order Enterobacterales, of the class *Gammaproteobacteria*. They are rod-shaped, facultatively anaerobic and non-spore forming.

1.8.1 *E. coli*

E. coli is named after the German/Austrian paediatrician Theodor Escherich who isolated *E. coli* (which he initially named Bacterium coli commune as it was a common gut commensal) from infant faeces[36]. *E. coli* has traditionally been classified according to the pattern of disease it causes e.g. uropathogenic *E. coli* (UPEC causing urinary tract infections (UTIs)/pyelonephritis), intestinal pathogenic *E. coli* (InPEC, causing diarrhoea which can be further classified according to virulence gene carriage e.g. into Shiga toxin-producing (carrying *stx1*/*stx2*), Enteroinvasive *E. coli*/*Shigella* (carrying the *pINV* plasmid, shiga toxin and enterotoxin on a chromosomal pathogenicity island)) or by its behaviour in animal models (e.g. extra-intestinal *E. coli* (ExPEC)) and commensal strains[37]. However, there is substantial overlap in population structure between these pathotypes. More recently, and predominantly for research/epidemiological purposes, *E. coli* isolates have been further classified using MLST/phylogroup schemes. The major groups causing invasive disease in humans in studies to date belong to phylogroups B2 (STs 131/73/95) and D (ST69)[37]; these isolates generally carry multiple virulence genes and are responsible for the majority of ExPEC disease in humans. Phylogroups A and B1 have been isolated from a wide range of vertebrate species[38] whereas those from other phylogroups are much more rarely seen in clinical isolates. There has been some suggestion that phylogroup G strains carry virulence genes associated with ExPEC and can exhibit multi-drug resistance[39].

ST131 has emerged as the predominant globally important ExPEC clade, variably reported as being responsible for between 10-24% of BSIs across the world[40–43]. It has subsequently been split into three clades (A, B and C, named in chronological order of emergence)[44]. Clade C, frequently associated with ciprofloxacin resistance, has been further sub-divided to sub-clades C0/1/2. *bla*$_{\text{CTX-M-15}}$ and *bla*$_{\text{OXA-1}}$ carriage has been

associated with IncF-type plasmids carried by C2 isolates whereas *bla*$_{CTX\text{-}M\text{-}27}$ carriage has emerged in conjunction with a subgroup of clade C1 isolates. Clade B is comparatively associated with carriage of significantly fewer antimicrobial resistance genes (ARGs)[45]. The reasons for the successful emergence and subsequent rapid global spread of ST131 are still not entirely clear, although various factors including global travel[46], antimicrobial resistance[41,43], effective colonisation of and persistence within the human gut[47,48] are likely to play a role. In comparison to ST131, STs 95 and 73 carry relatively fewer ARGs, though dihydrofolate reductase and beta-lactamase *bla*$_{TEM}$-type genes have been found to be associated with these lineages in some studies[40,42] and there have been reports of MDR ST73 isolates carrying *bla*$_{CTX\text{-}M}$ genes on large F-type plasmids[49]. The most common non-phylogroup B2 ExPEC sequence type is ST69 which was noted to be associated with co-trimoxazole resistant UTIs in the USA in the 1990s[50] and has since been reported to cause between 5-15% of BSIs[41,42] with increasing reports of association with MDR and extended-spectrum beta-lactamase (ESBL)-associated disease[51–53].

1.8.2 *Klebsiella* spp.

Klebsiella is named after the Swiss-German microbiologist Edwin Klebs, though interestingly it was actually Carl Friedlander who first described what was temporarily known as the Friedlander bacillus[54]. It is found in a wide variety of environmental niches (e.g. sewage, soil, surface water, vegetation[55]) and one large study estimated the prevalence of carriage in the general population to be 16.3%[56]. The *Klebsiella* species includes members of the *K. pneumoniae* species complex, members of what used to be known as the *Raoultella* complex, *K. aerogenes* and a group which includes *K. oxytoca*, *K. michiganensis*, *K. pasteurii* and *K. grimontii*[5]. The population structure of *K. pneumoniae* has, for epidemiological purposes, been split into clonal groups which can be summarised as i) causing locally problematic outbreaks (e.g. ST70/323/490 which have

caused MDR outbreaks in Kilifi[57]/Melbourne[58]/Oxford[59] respectively) ii) globally disseminated MDR clones (e.g. clonal group 258)[60] and iii) hypervirulent clones e.g. CG23[5]. The distinctions between these have more recently become blurred with a large number of descriptions of convergence of antimicrobial resistance and hyper-virulence emerging in recent years[61].

1.9 Epidemiology of *E. coli*/*Klebsiella* spp. BSIs

The incidence of Gram-negative bloodstream infections has steadily increased over the past eight years in England (although more recently the SARS-CoV-2 pandemic has changed this trend in a way that is not yet understood). UKHSA data[12] shows that for *E. coli*, reported cases went from 60.4 per 100,000 in 2012/13 to 77 per 100,000 in 2019/20; around 1/3 of cases were nosocomial. Similarly, *Klebsiella* spp. BSIs increased from 17.6 per 100,000 in 2017/18 to 19.6 per 100,000 in 2019/20 but the proportion of nosocomial cases was slightly higher than for *E. coli* (just under half). The incidence for both has tended to be higher in the North (and for *E. coli* the South West) of England and both species show a clear seasonal trend with cases peaking in the summer months[12]. It is possible that ascertainment bias (increased awareness of sepsis through high profile campaigns and a higher number of blood culture samples being taken as a result) might be responsible for some of these trends and we might thus be detecting more mild cases of less clinical significance which previously would not have been counted. Previous work in Oxfordshire suggested this was unlikely since there were no important changes in severity over time[62]. Another possible explanation is more accurate reporting by laboratories in England over time since the introduction of mandatory surveillance, although the same trends have been observed internationally[63,64].

For both *E. coli* and *Klebsiella* spp., the urinary tract is the most frequently reported portal of entry (48.2% for *E. coli* and 33.6% for *Klebsiella* spp. in 2019). Interestingly whilst for both species case numbers are higher in older age groups and men, UKHSA data suggests there has been little change to the age and sex distribution since the beginning of mandatory surveillance (in 2011 and 2017 respectively). The reasons for these increases are unclear and probably multifactorial but could be related to host factors e.g. the ageing population, use of indwelling urinary catheters, inappropriate prescribing and pathogen factors e.g. acquired antibiotic/virulence genes or genetic evolution to occupy either an environmental or anatomical niche.

Data from a large molecular epidemiology study of *E. coli* in England suggested that these cases may be largely caused by the four major globally disseminated STs (131/69/73/95) and that, at least until 2012, the population structure was relatively stable following the introduction and emergence of STs 131/69 in the early 2000s[41]. Smaller studies in Scotland and Australia found similar population structures[40,42]. To my knowledge there were no such longitudinal studies of isolates of *Klebsiella* spp. causing bloodstream infections available prior to the work conducted in this thesis and existing studies have primarily focussed on isolates selected for their AMR/virulence phenotype[61,65–67]. Part of the aim of this thesis (chapter 2) was therefore to better understand any changes in the population structure of *E. coli* that might have occurred since 2012 as well as to characterise that of *Klebsiella* spp. in an unbiased way. By studying the epidemiology of STs responsible for the majority of BSIs, I hoped to be able to inform future efforts to reduce the incidence of these infections.

1.10 Thesis outline

The goal of this thesis is to leverage large whole genome sequencing and electronic healthcare datasets in combination to try to understand factors which might be contributing to the observed increases in *E. coli*/*Klebsiella* spp. bloodstream infections in Oxfordshire and globally. In the first chapter, I conduct a detailed analysis of ten years of (Illumina short read) sequencing data from both species to look for genomic features associated with the observed epidemiological trends. Given that previous studies had identified the relatively recent emergence of globally successful clones, it is entirely possible that further expansion of one of these (or the introduction of a new one) could explain the incidence trends. I also test other hypotheses that might explain these trends including the successful dissemination of one or more virulence/antimicrobial resistance genes, niche adaptation (either anatomical or environmental) and the possible role of plasmids.

E. coli and *Klebsiella* spp. both express antigens on the cell surface and in chapter 3 I explore the possibility that changes in the seroepidemiology of these might be contributing to the observed incidence trends. Furthermore, given the emerging clinical trials of O-antigen targeted vaccines for *E. coli*, I also investigate the proportion of *E. coli* BSIs in Oxfordshire which would (at least theoretically) be covered by these. Whilst no such vaccines exist at the moment for *Klebsiella* spp., I also present data on O and K antigen seroprevalence to estimate how well equivalent vaccines might work and to inform future development.

Given the lack of clear lineage effects associated with incidence trends in chapter 2, chapter 4 seeks to identify virulence factors that might increase or decrease the invasiveness of these (often) commensal organisms by performing a GWAS against mortality. Finally, in chapter 4, I leverage recent advances in long-read sequencing

technology to investigate the role of plasmids in *E. coli*/*Klebsiella* spp. BSIs, particularly in the dissemination of antimicrobial resistance and virulence genes. I conclude the thesis with a discussion of implications for clinicians and public health practitioners as well as direction for future work.

2. Molecular epidemiology of *E. coli* and *K. pneumoniae* bloodstream infections in Oxfordshire 2008-2018

Note: This chapter has previously been published[68]. I authored the text and figures below (which are reproduced here largely unaltered from the published work) with input from supervisors commensurate with the amount of input/advice that would be considered appropriate for a DPhil thesis. All authors have agreed to the inclusion of this published work in my thesis.

2.1 Introduction

Gram-negative bloodstream infections (GNBSI), predominantly caused by *Escherichia coli* and *Klebsiella* spp., are a significant and increasing threat to public health. As noted in chapter 1, they are now the leading cause of bloodstream infection (BSI) in the UK with a substantial associated burden of morbidity and mortality[62,69]. Despite becoming a significant public health concern and policy focus, their incidence continues to increase and it is unclear how ambitious targets to reduce this can be achieved[70].

There is a significant association between GNBSIs and antimicrobial resistance (AMR), particularly in certain globally successful multi-locus sequence types (STs), such as *E. coli* ST131[4,43,71]. Enterobacteriaceae have relatively open pan-genomes and are able to

rapidly adapt to changing selection pressures (including antibiotic usage)[42,72]. Multidrug resistance (MDR)-associated STs have been linked with prolonged hospital stay and adverse outcomes[73]. Whilst infections caused by relatively susceptible isolates still represent the majority of cases, the potential for rapid proliferation of AMR-associated clones and the dissemination of AMR genes on mobile genetic elements between lineages and species is a major concern[74].

Recent molecular epidemiology studies in the UK have replicated global findings that most *E. coli* BSIs are caused by STs 131, 95, 73 (all phylogroup B2) and ST 69 (phylogroup D)[42,75]. Similar results were observed more recently in a large longitudinal study of isolates in Norway[43]. One large UK based study has shown that after the emergence and expansion of ST69 and ST131 in the early 2000s, the population structure reached an equilibrium, with STs 131, 95, 73 and 69 predominating[41]. However, this study only sequenced isolates cultured prior to 2012, which may represent a critical inflection point in the incidence rate,[62] which continues to increase year on year. It remains unclear whether the population structure remains in equilibrium or whether one or more STs are responsible for this increase.

For *Klebsiella* spp., it has been hypothesised that isolates broadly cause two categories of BSI[65,76], namely: (i) multidrug resistant healthcare-associated (HA) infections caused by low virulence strains with AMR gene-associated plasmids, and (ii) clinically hypervirulent community-associated (CA) infections caused by strains carrying virulence gene plasmids. The convergence of these two phenotypes might pose a significant risk to human health[66]. Whilst there are multiple studies describing in detail the epidemiology of globally distributed clones associated with AMR and hyper-virulence[71,77], such isolates represent a minority of BSI in the UK and most of Europe[78]. A study in Cambridgeshire

(UK, 2006-2012) of 162 *K. pneumoniae* isolates, enriched to over-represent MDR isolates, demonstrated a predominance of globally important STs with apparent cycling of relative incidence over a two-year period[67]. However, MDR infections represent approximately 18% of *K. pneumoniae* BSIs in England and the molecular epidemiology of most invasive disease caused by *Klebsiella* spp. has not been systematically studied[69].

In order to investigate underlying potential pathogen genetic factors driving the increasing incidence of these infections, I analysed all *E. coli* and *Klebsiella* spp. BSIs over a decade (2008-2018) in Oxfordshire, UK, linking electronic health records and mandatory reporting data with sequencing data for all isolates, to give a detailed picture of the regional genomic epidemiology of *E. coli* and *Klebsiella* spp. BSIs. I used other publicly available sequencing datasets to contextualise my findings.

2.2 Methods

2.2.1 Study setting, laboratory procedures and DNA extraction

All isolates causing BSI between 15/09/2008-01/12/2018 (de-duplicated to one BSI isolate per species per 90-day period per patient) were processed by the clinical microbiology laboratory at the John Radcliffe Hospital, Oxford, UK, using standard operating procedures, with sub-cultured stocks stored at -80°C in 10% glycerol nutrient broth. The microbiology laboratory serves all four hospitals and all community healthcare facilities within Oxfordshire, with a catchment population of ~805,000. Blood cultures processed by the laboratory were taken as part of routine clinical workup. We attempted to sequence all isolates with no pre-selection according to e.g. *in vitro* antibiotic susceptibility, patient

factors, date of acquisition etc; I hereby refer to this as an 'unselected' sampling frame. 3461 *E. coli* isolates and 886 Klebsiella spp. isolates were successfully sequenced. All isolates were included in all parts of the relevant analysis except where otherwise indicated below. For sequencing, isolate stocks were grown overnight on Columbia blood agar at 37°C and DNA was extracted from single colonies using the QuickGene DNA extraction kit (Autogen, MA, USA) as per the manufacturer's instructions with the addition of a mechanical lysis step (FastPrep, MP Biomedicals, CA, USA; 6m/s for 40 secs). Short-read sequencing was performed using the Illumina HiSeq 2500/3000/4000/MiSeq instruments as previously described[79]. Additional, publicly available short-read sequencing data (n=1494 sequences) was obtained to facilitate phylogenetic comparison with previous studies as follows:

- 436 *E. coli* BSI isolates from Sydney, Australia (January 2013-March 2016, PRJNA480723[40])
- 415 *E. coli* BSI isolates from Cambridge, UK (2006-2012, PRJEB4681[41])
- 481 *E. coli* BSI isolates from the Netherlands (2014-2016, enriched for ESBL isolates, PRJEB35000[52])
- 162 *E. coli* BSI isolates from Scotland (2013-2015, PRJEB12513[42]).

These isolates were mapped to ST-specific references (for STs 131/95/73/69) and phylogenetic trees generated, followed by permutation tests for geographical clustering as described in the statistical analysis section below.

2.2.2 Epidemiological linkage

Isolate data was linked to laboratory and electronic health record data via the Infections in Oxfordshire Research Database (IORD). Data on suspected infectious focus and patient provenance was acquired via linked local infection control records which had been

submitted to Public Health England as part of the mandatory surveillance programme; such data was available for 400 *Klebsiella* spp. and 2773 *E. coli* isolates (this is because mandatory surveillance for *E. coli* has been in place since 2011 whereas for *Klebsiella* spp. it was only introduced in 2017). Healthcare-associated (HA) BSI were defined as occurring >48 hours post-hospital admission or ≤30 days since hospital discharge; other cases were defined as community-associated (CA)[80].

2.2.3 Genomic analysis

All programs were run using default settings unless indicated. Raw reads were assembled using Shovill (v1.0.4)[81] with assemblies <4Mb and >7Mb excluded from further analysis, because these were thought to represent possible assembly errors/contamination given the typical genome size for these species (4-6.5Mb)[81]. De novo assemblies were annotated with Prokka (v1.14.6)[82]; AMR genes, virulence factors and plasmid replicons were identified using the Resfinder[83], VFDB[84], ISFinder[85] and PlasmidFinder[86] databases with Abricate (v0.9.8) (--min-id 95 --min-cov 95)[87], and Kleborate for *Klebsiella* spp.[65,88–90]. Integrons were located using the IntegronFinder[91] tool and annotated by Mash[92] comparison to reference sequences (minimum containment >= 0.95, accessions in Additional File 1 at https://doi.org/10.1186/s13073-021-00947-2). Multi-locus sequence types (MLST) were determined *in silico* using the MLST tool (v2.17.6) for *E. coli* and Kleborate for *Klebsiella* spp.[93,94], and isolates were assigned to sequence types (ST) based on allelic profiles catalogued in the Achtman (*E. coli*) and Pasteur (*Klebsiella* spp.) schemes[95,96]. These same schemes were used throughout this thesis. Phylogroups were determined using the ClermonTyping tool[97]. To fully utilise the resolution provided by whole genome sequencing I additionally used fastbaps[34] to partition the population below the level of ST designations using the core gene alignment produced by Panaroo[98].

Mapping to MLST-specific reference genomes was performed using Snippy (v4.6.0)[99]; genomes were acquired from NCBI for ST131 - HG941718.1, ST95 - NZ_CP012625, ST73 - AE014075.1. and ST69 - ERR3986909, whereas a custom *de novo* assembly (available at 10.6084/m9.figshare.14498206) was used for *K. pneumoniae* ST490. Gubbins (v2.3.4) was used to build recombination-corrected phylogenies[100]. Time-scaled phylogenies were created using the BactDating[101] library in R (version 3.6.0) under a relaxed gamma model with a minimum of 100,000 Markov chain Monte Carlo iterations[101]. Significance of temporal signal was assessed with root-to-tip plots and 10,000 random permutations of tip dates. Effective sample size was assessed using the Coda package[102]. Evolutionary distinctiveness (ED) was calculated using the Picante package in R[103]; low ED scores indicate closely related genomes. All bioinformatics was performed using the Oxford University Biomedical Research Computing Facility.

Contigs were classified as being of likely chromosomal or plasmid origin using MLplasmids with a 0.7 probability cut-off[104]. All plasmid/chromosomal contigs for a given isolate were binned into separate multi-fasta files and the distances between these calculated using Dashing[105]. The pairwise distance matrix was filtered on a distance of 0.71 (the median plasmidome similarity of $bla_{CTX\text{-}M\text{-}15}$ carrying *K. pneumoniae* ST490 isolates) and then clustered into “plasmidome groups” using the LinkComm package in R[106]. The ecology of categorical groups was compared with a permutational multivariate analysis of variance (permanova) test in the Vegan package in R[107].

A pangenome wide association study (PGWAS) was conducted using PySeer[108] with a linear mixed model utilising relatedness distances inferred from a core genome phylogeny to correct for population structure. A discriminant analysis of principal components

(DAPC) was performed using the Adegenet library in R[109]. Gene co-occurrence was examined using graphs constructed using the iGraph library in R[110]. For each species, edge lists with all genes/plasmids/insertion sequences with a Pearson correlation coefficient spp. >0.5 were created and communities detected using single linkage.

2.2.4 Statistical analysis

To describe molecular epidemiological trends, STs were arbitrarily categorised as “rare” (≤10 isolates over the study period, or untypeable STs), “intermediate” (11-100 isolates), and for *E. coli*, “sub-major “(101-300 isolates) and “major” (>300 isolates). For *Klebsiella* spp. I considered the most prevalent ST (i.e. ST490) in isolation. Stacked negative binomial regression models, assuming a constant population size (i.e. with no offset) with clustered standard errors were used to compare rates of change over time (STATA v16[111])[112]. Incomplete years (2008 and 2018) were excluded from this part of the analysis. All other statistical tests were performed in R v3.6 using the Stats package unless indicated[113]. Charlson score was calculated in the Comorbidity[114] R package using all ICD10 codes associated with episodes in the year prior to specimen collection dates. To test for geographical/healthcare setting structure in MLST groups, I conducted a permutation test similar to that previously described[115]. For this analysis Scotland was considered to be the same country as England (i.e. as part of the UK). In brief, the ratio of median distances between isolates from the same centre and different centres was calculated. This observed ratio was compared to a permuted distribution created by 1000 tip label randomisations. The number of permuted values at least as extreme as the observed value divided by the number of permutations carried out was calculated to give a one-sided test of statistical significance.

2.3 Results

2.3.1 *Escherichia coli* bloodstream infections are stably dominated by major lineages across community and healthcare-associated settings, but with diverse sporadic lineages accounting for almost half of cases, and evidence of sub-lineage replacement in major STs

From September 2008-December 2018, 3461 *E. coli* isolates from 3196 patients were sequenced. Major STs, namely ST73 (n=567), ST131 (455), ST95 (317) and ST69 (312) comprised 48% of all isolates (Table 1); sub-major STs, ST12 (144) and ST127 (113), 7% of all isolates; intermediate STs (32 STs) 24% of all isolates; and rare STs (304 STs) 21% of all isolates (Figure 1/Figure 2). The incidence rate ratio per year (IRRy, i.e. the relative increase in incidence per year) of all these categories increased over time (major IRRy=1.13 (95% CI:1.10-1.17), sub-major IRRy=1.16 (1.10-1.22), intermediate IRRy=1.15 (1.12-1.18), rare IRRy=1.10 (1.06-1.14) (Fig.1A). There was no evidence that the proportion of BSIs caused by major or sub-major STs changed over the study period (Fig.1B). There was evidence that the incidence of BSIs caused by intermediate STs increased slightly faster versus all other isolates and by rare isolates slightly slower ($p_{heterogeneity}$=0.006/<0.001 respectively).

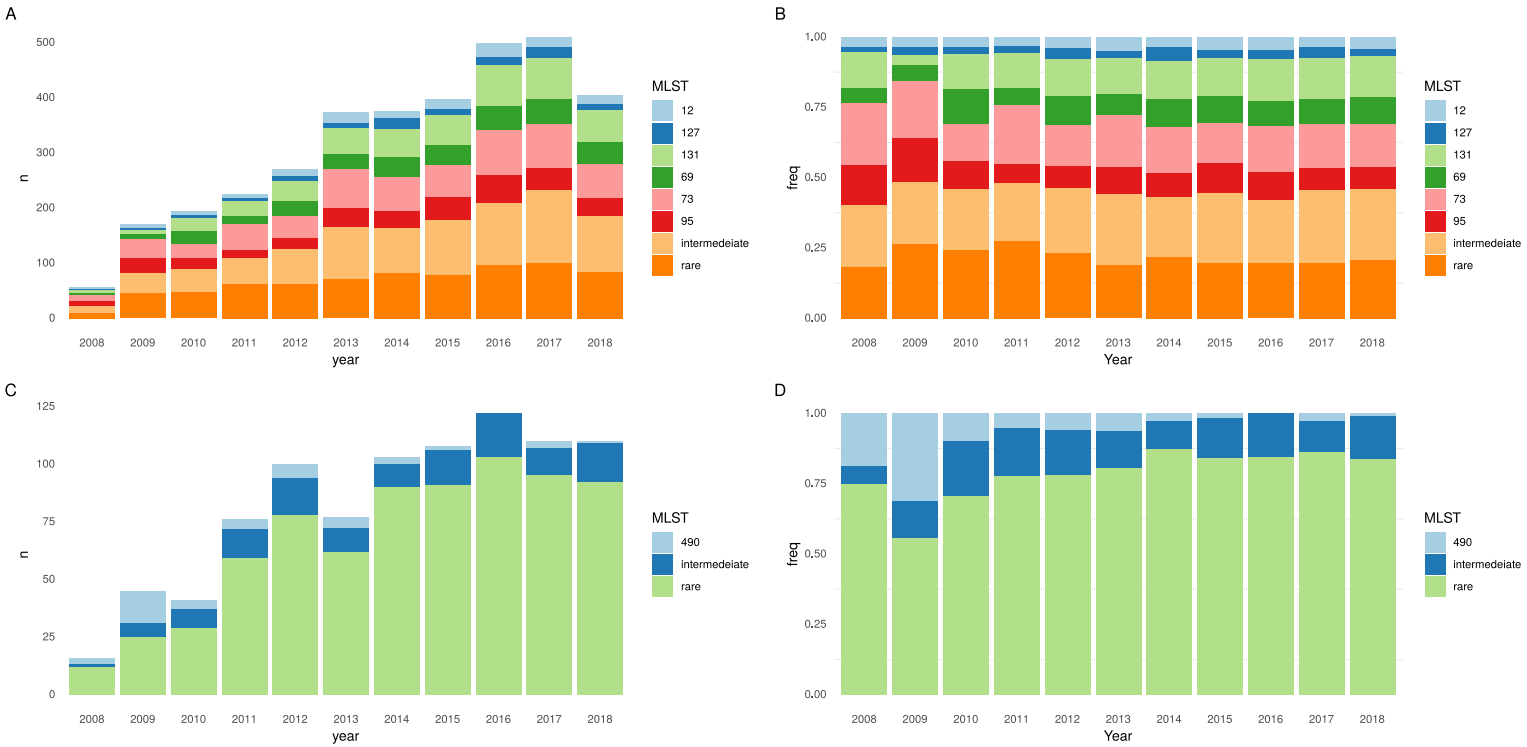


Figure 1 - Population dynamics of *E. coli* (A, B) and Klebsiella spp. (C, D) STs over time.

A) Absolute number and B) proportions of *E. coli* BSIs caused by the four major STs (STs 131/95/73/69), sub-major STs (STs 127/12), intermediate STs and rare STs. C) Absolute number and D) proportions of *Klebsiella* spp. BSIs caused by ST490, intermediate, and rare STs. N.B years 2008 and 2018 are incomplete (see methods).

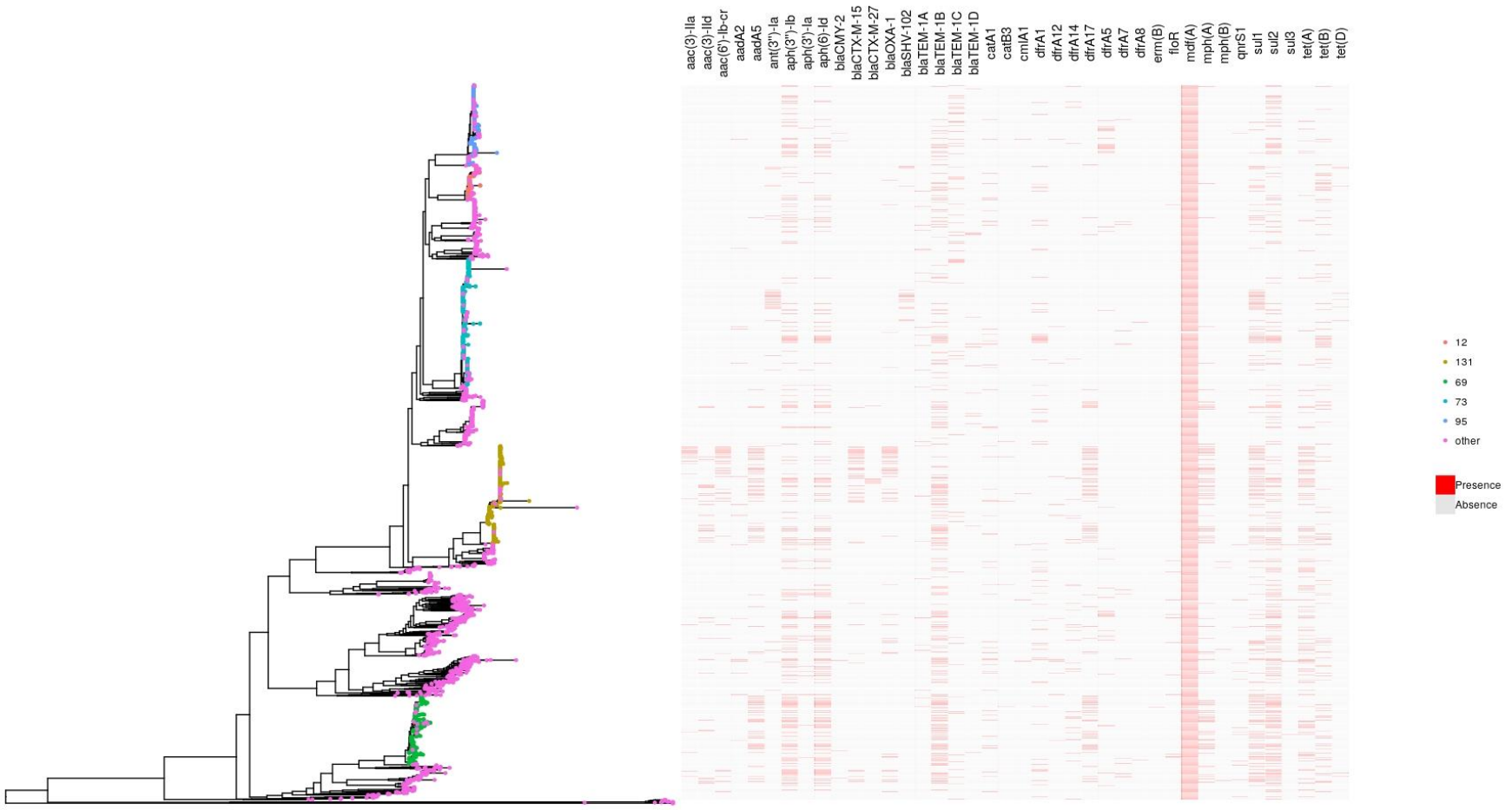


Figure 2 - Gene presence/absence heatmap showing AMR gene presence/absence against the core genome phylogeny for 3461 Oxfordshire *E. coli*.
Tip colours show the location of the more common STs in the phylogeny. Only genes occurring >10 times in the dataset are shown. The tree is midpoint rooted. This figure is also available in higher resolution at https://doi.org/10.6084/m9.figshare.14459139.v2

	E. coli		***Klebsiella spp.***	
Characteristic	Community N = 2,104	Healthcare associated N = 1,357	Community N = 375	Healthcare associated N = 511
MLST Category				
Major	990 (47.1%)	661 (48.7%)	0 (0.0%)	0 (0.0%)
Sub-Major	161 (7.7%)	96 (7.1%)	0 (0.0%)	0 (0.0%)
Intermediate (includes ST490 for *Klebsiella* spp.)	516 (24.5%)	302 (22.3%)	69 (18.4%)	102 (20.0%)
Rare	437 (20.8%)	298 (22.0%)	306 (81.6%)	409 (80.0%)
Source of Infection*				
Bone or Joint	4 (0.3%)	2 (0.2%)	0 (0.0%)	1 (1.1%)
Gastrointestinal	88 (6.3%)	153 (17.0%)	4 (6.0%)	10 (11.4%)
Hepatobiliary	271 (19.4%)	131 (14.5%)	20 (29.9%)	10 (11.4%)

	E. coli		***Klebsiella spp.***	
Characteristic	Community N = 2,104	Healthcare associated N = 1,357	Community N = 375	Healthcare associated N = 511
Other	38 (2.7%)	35 (3.9%)	0 (0.0%)	2 (2.3%)
Respiratory	25 (1.8%)	34 (3.8%)	2 (3.0%)	12 (13.6%)
Skin/Soft Tissue	12 (0.9%)	13 (1.4%)	0 (0%)	0 (0%)
Unclear	234 (16.7%)	186 (20.6%)	13 (19.4%)	31 (35.2%)
Urinary Tract	728 (52.0%)	344 (38.2%)	27 (40.3%)	18 (20.5%)
Vascular Implant	0 (0.0%)	3 (0.3%)	0 (0.0%)	0 (0.0%)
Invasive Line	0 (0%)	0 (0%)	1.0 (1.5%)	4.0 (4.5%)
Unknown	704	456	308	423
Age	76 (63.0, 85.0)	70 (55.0, 80.0)	76 (64.2, 84.8)	65 (49.0, 75.0)
Sex				
Female	1,156 (54.9%)	582 (42.9%)	158 (42.1%)	187 (36.6%)

	E. coli		*Klebsiella spp.*	
Characteristic	Community N = 2,104	Healthcare associated N = 1,357	Community N = 375	Healthcare associated N = 511
Male	948 (45.1%)	775 (57.1%)	217 (57.9%)	323 (63.2%)
Not Specified	0 (0.0%)	0 (0.0%)	0 (0.0%)	1 (0.2%)
Death within 30 days	280 (13.3%)	267 (19.7%)	86 (22.9%)	112 (21.9%)

Table 1 – MLST/clinical characteristics of isolates
"Major" STs > 300 isolates, "sub-major" 101-300, "intermediate" 11-100, "rare" <=10. Number represent proportions and percentages for categorical variables/medians and interquartile ranges for continuous variables. * data available from 2011 for *E. coli* and 2017 for *Klebsiella* spp.

After sub-stratifying the major STs (69, 73, 95 and 131) using fastbaps, there was some evidence of replacement at the sub-lineage level (i.e. genomic groups below the level of STs). Whilst the biological significance of this is not clear, it is interesting that the two sub-lineages of ST69 displaying a relative increase/decrease in incidence had different O-antigen serotypes (fastbaps cluster 16, 38/38 O17, IRRy 0.95 (95% CI 0.87-1.05) and fastbaps cluster 19, 50/67 O15, IRRy 1.31 (95% CI 1.19-1.43). However this was not the case for the other two sub-lineages with some evidence of replacement in ST95 (fastbaps clusters 48 and 49; Figure 3/

Table 2).

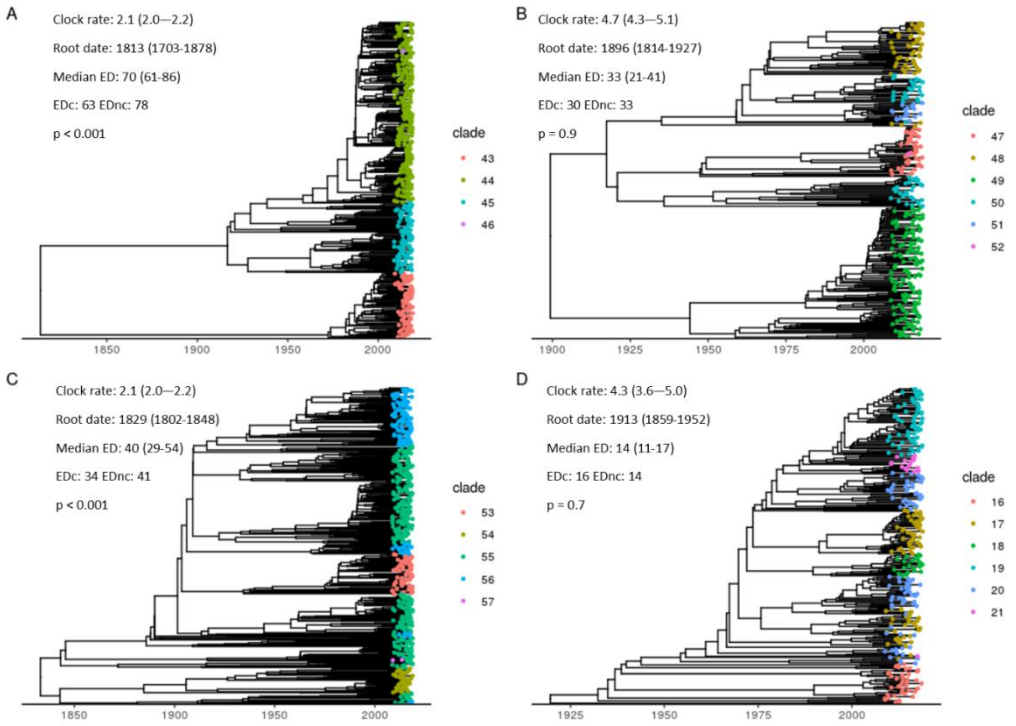


Figure 3 - Time-scaled phylogenies for A) ST131 B) ST95 C) ST73 and D) ST69.
Tip colours represent fastbaps clusters, which were computed from a multiple sequence alignment using the maximum likelihood tree to reduce search space. The text inset shows estimates for the molecular clock (mean, 95% highest posterior density interval), SNPs/site/year), root date (mean, 95% highest posterior density interval), evolutionary distinctiveness (ED, median and interquartile range) and median evolutionary distinctiveness for isolates with (EDc) and without (EDnc) a gene conferring ceftriaxone resistance. The p-value represents a Kruskal-Wallis test of the ED scores between the EDc and EDnc groups.

Table 2 Incidence rate ratios for sub-lineage of major STs identified by fastbaps.

Stacked negative binomial regression models were used to estimate incidence rate ratios per year (IRRy). The p-value represents a Wald test for heterogeneity between the incidence rate for isolates of a sub-lineage defined by fastbaps compared to the incidence rate for all other isolates in a given ST (Bonferroni corrections to account for the number of tests performed per ST are shown to the right of this).

MLST	**Fastbaps cluster**	**IRRy**	**95% CI**	**$p_{heterogeneity}$**	**Bonferroni corrected**
131	43	1.18	1.09-1.27	0.96	1
131	44	1.20	1.10-1.32	0.08	0.25
131	45	1.12	1.01-1.25	0.03	0.1
95	47	1.21	1.00-1.46	0.19	0.93
95	48	1.18	1.11-1.25	0.01	0.05
95	49	1.05	1.00-1.10	0.01	0.03
95	50	1.06	0.96-1.16	0.32	1
95	51	1.00	0.89-1.10	0.07	0.35
73	53	1.17	1.10-1.26	0.13	0.52
73	54	1.02	0.92-1.14	0.13	0.52
73	55	1.11	1.04-1.18	0.47	1
73	56	1.07	0.98-1.16	0.29	1
69	16	0.95	0.87-1.05	<0.001	<0.001
69	17	1.18	1.11-1.24	0.14	0.86
69	18	1.29	1.16-1.44	0.04	0.25
69	19	1.31	1.19-1.43	0.002	0.01

69	20	1.05	0.95-1.17	0.04	0.24
69	21	1.15	0.93-1.42	0.87	1

The majority of *E. coli* BSIs were CA-BSIs (2104/3461, 61%), with no evidence of variation in the proportion caused by major STs between CA-BSI and HA-BSI (990/2104, 47%, vs 661/1357, 49%, p=0.4). Relatively few isolates came from patients resident in a nursing/care home (169 (7%) of 2301 with data available). The four major *E. coli* STs accounted for a significantly greater proportion of these cases (97/169, 57% vs 1009/2132, 47% p=0.01); however, there was no evidence of large-scale nursing/care home-associated BSI outbreaks (Figure 4). *E. coli* CA-BSIs occurred in slightly older (median age 76 (IQR 63-85) vs 70 (55-80), p<0.001) and less comorbid individuals (median Charlson score 1 (IQR 0-2) vs 1 (IQR 1-2), p=0.003) than HA-BSIs.

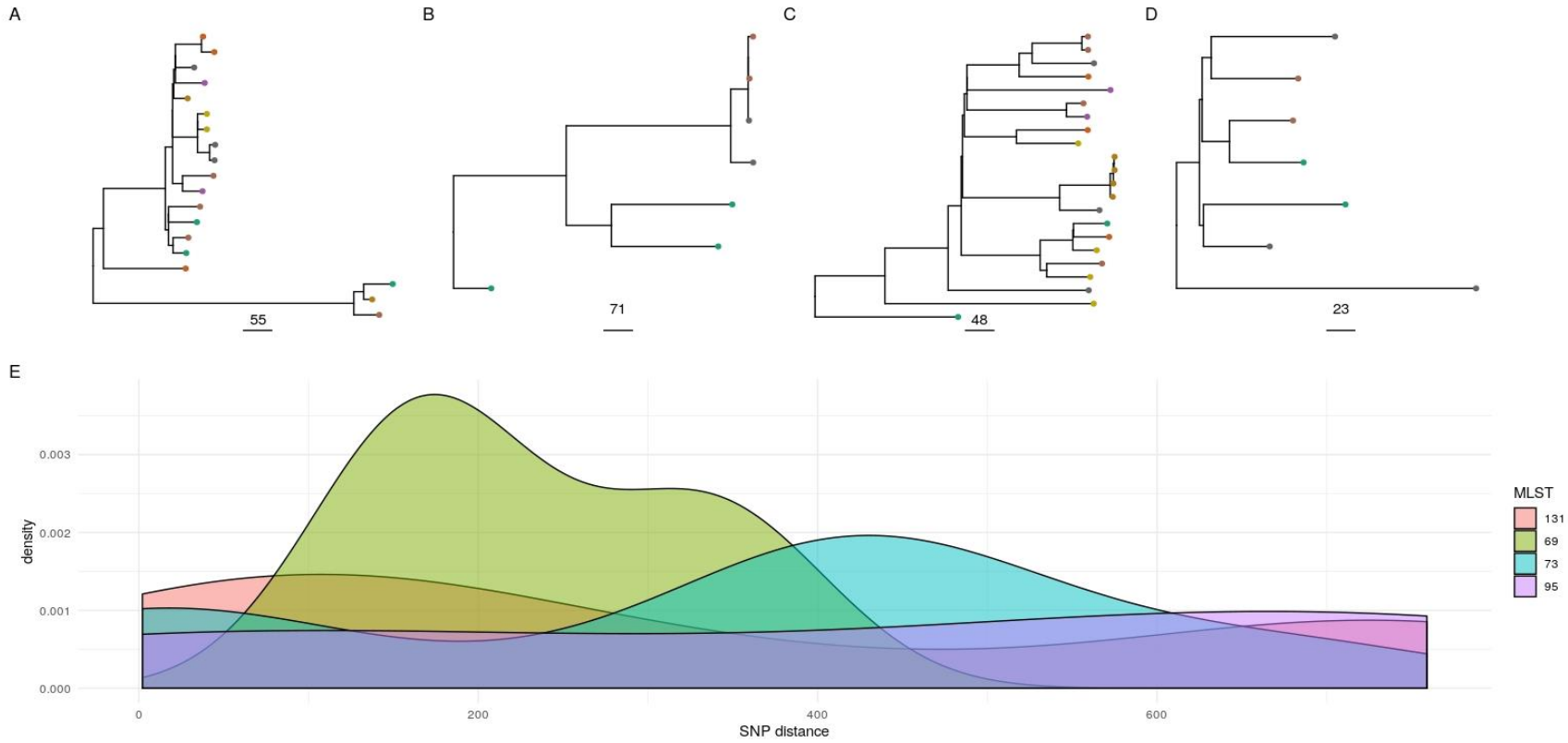


Figure 4 – Minimal evidence of transmission within nursing homes for *E. coli*

A-D - Phylogenies for the major STs containing isolates originating from nursing home residents where there was more than one patient from the same GP practice in the study (the assumption being that they may therefore reside in the same nursing home); panel A - ST131, B - ST95, C - ST73, D - ST69. Tip colours represent GP practices. E - density plot for cophenetic SNP distances for the major ST plots shown in A-D for patients presumed to be in the same nursing home.

Out of 54 total pairwise comparisons of isolates in the same ST and presumed same nursing home, 4 were <= 10 SNPs, 5 11-20 SNPs, 2 21-50 SNPs, 3 51-100 SNPs and 40 >100 SNPs.

2.3.2 Klebsiella spp. bloodstream infections are caused by a diverse representation of sub-species with significant intra-species diversity, with the exception of *K. pneumoniae* ST490, causing a transient, clonal, local outbreak

Amongst the 886 successfully sequenced *Klebsiella* spp. isolates, a large number of sub-species were observed, including *K. pneumoniae* (n=528 [60%]), *K. variicola* (n=112 [13%]), *K. michiganensis* (n=90 [10%]), *K. oxytoca* (n=59 [7%]), *K. aerogenes* (n=38 [4%]), *K. grimontii* (n=30 [3%]), *K. quasipneumoniae* (n=28 [3%]) and *K. africana* (n=1 [0.1%]). In stark contrast to *E. coli*, 738/886 (83%) of *Klebsiella* spp. isolates belonged to rare STs (Figure 1/ Figure 5). The multidrug resistant *K. pneumoniae* ST490 was the most prevalent *Klebsiella* ST and the only one with >40 isolates; its incidence decreased over the study period (IRRy 0.78, 95%CI:0.68-0.89). This ST is rarely seen in other studies (though has been found in a small number of carbapenem resistant isolates in China[116]), consistent with a transient but relatively large, local outbreak. It has previously been shown that prior antibiotic exposure is a risk factor for acquisition of this clone in Oxfordshire[59]. Notably 16/45 (36%) of isolates from this ST were community-onset.

The majority of *Klebsiella* spp. BSIs were HA-BSI (511/886, 58%). As with *E. coli,* few *K. pneumoniae* cases occurred in patients resident in a nursing/care home (10/151, 7%) though these data were not available for most isolates. There was no difference in the proportion of intermediate STs amongst HA-BSI vs CA-BSI (102/511 20% vs 69/375 18%, p=0.6). *Klebsiella* spp. CA-BSI cases were older (median age 76 years (IQR 64-85) vs 65 years (49-75) for HA-BSI; p<0.001) but also less comorbid than HA-BSI cases (median Charlson score 1 (IQR 0-2) vs 2 (1-3); p<0.001). A large number (265/341, missing data for 31) of CA-BSIs occurred in relatively healthy individuals (Charlson score ≤2 i.e. predicted 10 year survival ~90%). In contrast to previous reports suggesting *K.*

quasipneumoniae/*K. variicola* may be less virulent than *K. pneumoniae*[76], in my study there was no evidence of differences in crude 30-day mortality between the *Klebsiella* subspecies with ≥20 isolates (111/528 (21%) *K. pneumoniae*, 29/112 (26%) *K. variicola*, 23/89 (26%) *K. michiganensis*, 10/59 (17%) *K. oxytoca*, 10/38 (26%) *K. aerogenes*, 7/28 (25%) *K. quasipneumoniae,* and 7/30 (23%) *K. grimontii;* exact p=0.5, missing data for 9 isolates), nor the proportion of CA-BSI (230/527 (44%), 48/112 (43%), 34/89 (38%), 27/58 (47%), 11/28 (39%), 9/27 (25%) and 12/30 (40%) respectively; exact p=0.6.

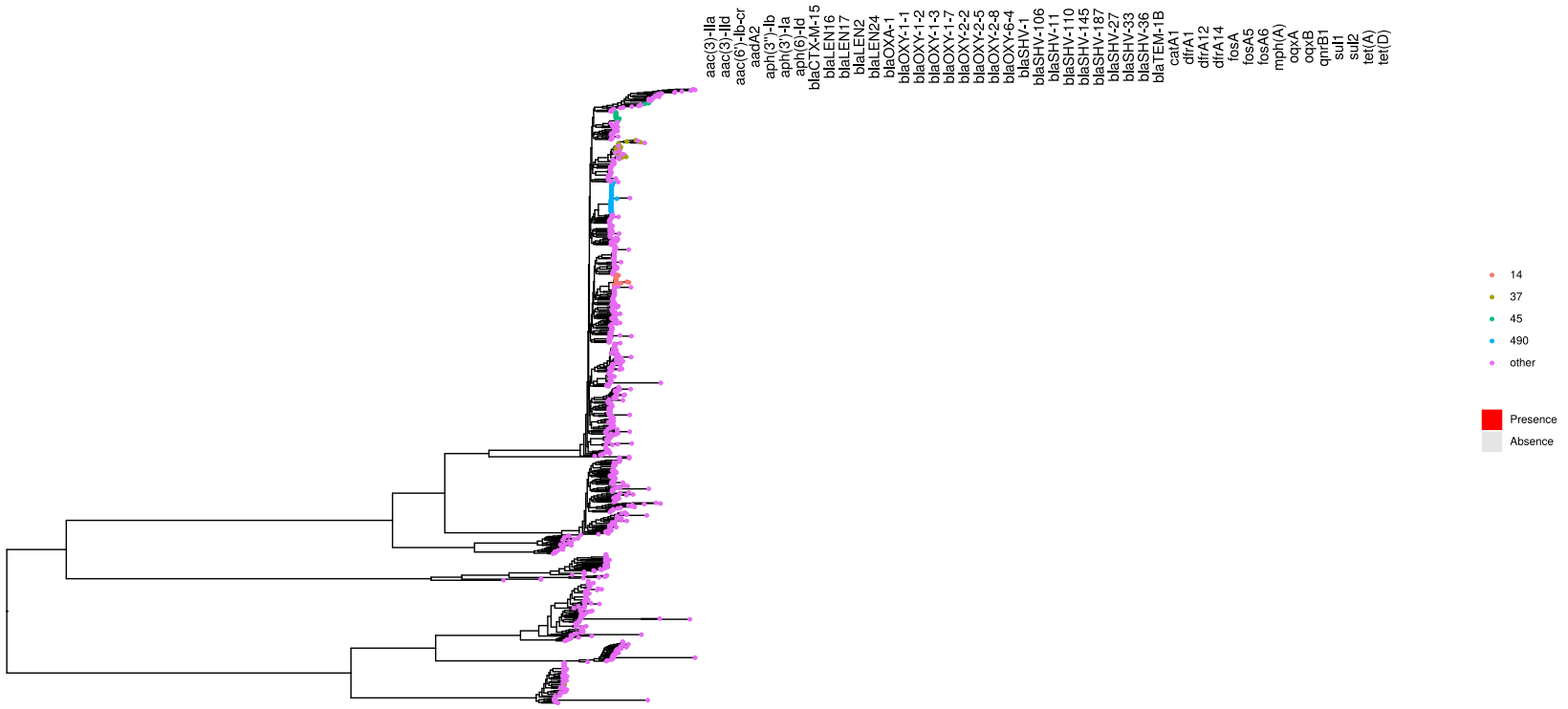


Figure 5 - Gene presence/absence heatmap showing AMR gene presence/absence against the core genome phylogeny for *Klebsiella* spp.
Only genes occurring >10 times in the dataset are shown. Tip colours show the location of the more common STs in the phylogeny. This figure is also available in higher resolution at https://doi.org/10.6084/m9.figshare.14459139.v2

2.3.4 No evidence of intra- or inter-regional clustering of major *Escherichia coli* lineages causing BSI in either community- or healthcare-associated settings

For the four largest *E. coli* STs, I compared the phylogeny of BSI isolates in this study to those in previous studies over a similar timeframe (see Methods). There was no obvious phylogenetic clustering of HA or CA isolates (Figure 6), supported by permutation testing, confirming that for the major STs these were indeed randomly distributed across the phylogeny by geography and healthcare/community setting (Table 3). Additionally, for all four of the major *E. coli* STs, there was no difference in the evolutionary distinctiveness scores (ED, where one might hypothesise there would be low scores amongst isolates involved in clonal nosocomial outbreaks), between HA and CA cases, and therefore no evidence of healthcare-associated adaptation and subsequent transmission (

Table 4). My PGWAS (including 10,095 genes present at minimum 1% population frequency, Bonferroni corrected significance threshold = $2.50x^{10-6}$) did not identify any significant hits separating CA and HA isolates (Figure 7), including those that had been previously identified in other studies (*bglH*, *rhs*, *higB*, *chpB* and *chpS*)[42]. The top five hits in this analysis included four associated with nosocomial onset (three hypothetical proteins (p values $1.86x10^{-5}$, $1.78x10^{-4}$ and 2.24x10-*4*), the flagella export apparatus gene *flip* ($p=2.45x10^{-4}$)) and *papGII* ($p=3.24x10^{-4}$) which was associated with community onset.

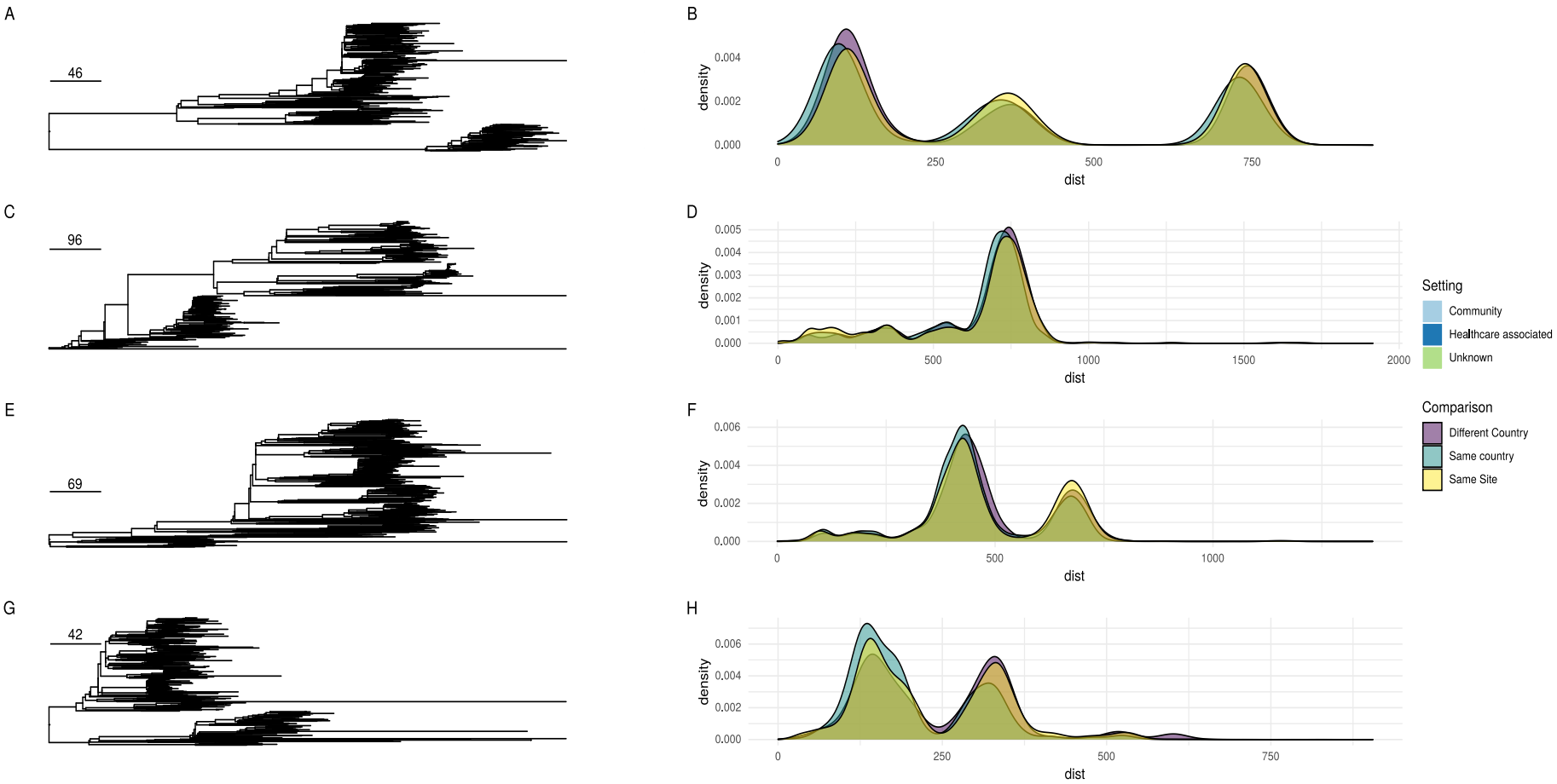


Figure 6 – Genomic relatedness of community/healthcare onset *E. coli* isolates and those from the same/different countries.

Left panel: Core genome phylogenies (corrected for recombination) for *E. coli* ST131/95/73/69 (top to bottom). Origin of isolates is denoted by the colour in the bar to the right of the tree. Scale bar shows SNPs. Right panel shows the distribution of cophenetic distances for the trees shown for isolates from the same site, same country and different countries. The original file for rendering in higher resolution is available at https://doi.org/10.6084/m9.figshare.21076933.v1.

Table 3 - No evidence of genomic structuring of *E. coli* isolates by geography or healthcare exposure

SNP ratios (median within/between region and median HA/all) were calculated for each ST. Permuted distributions were created by randomising tip labels 1000 times and recalculating ratios. P-values represent the number of permuted values at least as extreme as the observed value/1000. Values closer to zero imply greater geographical/nosocomial clustering of isolates, respectively. In all cases the observed values were compatible with the null hypothesis.

	Geographic distribution			**Healthcare setting**		
ST	**Observed ratio**	**Expected ratio**	**p-value**	**Observed ratio**	**Expected ratio**	**p-value**
95	1.00	0.99 - 1.01	0.9	1	0.99 - 1.01	0.2
73	1.00	0..98 - 1.02	0.6	0.99	0.98 - 1.02	0.07
69	1.01	0.90 - 1.10	0.6	1.14	0.92 - 1.12	1
131	1.12	0.93 - 1.07	1	1	0.95 - 1.04	0.5

Table 4 - Evolutionary distinctiveness (ED) scores for community-associated (CA) and healthcare-associated (HA) isolates amongst major *E. coli* STs. P-values represent Wilcoxon rank sum tests between the groups.

ST	Median ED CA (IQR)	Median ED HA (IQR)	p-value
95	40 (31 - 49)	41 (33 - 47)	0.7
73	49 (36 - 63)	45 (35 - 57)	0.09
69	18 (15 - 22)	17 (14 - 21)	0.06
131	75 (68 - 91)	76 (68 - 91)	0.8

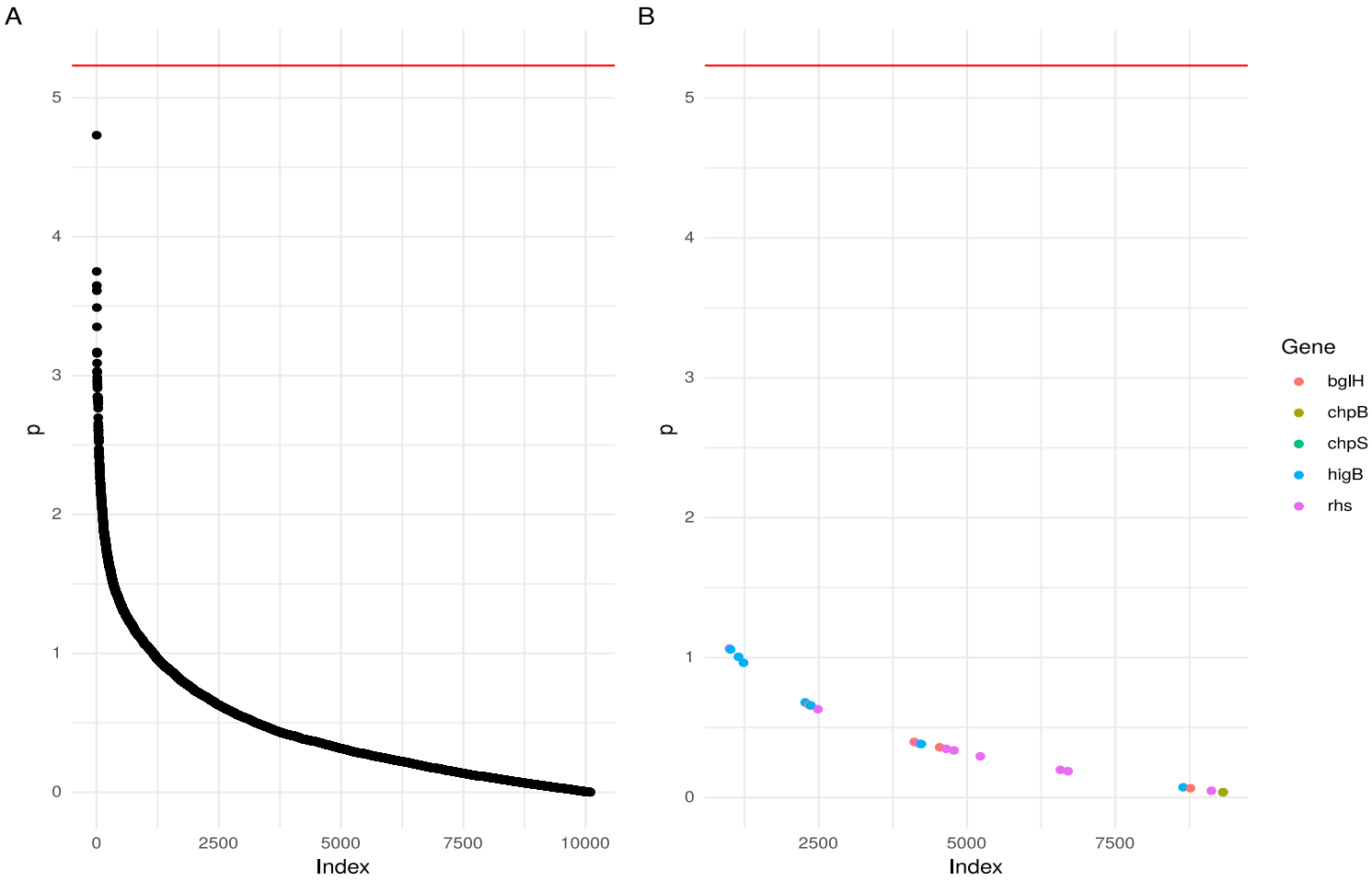


Figure 7 - Manhattan plots of a pangenome wide association study of the association of genes with community/healthcare associated onset in *E. coli*.
A) all genes shown in descending order of -log10(p value). B) genes previously shown to be significantly associated with either healthcare/community onset cases[42]. Red horizontal bar represents the Bonferroni adjusted significance threshold.

2.3.5 Urinary and hepatobiliary sources of *E. coli* and *Klebsiella* spp. BSI predominate, with major *E. coli* STs over-represented in urinary-associated *E. coli* BSIs

The urinary tract was the most common physician-identified source of infection in both species. In *E. coli*, these UTIs were strongly associated with the presence of the *pap* group of genes (Table 5) which were over-represented in major STs (916/1651 (55%) vs 394/1810 (22%) isolates, p<0.001). The hepatobiliary tract was the second most common source for both *E. coli* and *Klebsiella* spp., followed by other gastrointestinal infections for *E. coli* and the respiratory tract for *Klebsiella* spp. (Figure 8a/b). For *E. coli* BSIs, there was some evidence that the incidence of BSIs not attributed to the urinary tract increased faster than those thought to be of urinary origin (IRRy 1.45, 95%CI 0.98-2.17 vs IRRy 1.36, 95%CI 0.88-2.08 respectively; $p_{heterogeneity}$=0.005, Figure 8c). The proportion of BSIs caused by the five major *E. coli* STs was greater for BSIs with urinary compared to other sources (586/1071 [55%], vs 520/1229 [42%], p<0.001); this was not the case for *Klebsiella* spp. where the STs causing BSI attributable to all sources were diverse.

Table 5 – Top genes associated with a urinary source of *E. coli* bloodstream infection

A pangenome-wide association study (PGWAS) was performed using Pyseer[108] to test the assocation of gene presence/absence with pysician identified bloodstream infection source. Shown are the first 8 hits in order of likelihood ratio test p value. The Bonferroni corrected significance threshold was 1.10×10^{-6}. For each test the comparison is isolates of a Urinary source vs. all other isolates.

Gene	**Allele frequency**	**Likelihood ratio test p-value (adjusted for population structure)**	**Beta**	**Standard Error**	**Gene H^2**
papG	0.39	3.43×10^{-29}	0.28	0.02	0.23
papC	0.50	3.61×10^{-20}	0.23	0.02	0.19
papH	0.50	4.97×10^{-20}	0.23	0.02	0.19
papD	0.50	6.19×10^{-20}	0.23	0.02	0.19
papK	0.43	3.83×10^{-19}	0.21	0.02	0.19
papF	0.47	3.11×10^{-16}	0.20	0.02	0.17
papJ	0.40	2.91×10^{-11}	0.17	0.03	0.14
papE	0.32	1.86×10^{-9}	0.15	0.03	0.13

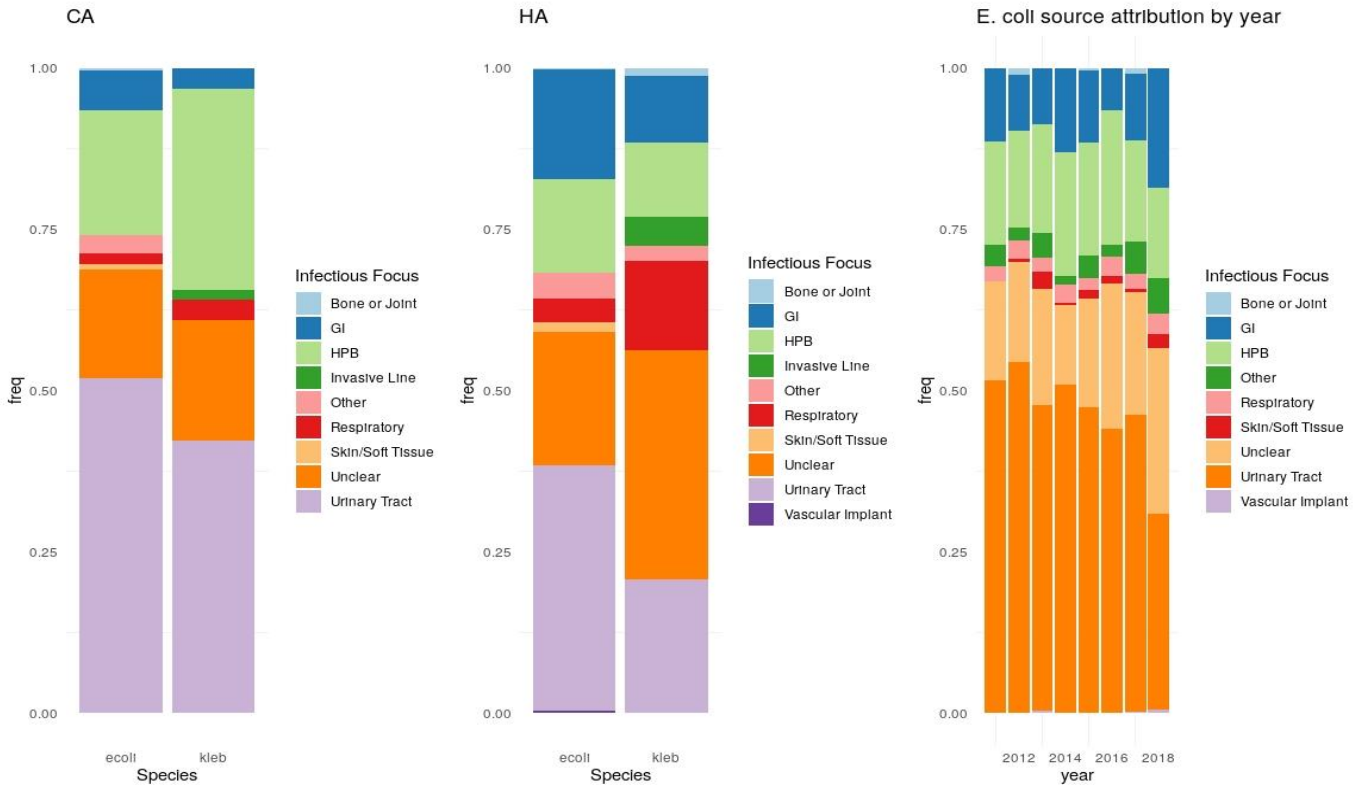


Figure 8 - Proportions of presumed infectious foci for CA (left panel) and HA (central panel) *E. coli* BSI. Source attribution by year for *E. coli* (right panel).

2.3.6 Multi-drug resistant *E. coli* BSIs are increasing, likely driven by the association of AMR genes within lineages, and ceftriaxone resistance is driven by the expansion of resistant ST131 and ST73 sub-lineages

For E. coli BSIs, considering total effects of resistance to individual antibiotics, regardless of mechanism, there was evidence of increasing incidence of BSI carrying AMR genes conferring resistance to amoxicillin IRRy (1.17 (95%CI 1.14-1.20)), co-amoxiclav (IRRy 1.15 (1.08-1.24)), ceftriaxone (IRRy 1.26 (1.17-1.36)), gentamicin (IRRy 1.24 (1.15-1.33)), trimethoprim (IRRy 1.17 (1.14-1.20)) and ciprofloxacin (IRRy 1.23 (1.17-1.29) Fig.3); there was only a single isolate carrying a gene encoding for carbapenem resistance (ST10, bla_{OXA-48}) in 2014. For ceftriaxone ($p_{heterogeneity}$=0.003), gentamicin ($p_{heterogeneity}$=0.03), amoxicillin ($p_{heterogeneity}$=0.003), and ciprofloxacin ($p_{heterogeneity}$<0.001), there was evidence that the incidence of BSI from isolates with AMR genes encoding resistance to these antimicrobials was increasing faster than those without. Genes conferring resistance to ceftriaxone in *E. coli* were over-represented in ST131 (180/346) and ST73 (56/346), with significantly lower ED scores for isolates carrying ceftriaxone-resistance genes in these STs (accounting for 236/346 (68%) of these genes, Figure 2), consistent with genomes from resistant sub-lineages being more closely related within these sub-lineages, and therefore with their expansion. Findings were similar restricting to CA-BSI. ST127 was the only major/sub-major ST in which no genes conferring resistance to ceftriaxone were detected.

2.3.7 Drug-susceptible *Klebsiella* spp. BSIs are increasing faster than drug-resistant strains, with ceftriaxone-resistance largely healthcare-associated

For *Klebsiella* spp., there was only evidence of increasing incidence of BSI from isolates carrying genes/mutations conferring resistance to co-amoxiclav (IRRy 1.07 (95%CI:1.02-1.12)) and ciprofloxacin (IRRy 1.11 (1.06-1.16)) (Figure 9). In contrast, the incidence of genes conferring resistance to ceftriaxone (IRRy 0.99 (0.92-1.09)) and trimethoprim (IRRy 1.03 (0.93-1.15) were stable while gentamicin (IRRy=0.86 (0.76-0.97)) decreased. For ceftriaxone, co-amoxiclav and gentamicin, the incidence trends of isolates not carrying genes conferring resistance to these antibiotics increased faster than those carrying resistance ($p_{heterogeneity} \leq 0.01$).

A major contributing factor to this finding was the overall decline of the MDR ST490 lineage and the emergence of a more susceptible ST490 sub-lineage between 2005-2010 which had lost the *aac(3)-IIa*, *aac(6')-Ib-cr*, bla_{OXA-1} and *tet(a)* genes (Figure 10). Genes encoding for resistance to ceftriaxone were significantly more common in intermediate STs (STs with >10 isolates in the dataset) (54/148, 36% vs 73/738, 10%, p <0.001) and in HA-BSIs (85/510, 17% vs 42/372, 11% p=0.03). Similarly, MDR isolates (resistant to >=3 antibiotic classes) were more common in intermediate vs rare STs (89/366, 24% vs 80/520, 15%, p=0.001) and HA-BSI vs CA-BSI (109/510, 21% vs 59/372, 16%, p=0.049).

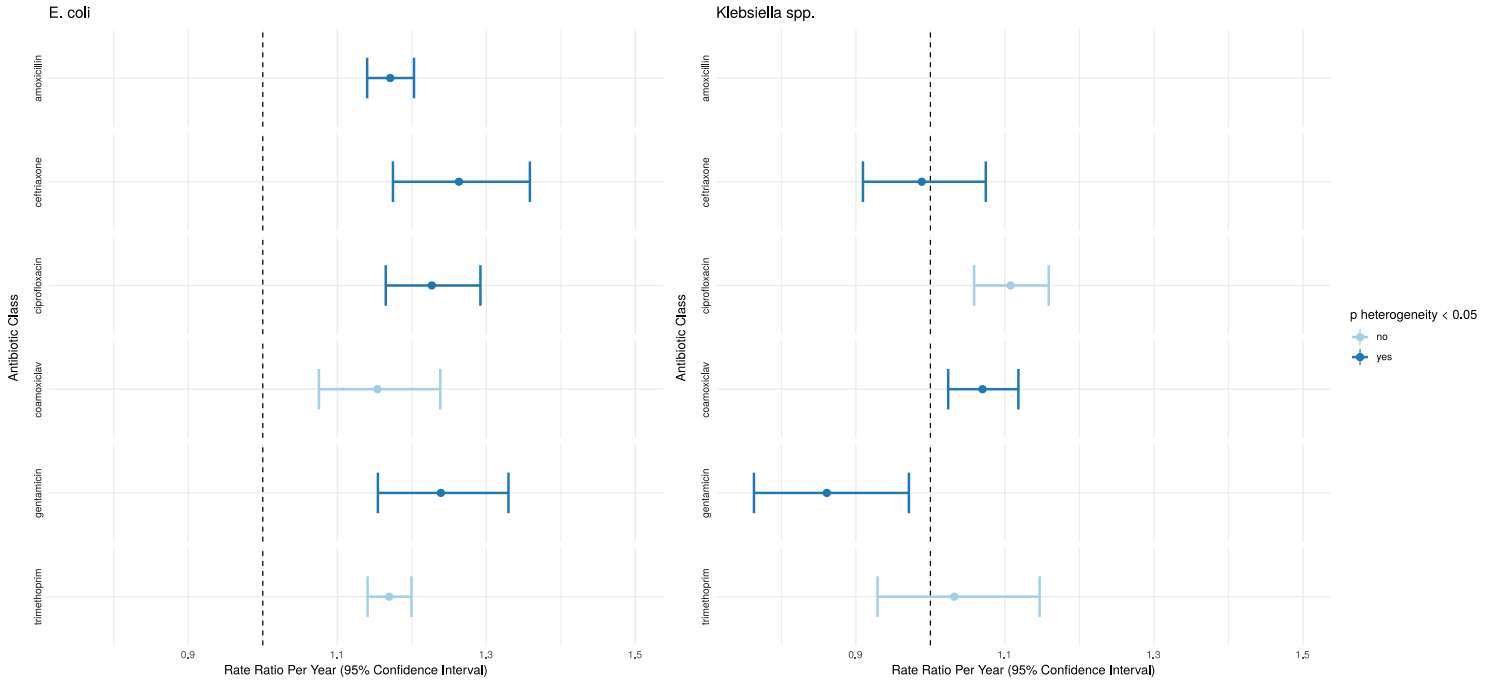


Figure 9 - Changes in the incidence of clinically relevant groups of antimicrobial resistance genes considering total effects of resistance to individual antibiotics, regardless of mechanism, over the study period.

Gene groupings are as defined by ResFinder. Amoxicillin is excluded for *Klebsiella* spp. which are generally considered intrinsically resistant to this drug. Bars show univariate point estimates and 95% confidence intervals for the incidence rate ratio (per year) of presence of genes belonging to the classes shown. The point estimate and error bar is shown in a darker blue if a Wald test between the incidence rate ratios for the presence and absence of a given set of genes was significant at a 0.05 threshold (i.e. there was evidence that the rate of increase of isolates with a group of genes was different to that for isolates without.)

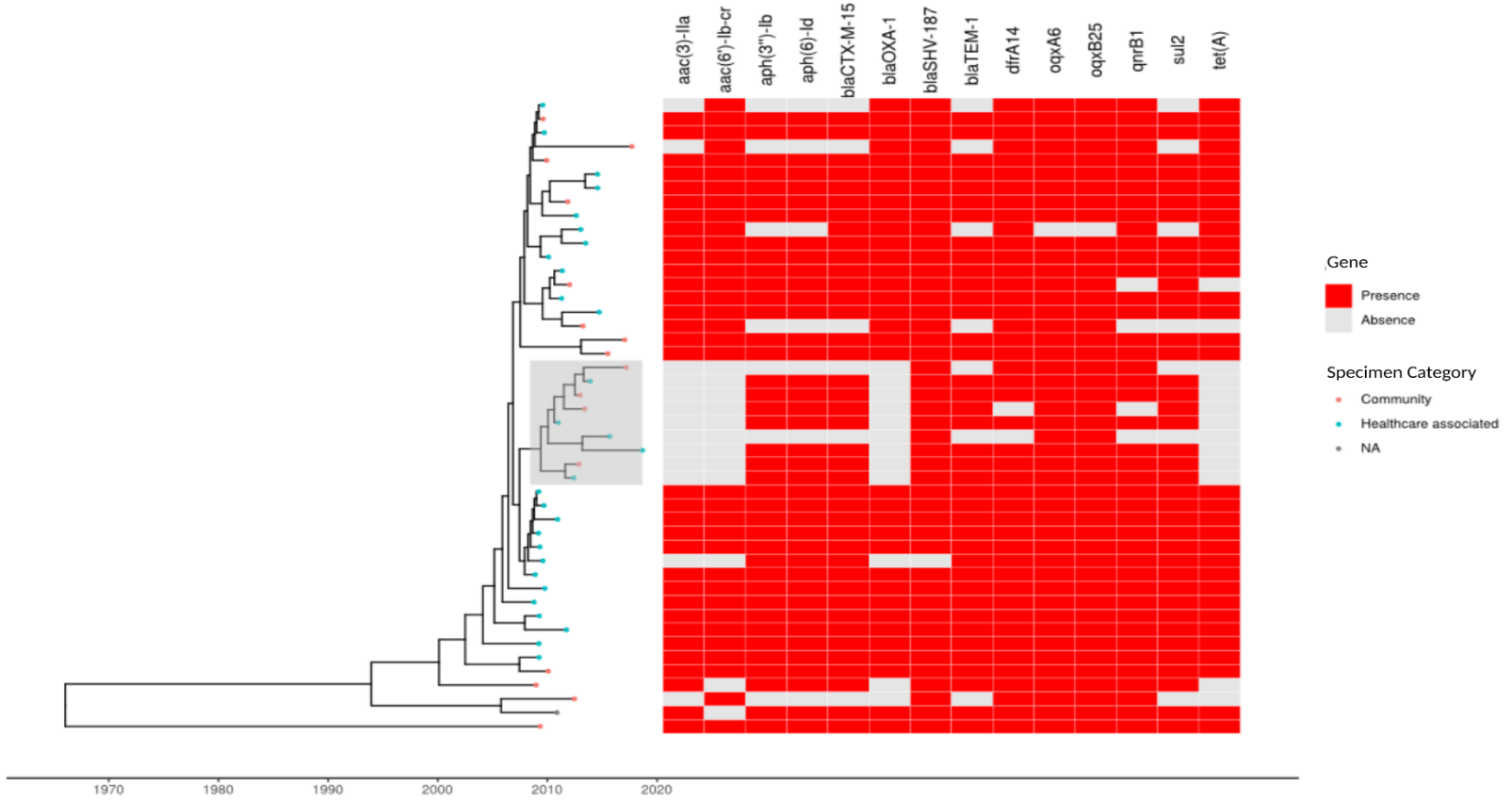


Figure 10 - Timescaled phylogeny of *Klebsiella pneumoniae* ST490 with a heatmap of AMR genes.

Tree tip colours indicate whether the isolate is community/healthcare associated. Only genes occurring >10 times are shown. The subclade which emerged without *aac(3)-IIa*, *aac(6')-Ib-cr*, bla_{OXA-1} and *tet(A)* is shown with a grey highlighted box. The x axis shows time in years.

2.3.8 Virulence factors are strongly structured by ST amongst *E. coli* but not *Klebisella* spp. BSI isolates, with no evidence in *Klebsiella* spp. of a difference in virulence gene carriage between community vs hospital acquired BSIs

The distribution of virulence factors strongly reflected the underlying clonal population structure in *E. coli* isolates, with segregation by ST, confirmed with a discriminant analysis of principal components (DAPC) (Figure 11). Most notably phylogroup B2 isolates were separated from the rest of the population by the presence of *chuA*, *chuX* (involved in haem utilisation) and *espL1/espX4* (elements of the type 3 secretion system). Given this and the broad equilibrium in the *E. coli* BSI population structure demonstrated above (Figure 1), it is unlikely that increased carriage of certain virulence factors explains the increasing incidence of these infections. In keeping with this, there was a near-perfect correlation between frequency of carriage of any given virulence gene across all isolates in 2009 vs 2018 (Pearson correlation=0.99; $p<0.001$) suggesting that the proliferation of one or more virulence factors over time is an unlikely explanation for the increasing incidence of BSIs observed (although I cannot exclude this possibility as the BSI isolates in this dataset are virulent by definition and there is no control set of commensal isolates).

No such structuring by virulence factors was observed for *Klebsiella* spp.; instead, virulence factors were widely distributed amongst 79 known STs and 14 isolates with no assigned ST (Figure 12). Notably, of the 372 CA-BSIs, only 65 (17%) carried known virulence factors. There was no difference in the proportion of CA-BSI and HA-BSI isolates carrying ≥1 virulence factor (65/372 [17%] vs 92/510 [18%] respectively; p=0.9) nor carrying colibactin and/or aerobactin (thought to be associated with hypervirulence, 15/372 [4%] vs 25/510 [5%]; p=0.7) (Figure 13).

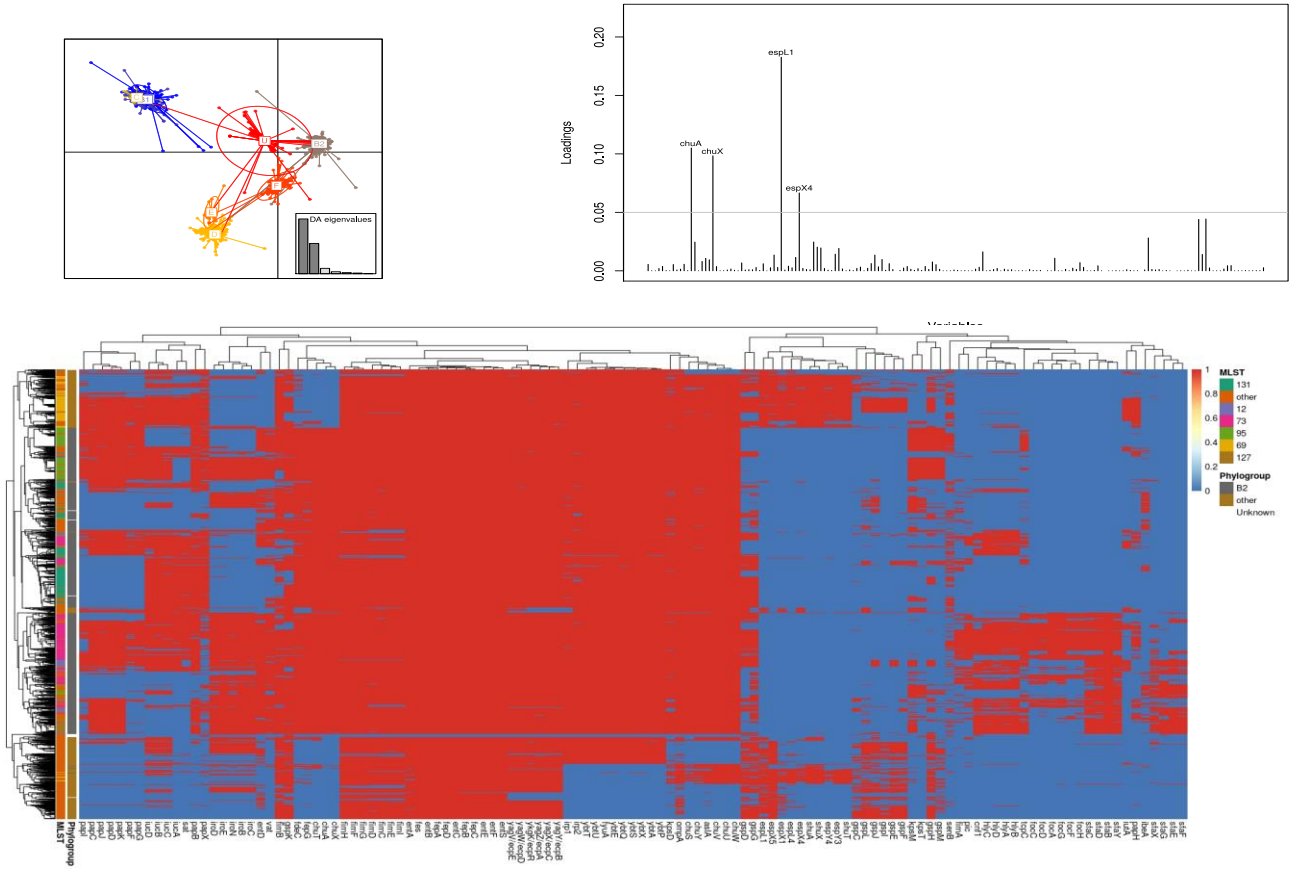


Figure 11 - Virulence elements are structured by ST/Phylogroup in *E. coli*.

Top left - Discriminant analysis of principal components plot showing clear separation of phylogroups by their virulence factor content. Top right - Loading plot showing genes contributing most to the discrimination of phylogroup B2 from other phylogroups. Bottom - Heatmap of presence absence of virulence genes (x-axis) by major ST and phylogroups (y-axis).

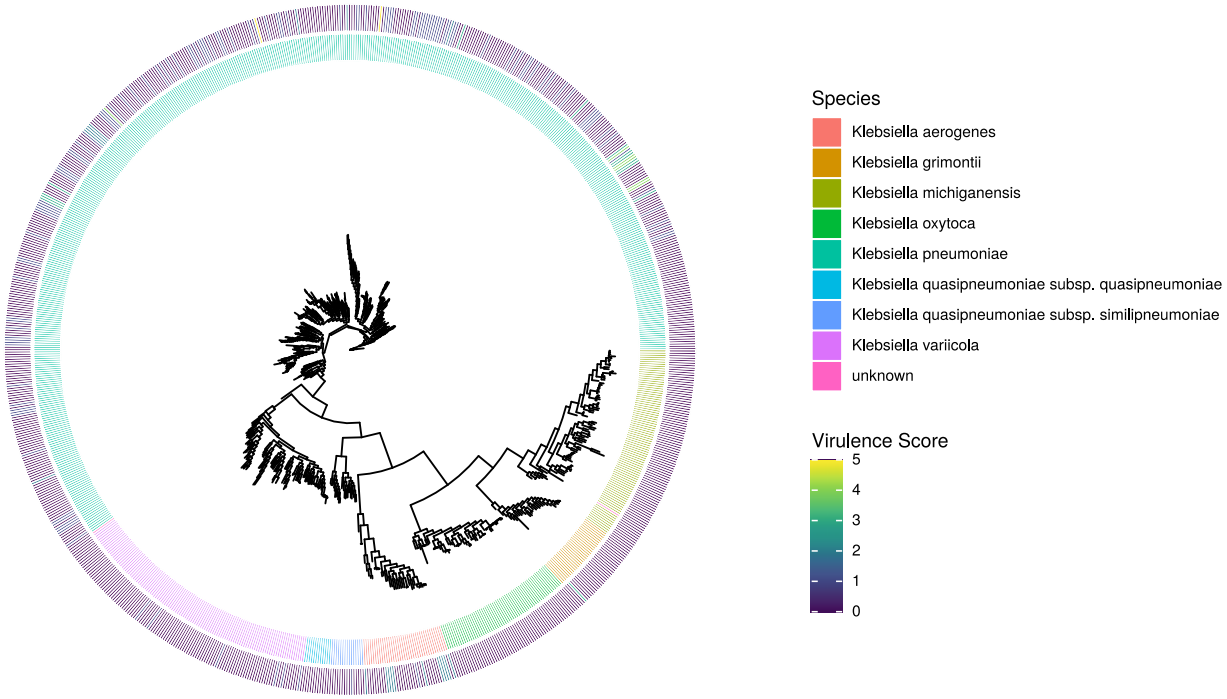


Figure 12 - Phylogenetic tree of *Klebsiella* spp annotated with species (inner ring) and virulence score (outer ring).
Virulence score is defined by Kleborate[94] as follows: 0 – no acquired virulence loci, 1 – yersiniabactin only, 2 – yersiniabactin and colibactin (or colibactin only), 3 – aerobactin only, 4 – aerobactin and yersiniabactin (no colibactin), 5 – yersiniabactin, colibactin and aerobactin. This figure is also available in higher resolution at https://doi.org/10.6084/m9.figshare.14459139.v2

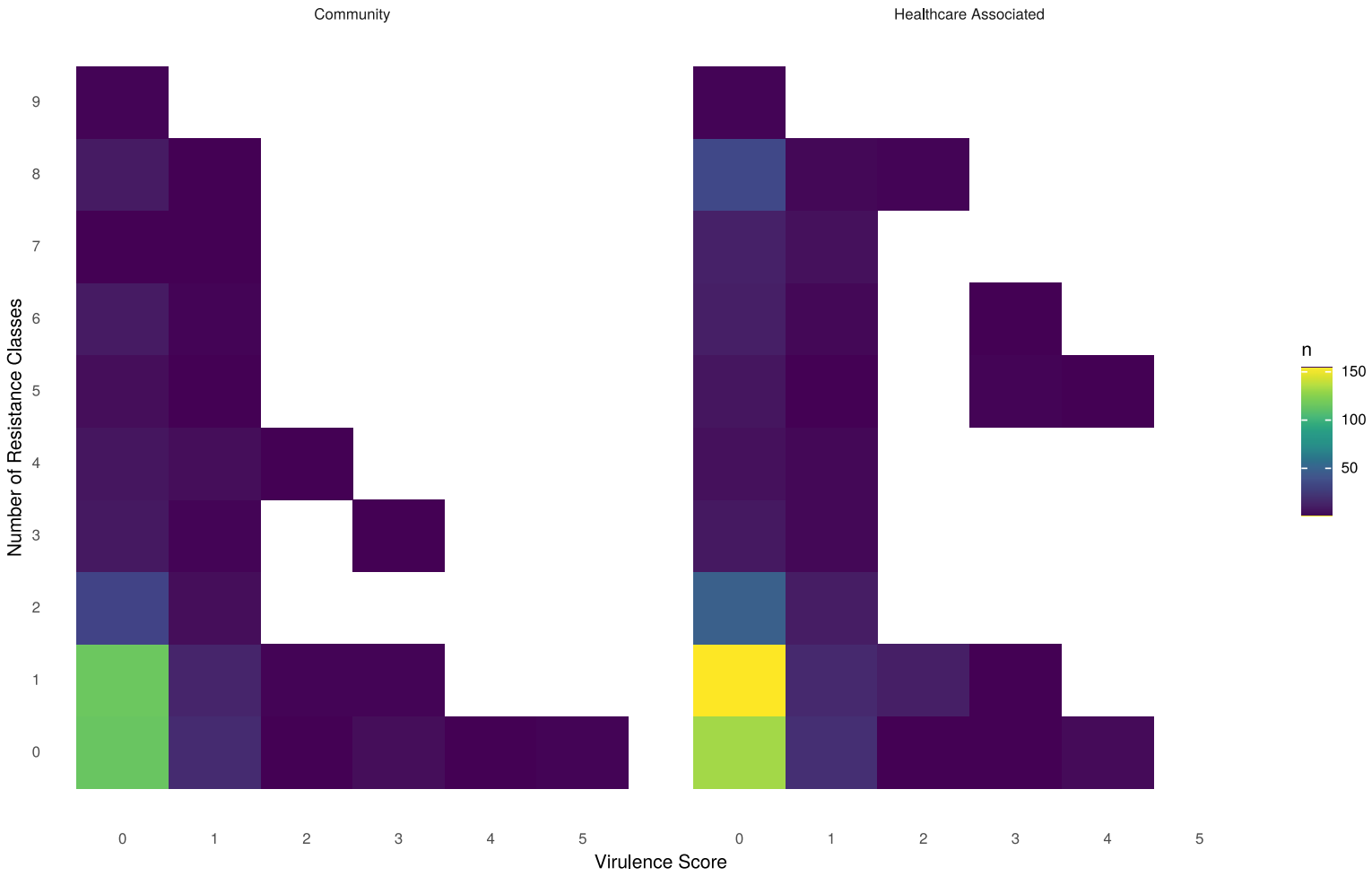


Figure 13 - Distribution of virulence factors and antimicrobial resistance classes across community acquired (left) and healthcare associated (right) *Klebsiella* spp. isolates. Virulence factor scores were assigned by Kleborate as follows: Virulence Score 0 = No acquired loci, 1= yersiniabactin, 2= yersiniabactin and colibactin (or colibactin only), 3 = aerobactin (without yersiniabactin or colibactin), 4 = aerobactin with yersiniabactin (without colibactin) and 5 = yersiniabactin, colibactin and aerobactin.

2.3.9 Plasmid replicon and plasmidome analyses reflect diversity consistent with a highly mobile accessory genome within species

Given the known importance of plasmids in the carriage and transmission of AMR genes, I sought to understand the role they might play in the relative success of major STs. Analysis of plasmid replicon profiles showed these were largely genus-restricted, with some overlap (Figure 14). However, plasmid replicons were not structured by host lineage in *E. coli* or *Klebsiella* spp., with the exception of *K. pneumoniae* ST490 which was discriminated by the presence of the IncFIA(HI1) plasmid type (also identified in 23 *E. coli* isolates; Figure 15).

Using my approach to predict plasmidome population structure based on k-mer similarity of contigs binned as plasmid, I observed a striking degree of plasmidome diversity within highly genomically-related isolates/STs for all species; for 656 isolates with a core genome similarity >0.99 to another isolate in the dataset, the median plasmidome similarity was only 0.51 (IQR 0.28-0.86). Isolates with a near-identical plasmidome (>0.99 similarity, n=115 isolates) did however have highly similar chromosomes (median plasmidome similarity: 0.97 [IQR: 0.94-0.99]). Compared with other STs, major *E. coli* STs had larger plasmidomes (median size 106,766 vs 97,432, $p<0.001$) which belonged to more “plasmidome groups” (see methods) (median 3 vs 2, $p<0.001$). For both *E. coli* and *Klebsiella* spp., there was no evidence of different plasmid populations between those associated with HA- vs CA-BSI ($p=0.4$).

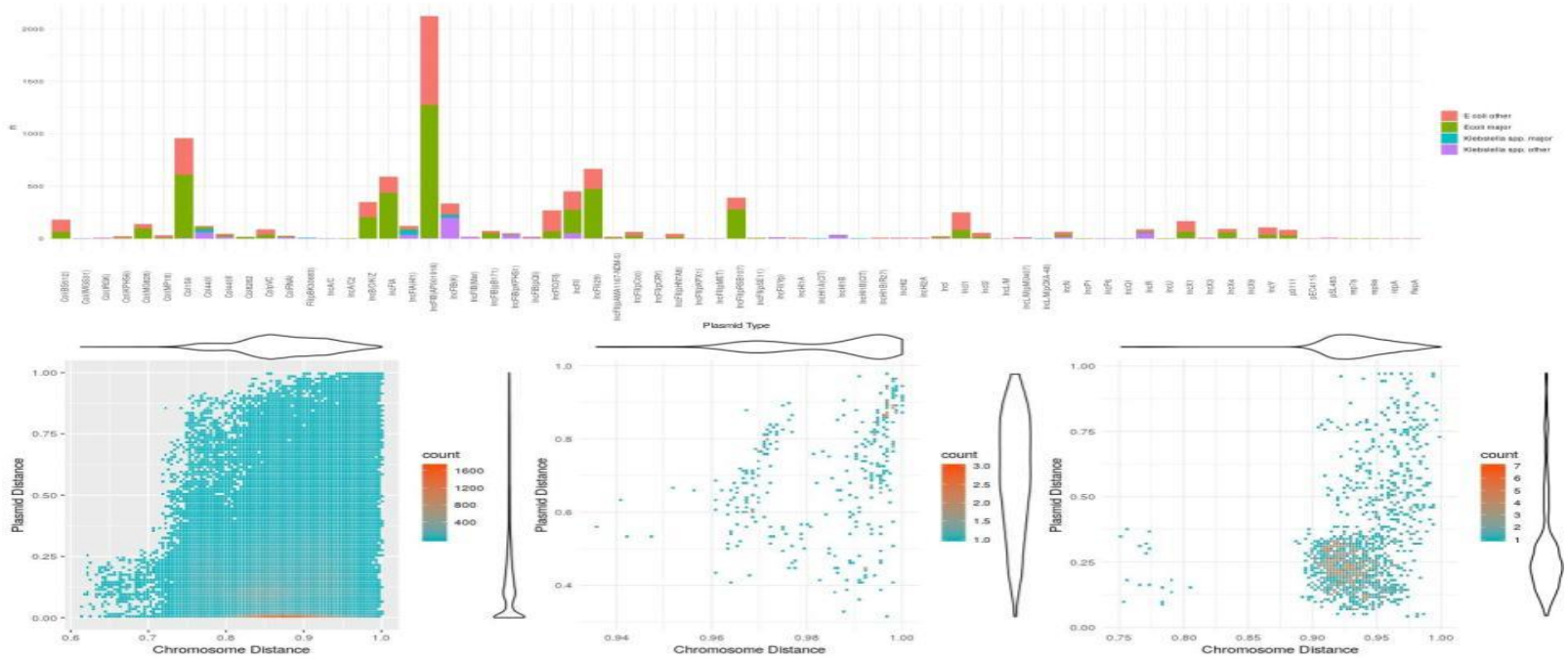


Figure 14 Plasmid replicon types structured by species but with substantial apparent intra-species diversity

Top panel - plasmid types in the PlasmidFinder database identified in the major/other *E. coli*/*Klebsiella* spp., bottom left - plot showing kmer based plasmidome similarity (y-axis) against chromosome similarity (x-axis) for isolates of the same MLST. Violin plots show the distribution of values for each axis.

Origin of contigs was predicted as being plasmid/chromosome associated by MLPlasmids and distances calculated by Dashing. The bottom middle and right plots show the same characteristics for isolates of *K. pneumoniae* ST490 (middle) and *E. coli* ST131 (right) carrying a *bla*$_{\text{CTX-M-15}}$ gene. This figure is also available in higher resolution at https://doi.org/10.6084/m9.figshare.14459139.v2

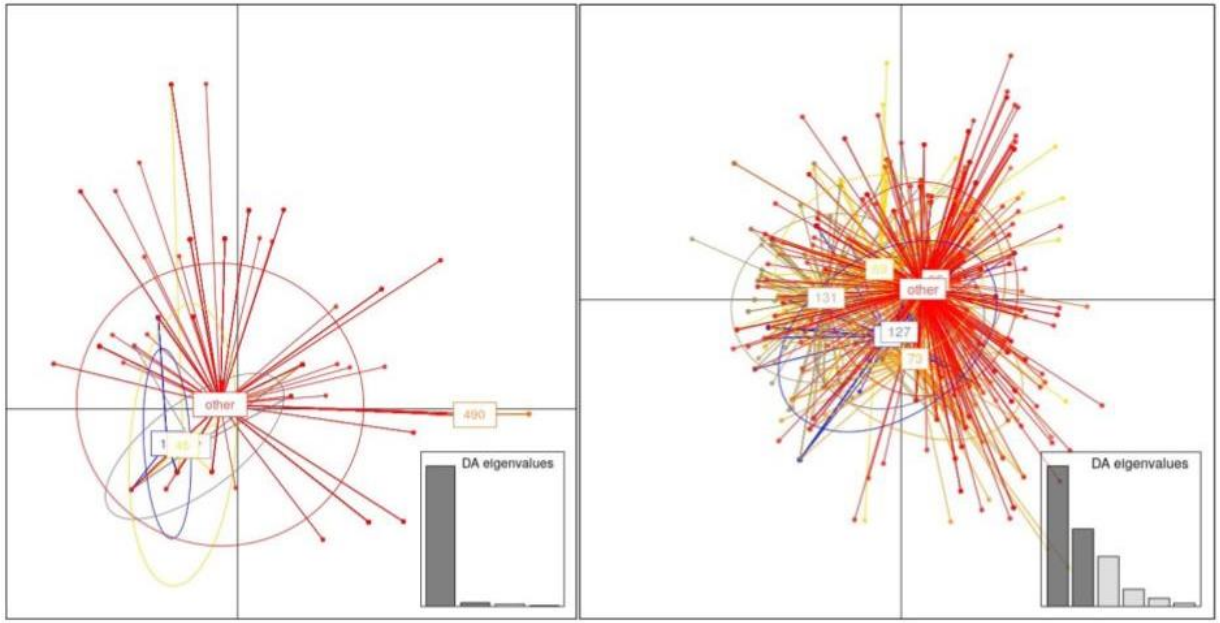


Figure 15 – No evidence of intra-species structure by replicon type

DAPC plots for *Klebsiella* spp. (left) and *E. coli* (right) showing discriminant factors which best discriminate isolates into STs based on plasmid types in the PlasmidFinder database.

Finally, I analysed the co-occurrence of AMR genes, plasmid replicon types, insertion sequences (ISs) and integron-associated markers within *E. coli* and *Klebsiella* spp. The networks formed were notable for the widespread co-occurrence of AMR genes, insertion sequences and (for some classes of AMR genes) integrons, but not usually specific plasmid types (Figure 16/Figure 17). This suggests that most common AMR genes are found in a diverse range of genetic contexts (e.g. multiple plasmid types, chromosomally integrated), and that horizontal gene transfer of important AMR genes in these isolates is largely facilitated by smaller, non-plasmid mobile genetic elements (MGEs) such as transposons/ISs/integrons, which are difficult to reliably evaluate with short-read data[117].

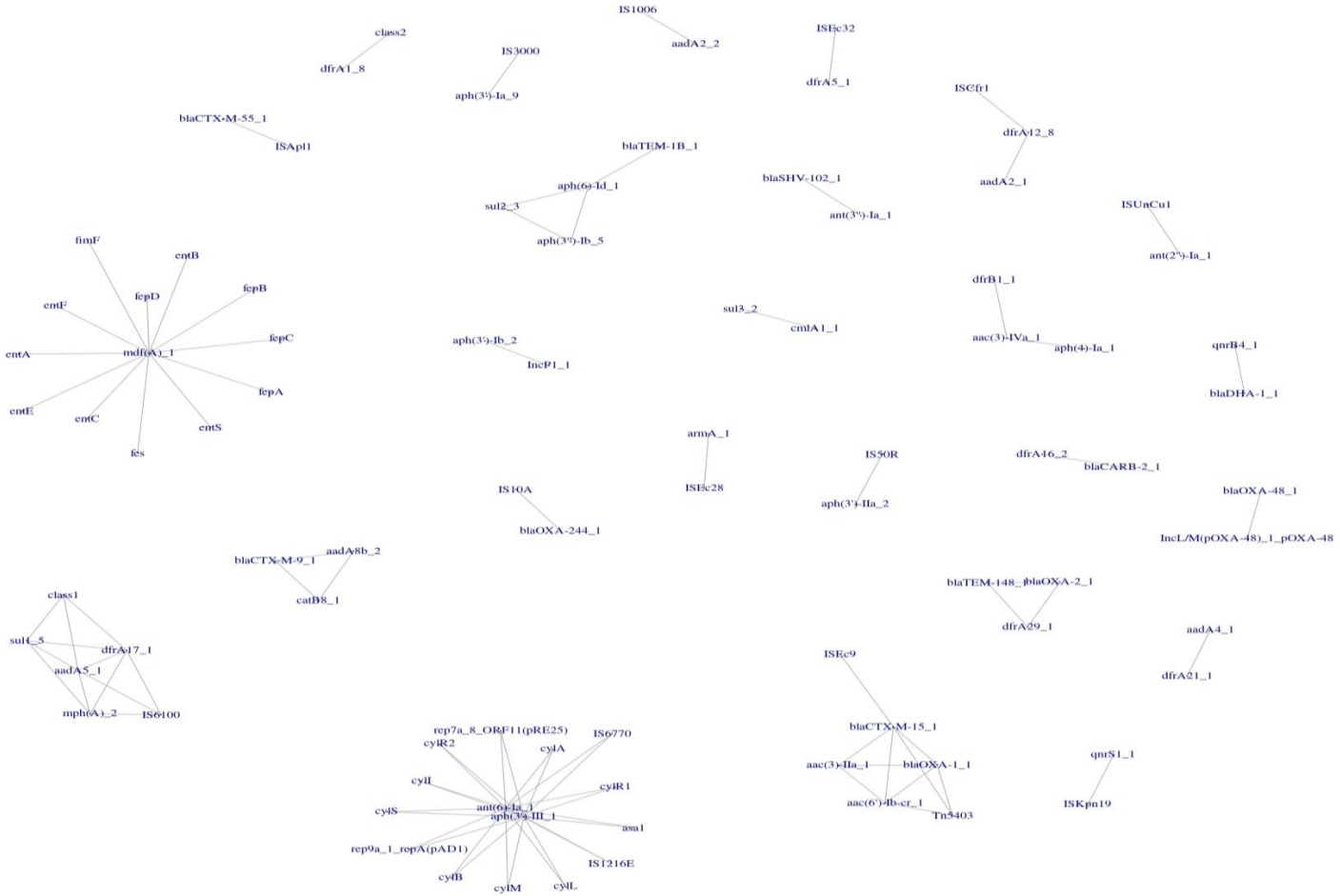


Figure 16 - Networks of genes/plasmids/insertion sequences commonly co-occurring in *E. coli.*

Edge lists were created from elements with a Pearson correlation coefficient >0.5. This figure is also available in higher resolution at https://doi.org/10.6084/m9.figshare.14459139.v2

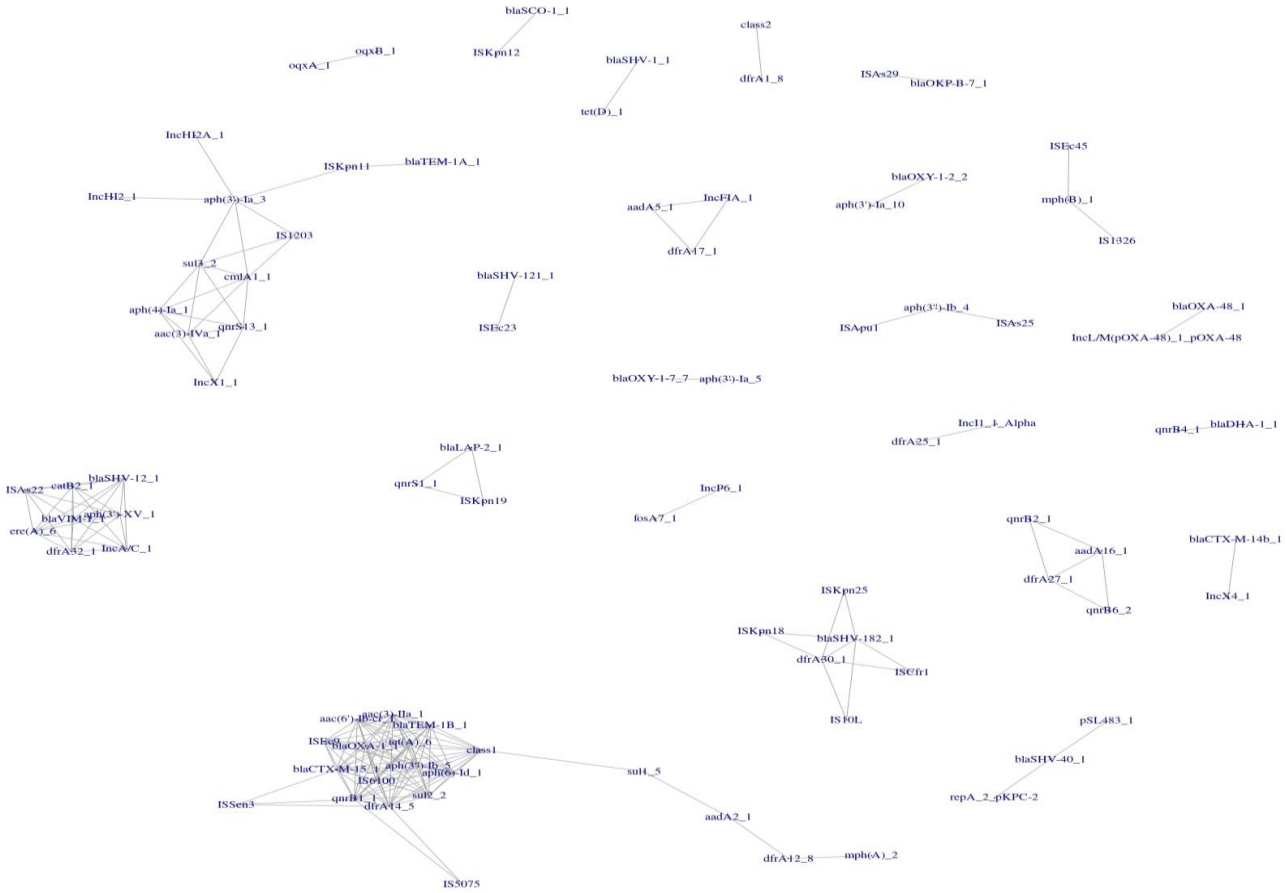


Figure 17 - Networks of genes commonly co-occurring in *Klebsiella* spp.
Edge lists were created from elements with a Pearson correlation coefficient >0.5. This figure is also available in higher resolution at https://doi.org/10.6084/m9.figshare.14459139.v2

2.4 Discussion

In this unbiased longitudinal sequencing study of all (90 day-deduplicated) *E. coli* and *Klebsiella* spp. BSIs in Oxfordshire over a decade (2008-2018), I highlight the similarities and differences in the molecular epidemiology of these species. Overall, the increasing incidence of *E. coli* and *Klebsiella* spp. BSIs in Oxfordshire is not explained by the expansion of a single ST, and much of the burden (40%) of *Klebsiella* spp. disease is caused by non-*K. pneumoniae* species. Although six lineages stably account for nearly half of all *E. coli* BSIs and a clonal outbreak was observed in *K. pneumoniae*, many BSIs in this setting are caused by diverse strains with diverse accessory genomes.

I found no genomic evidence supporting the stratification of *E. coli* isolates into HA-BSI vs CA-BSI. However, strains causing healthcare-associated BSIs may not be specifically healthcare-acquired but may still be healthcare-provoked (for example by the presence of indwelling devices or relative immunosuppression). Interventions targeting only HA infections (such as that of the UK government to reduce these by 50% by 2021[70]) might therefore have limited efficacy without considering the wider ecology of these species. Whilst the incidence of *E. coli* BSIs attributed to the urinary tract increased over the study period, there was evidence that the rate of this increase was slower compared to other sources. This possibly reflects some success of infection prevention interventions such as catheter care (though this is just a hypothesis given the intensive focus on catheter care in the UK in recent years[118]; I did not have any data on urinary catheters), but also highlights the multi-faceted approach required to reduce the overall incidence rate. More work is required to understand CA-*Klebsiella* spp. BSIs because known markers of virulence found in CA disease (e.g. aerobactin, yersiniabactin, colibactin) in other studies[66] were not seen in most of my cases. Importantly therefore, existing markers may

therefore be insensitive to detect the emergence of clinically hypervirulent, CA, multidrug-resistant strains.

For *E. coli*, my analysis is consistent with previous studies demonstrating a broadly stable population structure at the MLST level with the predominance of STs 69, 73, 95 and 131[41,42]. Within this however, I demonstrated some replacement in sub-clades of major STs over the study period. This may represent an adaptive strategy within major STs that allows them to continue to maintain their niche and evolve to survive changing environmental and immunological selection pressures. As has been previously noted, virulence factors in *E. coli* were structured by the underlying phylogeny[119,120]. Differences in carriage of haem utilisation genes were a notable factor discriminating isolates in major STs from the rest of the population, suggesting that iron metabolism might be an important selective pressure in determining invasive potential, a finding supported by a recent GWAS of a mouse model[121]. In contrast, for *Klebsiella* spp. I demonstrated that most infections are caused by sporadic STs which individually only accounted for only a small proportion of the overall population causing disease.

In *E. coli*, the incidence of isolates carrying genes conferring resistance against commonly used antimicrobial classes increased faster than those without; in general the opposite was true for *Klebsiella* spp.. The most commonly isolated MDR-ST in *Klebsiella* spp. (ST490) significantly decreased in incidence over the study period, lending credence to the idea that the relentless expansion of MDR clades is not inevitable, although it is unclear what interventions may have helped in causing its decline, particularly given that approximately a third of cases were community onset. The increasing incidence of gentamicin/ceftriaxone resistance in predominantly CA-*E. coli* BSIs seems paradoxical given that these antibiotics are rarely used in this setting (although the use of more narrow-spectrum beta-lactamases still provides some selective pressure). One explanation is that they are co-located on MGEs with other AMR genes encoding

resistance to antibiotics more commonly used in the community (e.g. amoxicillin, trimethoprim), or to genes that are facilitating selection in response to another evolutionary pressure. An alternative hypothesis might be increased exposure to third-generation cephalosporins due to the rise of ambulatory/"hospital-at-home" medical pathways (the Oxford University Hospitals NHS Foundation Trust opened two Ambulatory Assessment Units in 2015 and two community based Emergency Medicine Units in 2010 and 2012[122]) where the once-a-day dosing of ceftriaxone makes it a relatively widely prescribed antibiotic in these settings. Regardless, overall, *Klebsiella* strains causing BSI appear to be exposed to declining antibiotic/associated selection pressures compared to *E. coli*, suggesting that they are maintained and selected for in distinct ecological niches.

Whilst my analysis did not generate a clear signal of pathogen genomic factors responsible for the observed incidence trends this does not mean that no such factors exist as the analysis is limited by the fact that the vast majority of genes are not annotatable using current databases and as such their function is unknown. Other more agnostic approaches such as the recently described "Temporal GWAS" technique may identify novel factors which have a temporal trend[123]. There may also be a range of non-pathogen factors (e.g. antibiotic exposure, age and comorbidities) not explored in detail here driving the observed trends. Similarly ascertainment bias (driven by e.g. the huge publicity drive around sepsis in the UK in recent years, greater awareness of the need for blood cultures in atypical presentations, improved data capture following the introduction of mandatory surveillance) may explain some of the increases although a recent study in Oxfordshire found that this was unlikely to be a significant factor[62]. The observed trends are similar to those observed globally (at least in high income countries), suggesting that the factors driving the observed trends are not specific to the UK[63,124].

In *E. coli* BSI isolates, there was no genetic signal of adaptation to either healthcare or community environments, suggesting that these are not relevant niches for selection,

contrary to a previous study in Scotland (2013-2015)[42]. However, this previous study included small numbers of isolates (n=162) and may have been unable to fully account for population structure. Furthermore my analysis suggests that, unlike true nosocomial pathogens[125], the ecology of the *E. coli* plasmidome is similar for HA/CA infections. My findings support the hypothesis that the diversity observed in any given epidemiological strata (e.g. age groups/infection focus/healthcare setting) is sampled from the same common ecological pool of isolates with invasive potential rather than representing specialised adaptation to any given setting. There are of course exceptions to this, such as localised outbreaks within care facilities and hospital environments.[126] However, my data suggests that at least in Oxfordshire, these outbreaks are do not contribute significantly to the epidemiology of *E. coli* BSIs. Importantly, iatrogenic, non-pathogen-associated factors promoting invasive infection (e.g. urinary catheterisation) should be minimised to reduce the incidence of *E. coli* BSIs.

For *Klebsiella* spp., the traditional view that HA infections are opportunistic and caused by MDR isolates and CA infections are caused by isolates carrying one or more specific virulence genes (e.g yersiniabactin, colibactin, aerobactin, salmochelin) appears to be an over-simplification in this setting[65,76]. The majority of CA *Klebisella* spp. infections contained none of the known genetic markers of hypervirulence, and these genes are unlikely to be a reliable method of surveillance for emerging strains with a propensity to cause invasive CA disease, at least in Oxfordshire. Whilst ESBL and MDR isolates were significantly more common in HA isolates, about a third were CA (even using my fairly conservative definition) which significantly challenges the prevailing dogma that these are largely opportunistic HA strains.

My study was limited by the inability to reconstruct closed genomes/plasmids using short-read sequencing data and my plasmidome analysis should therefore be interpreted with caution. The widely used PlasmidFinder database allows some inferences to be made using short read data, however there are many untypeable plasmids that are not reflected in this database[127]. Similarly the MLplasmids classification algorithm used is trained on a limited reference set which may lead to some erroneous contig classifications. Additionally, analysis of the entire plasmidome may be too crude to identify the nuances of similarities/differences, particularly for small, shared plasmids as these represent only a small proportion of the overall plasmidome. Whilst I used a relatively conservative definition of healthcare-associated (i.e. within 30 days of hospital admission), this may have failed to capture the longer-term impacts of earlier hospital admissions/other healthcare exposures.

2.5 Conclusions

The contrasting epidemiology of *E. coli* and *Klebsiella* spp. BSIs suggests different reservoirs, selection pressures and modes of acquisition for these genera. Separate strategies to reduce the incidence of these infections are likely to be required, and consideration of the community as a reservoir is important. Given the stable population structure demonstrated here, vaccines may be a promising prospect to reduce the incidence of *E. coli* and *K. pneumoniae* BSI. For *E. coli* such a vaccine (ExPEC4/10V, Janssen Pharmaceuticals[128]) is in clinical trials and in chapter 3 I investigate its potential to provide protection against most *E. coli* isolates causing BSI in Oxfordshire[129]. The lack of reproducibility of several findings from previous studies and the poor sensitivity of current molecular markers for hyper-virulent *Klebsiella* surveillance highlights the critical importance of unselected sampling frames when making epidemiological inferences.

3 Seroepidemiological study of *E. coli*/*Klebsiella* spp. BSIs in Oxfordshire UK

Note: This chapter has previously been published[129]. I authored the text and figures below (which are reproduced here largely unaltered from the published work) with input from supervisors commensurate with the amount of input/advice that would be considered appropriate for a DPhil thesis. All authors have agreed to the inclusion of this published work in my thesis.

3.1 Introduction

As described at the start of chapter 2, the incidence of bloodstream infections (BSIs) caused by *Enterobacteriaceae* (e.g. *E. coli*, *K. pneumoniae*) continues to increase globally[63,130,131]. *Enterobacteriaceae* are a significant antimicrobial resistance (AMR) threat as they can acquire new genes horizontally and therefore rapidly adapt to changing selection pressures, including antimicrobial use. With the limited development of new antimicrobials to treat multidrug-resistant (MDR) infections, the threat of untreatable disease is a significant risk[69]. To date, a nationally mandated set of infection prevention and control standards arising from the Health and Social Care Act of 2008[132] has had limited success in reducing the incidence of BSIs caused by *Enterobacteriaceae*, as exemplified by national surveillance data[15,69].

Prophylactic vaccines represent an alternative approach to combating antimicrobial resistance (AMR) by reducing antibiotic usage and preventing infections caused by AMR-associated strains. For *E. coli*, the O-antigen, a component of lipopolysaccharide (LPS),

has been considered the most promising vaccine target[133]. A recently developed bioconjugate vaccine (ExPEC4V; Janssen Pharmaceuticals) targeting four *E. coli* O-antigens (O1A, O2, O6A and O25B) has subsequently shown promise in a Phase II study[134]. Janssen are now focussing on the development of a 9-valent vaccine in a large phase-3 trial (NCT04899336). For *K. pneumoniae,* vaccine development is lagging behind[135], and there are no vaccines currently in clinical trials.

For *E. coli*, recent serotyping studies have described limited O-antigen diversity in foodborne[136], Enterotoxigenic/Enteropathogenic *E. coli* (ETEC/EPEC)[136] and extended-spectrum beta-lactamase (ESBL)-producing BSI isolates[137]. To my knowledge however, there are no recent systematic and longitudinal studies characterising the seroepidemiology of unselected BSIs. Recent studies of multi-country isolate collections which were non-systematically collected and/or enriched for antibiotic resistance have suggested that the O-antigen could also be a promising vaccine target for *K. pneumoniae*. Follador and colleagues analysed four sequencing datasets (one pre-1950, two others enriched for AMR/clonal outbreaks) and found that three O-antigens (O1/2/3) accounted for 80% of isolates whereas the diversity of K-antigens was an order of magnitude higher[138]. Similarly the O1/2/3/5 antigens accounted for 90.1% of isolates in a recent global sequencing study[139], though there were only relatively small numbers of isolates over limited timeframes from each country and it was unclear how these had been selected.

Large scale, systematic typing studies could help to inform the rational selection of vaccine targets across population groups which best mitigate the risk of BSI (including those associated with AMR) and provide a baseline for monitoring the emergence of serotype variation in response to vaccine roll-out. I performed a large, systematic study

over ten years in Oxfordshire and combined this with an analysis of publicly available global datasets to provide a robust basis for future vaccine development.

3.2 Methods

3.2.1 Study description

All blood culture isolates identified by the Oxford University Hospitals clinical microbiology laboratory between September 2008-December 2018 as being *K. pneumoniae*/*E. coli* (by matrix-assisted laser desorption/ionisation time-of-flight (MALDI-TOF)) analysis (from 2013) and by growth on chromogenic/MacConkey agar and analytical profile index tests (prior to 2013) were included in the study. One *E. coli* isolate was excluded due to a cataloguing error. Infection control procedures were in place for inpatients known to be colonized with ESBL/carbapenemase-producing *Enterobacteriaceae* (CPE) during the study, including isolation in side rooms where feasible, enhanced contact precautions, and screening of high-risk patients on admission (for CPE). To avoid double counting instances where patients had several positive blood cultures taken during a single episode of infection, I considered only the first available isolate per 90-day period per patient.

Isolate processing, DNA extraction, library preparation and sequencing was performed as described in chapter 2. To place my results in a global context, I additionally analysed 3678 isolates from existing studies selected to encompass the recent national/international longitudinal datasets available in the literature as well as isolates that were relatively rare in the Oxfordshire collection (e.g. carbapenemase

producing/genotypically hypervirulent isolates; data available through NCBI project accession numbers: PRJNA480723[40], PRJEB4681[41], PRJEB35000[52], PRJEB12513[42], PRJEB30913[66], PRJEB1271[67], PRJEB10018[71,140]).

3.2.2 Statistics

Stacked negative binomial regression was used to compare incidence rate ratios over time (per year longer, IRRy) of BSI belonging to different antigen groups[112]. For each serotype or group of serotypes examined, I modelled the incidence of isolates with and without this trait over calendar time. In order to include as many isolates as possible in these models, and to enable winter seasons to be grouped together rather than arbitrarily split between calendar years, I defined calendar years as beginning in November (i.e. isolates before November 2008/after November 2018 are excluded from this part of the analysis). A Wald test for heterogeneity was performed to test the null hypothesis that the incidence trend did not differ between isolates with and without the trait (e.g. carriage of particular O-antigens). Healthcare-associated (HA) infection was defined as either nosocomial infection (onset >48 hours after admission) or onset within 30 days of last discharge and community-associated (CA) as that occurring >30 days since last admission[62,141]. Chi-Square or Fisher's exact tests were used to compare proportions between groups. Statistical analysis was performed using R version 4.0.2, except for stacked negative binomial regression models which were implemented in STATA version 15. Sequencing data was linked to electronic healthcare records via the Infections in Oxfordshire Research Database (IORD), as previously (see chapter 2).

3.3.3 Bioinformatics

De novo assembly was performed using Shovill (version 1.0.4)[81]. *In silico* serotyping was performed using the Ectyper tool[142–144] for *E. coli* and with Kleborate for *Klebsiella* spp.[94] For *E. coli* serotypes I included only those isolates which passed the ECtyper quality control (QC) filter and had a high confidence locus call; similarly for *Klebsiella* spp. I included only isolates where the Kleborate locus confidence call was at least 'good' (found in a single piece or with ≥95% coverage, with ≤3 missing genes and ≤1 extra genes)[94]. Genes (yersiniabactin, colibactin and aerobactin) associated with hypervirulence (defined as the ability to cause severe invasive disease in otherwise healthy individuals in the community[5]) were detected using the Kleborate tool. Virulence score was part of the Kleborate output, namely: 0 = no acquired virulence factors, 1= yersiniabactin only, 2 = yersiniabactin and colibactin (or colibactin only), 3 = aerobactin only, 4 = aerobactin and yersiniabactin, 5 = yersiniabactin, colibactin and aerobactin.

3.3.4 Antibiotic resistance

To estimate the potential impact of vaccines to protect against antibiotic resistant disease I compared the carriage of antimicrobial resistance genes (ARGs) between isolates with different serotypes using *in silico* predictions of gene carriage (as called using Kleborate[60]).

3.4 Results

A total of 3479 *E. coli* isolates were sequenced of which 3278 passed *in silico* serotyping quality control. From these, 106 unique O-antigens were identified, though a relatively small number of these accounted for the majority of isolates (Figure 18). The overall incidence of *E. coli* BSIs increased over the study period IRRy=1.14; 95% CI:1.11-1.16. The four most common serotypes (O1A, O2, O6A and O25B; i.e. those covered by the ExPEC4V vaccine) were identified in 1499/3278 (46%) isolates. Their incidence increased over time (IRRy=1.14; 95% CI:1.11-1.16); as did non-ExPEC4V serotypes (IRRy=1.13; 95% CI:1.10-1.16) ($p_{heterogeneity}$=0.7, Fig.S1). There was no evidence that O1/O2/O6/O25 serotypes differed in prevalence between CA and HA isolates (912/1969, 46% vs 576/1292, 45% respectively; p=0.3), or across five age-groups (<18 years 52/122 (43%), 18-39 years 98/222(44%) , 40-59 years 204/477 (43%), 60-79 years 577/1282 (45%), >80 years 557/1158(48%); p=0.3). The ExPEC10V serotypes (O1A, O2, O4, O6A, O8, O15, O16, O18A, O25B and O75) were identified in 2347/3278 (72%) isolates. This vaccine does not include O17 which was the 5th most common O-antigen type in my study, accounting for 181/3278 (6%) isolates.

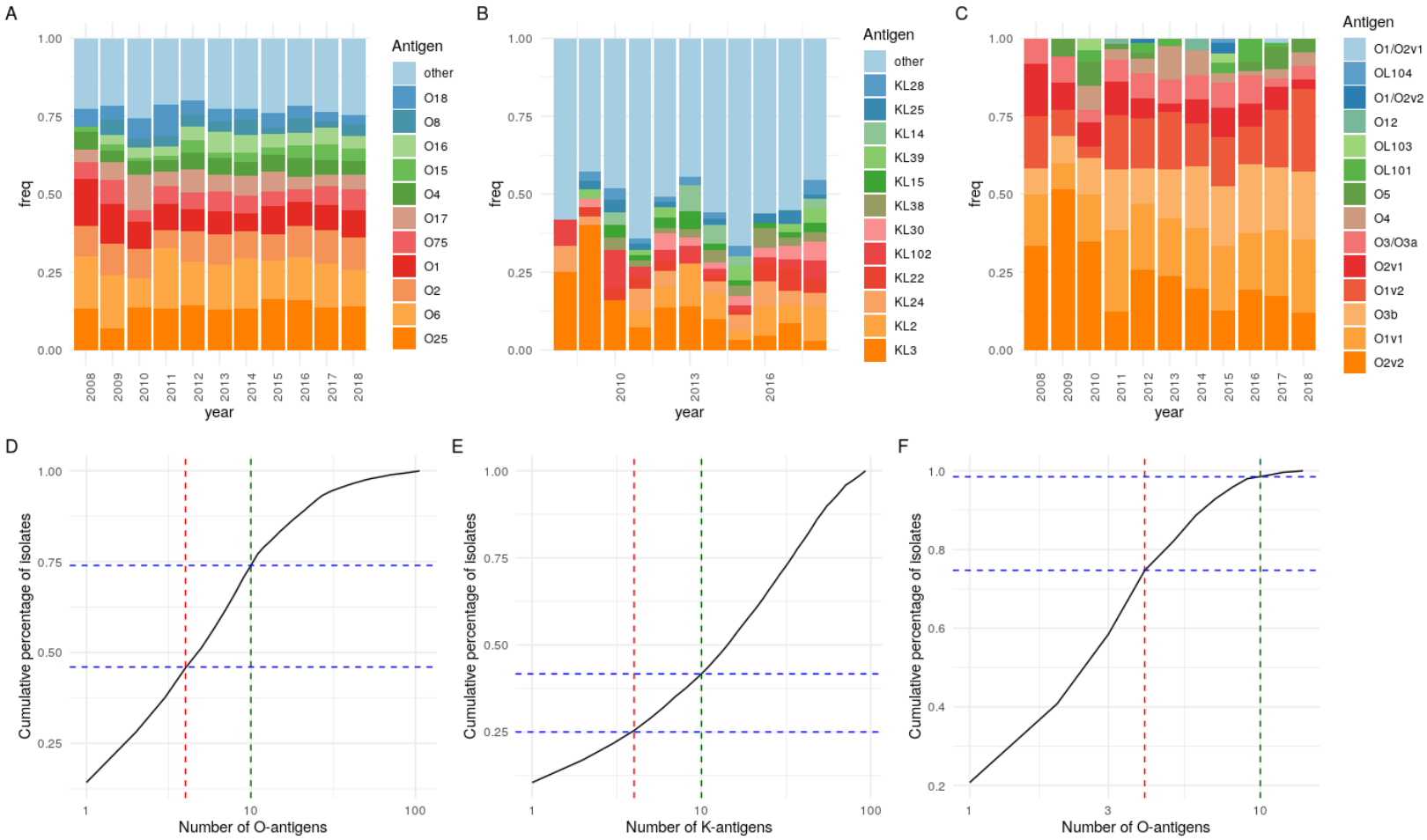


Figure 18 Limited O-antigen diversity in Oxfordshire *E. coli* and *K. pneumoniae* BSIs with a broadly stable population structure over time.
Proportion of: A) *E. coli* O-antigens; B) *Klebsiella* spp. capsular antigens; and C) *Klebsiella* spp. O-antigens observed in all isolates over the ten-year period.
In D-F), the red/green dashed lines denote the x-axis position of the four/ten most prevalent O/K-types respectively, and the blue dashed line demonstrates the cumulative percentage of isolates which have these antigens, for *E. coli* O-antigens (panel D), *Klebsiella* spp. capsular (K) antigens (panel E), and *Klebsiella* spp. O-antigens (panel F).

Of the 3251 isolates for which *in vitro* antimicrobial sensitivity data was available, the ExPEC4V serotypes accounted for 191/331 (58%) of ESBL-producing and 394/834 (47%) of multi-drug resistant (MDR; resistance to ≥3 antibiotic classes) isolates (vs all isolates 1499/3278 (46%) as above, p<0.001). The ExPEC10V serotypes accounted for 592/834 (71%) of MDR and 255/331 (77%) of ESBL isolates (similar to the value observed for all isolates 2347/3278 (72%), p=0.09). O25 was particularly associated with both ESBL (162/450, 36%) and MDR (249/450, 55%) phenotypes (Figure 19). Of the two carbapenem-resistant Oxford *E. coli* BSI isolates, neither had an O-antigen contained in the ExPEC4V/10V vaccines (one was O17 and the other O9).

Similarly, we sequenced 556 *K. pneumoniae* (and *quasipneumoniae)* isolates; 93 unique K- and 14 unique O-loci were confidently identified in 535 (K-antigen)/549 (O-antigen) isolates passing quality control. The incidence of *K. pneumoniae* BSIs also increased overall over the study period (IRRy=1.09; 95%CI:1.05-1.12). To determine the theoretical feasibility of developing a *K. pneumoniae* vaccine and for comparison with *E. coli*, I again considered the top four K- and O-loci. KL2, KL22, KL24 and KL3 were identified in 136/535 (25%) isolates (Figure 18). There was no evidence that the incidence of BSIs caused by this group of capsular types changed over time (IRRy=1.03; 95%CI:0.94-1.13), nor that incidence trends varied between these and other *Klebsiella* spp. BSIs ($p_{heterogeneity}$=0.23). By contrast, O2v2, O1v1, O3b and O1v2 accounted for 410/549 (75%) isolates and the incidence of BSIs caused by these did increase over time (IRRy=1.09; 95%CI:1.05-1.12), though again there was no evidence that these incidence trends differed from all other O-types ($p_{heterogeneity}$=0.8). A comparable Klebsiella vaccine to the ExPEC10V (i.e. composed of the ten most common O-antigens, O2v2, O1v1, O3b, O1v2, O2v1, O3/O3a, O4, O5, OL101, and one of O1/O2v2, O12, OL103 (which were tied for 10th most common O-antigen)), would cover 541/549 (99%) of isolates in my study.

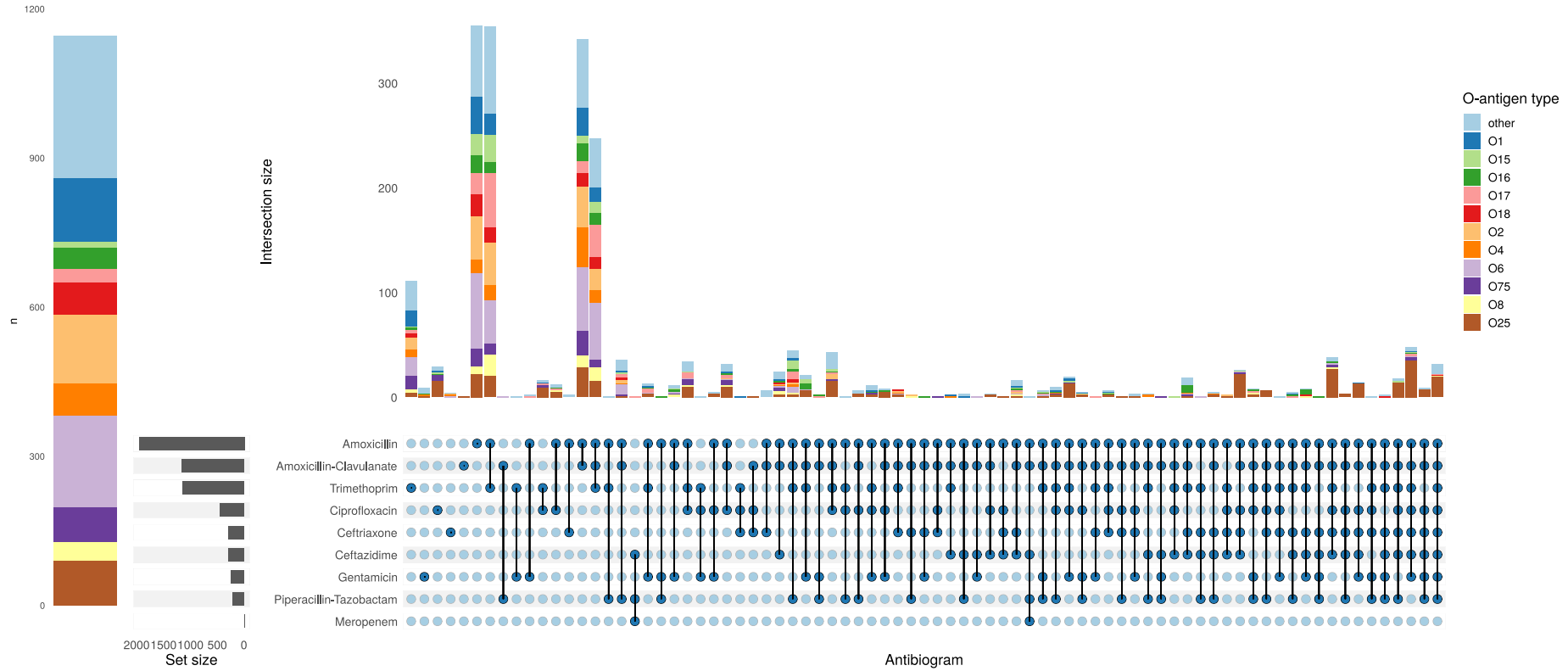


Figure 19 – Distribution of AMR phenotypes by *E. coli* O-antigen serotype

Left: Number of fully sensitive (to antibiotics shown in the righthand upset plot) *E. coli* BSI isolates by O-type. Right: Upset plot showing the number of *E. coli* isolates with phenotypic resistance to the antibiotics shown, stratified by O-antigen type (colours) and ordered by number of intersections.

Horizontal bar plot shows the total number of isolates phenotypically resistance to each antibiotic in any combination. Abbreviation: BSI, bloodstream infection.

KL3 and KL30 were the most common capsular antigens associated with the ESBL phenotype (43/123 [35%] and 12/123 [10%] respectively). KL3 was also the K-antigen most commonly associated with MDR (45/148 [30%]). In total the four most common K-antigens overall (KL2/242/24/3) accounted for 50/123 (41%) of ESBL and 65/148 (44%) of MDR isolates (compared to 136/535 (25%) for all isolates as above, $p<0.001$ i.e. significant evidence of enrichment of resistant isolates). Similarly, the four most common O-antigens accounted for 106/123 (86%) of ESBL and 122/148 (82%) MDR isolates (vs 410/549 (75%) overall $p=0.007$), with no evidence that the proportion of HA/CA BSI caused by these four antigens differed (179/238 [75%] vs 230/310 [74%], respectively; $p=0.9$).

Previous studies have suggested that clinically hypervirulent *K. pneumoniae* isolates (i.e. causing CA invasive infection) are particularly associated with the KL1/2 capsular types[145]. There were only two isolates with the KL1 capsular antigen in my study, only one of which carried the virulence-associated genes encoding for yersiniabactin, colibactin, aerobactin, salmochelin, and *rmpA*. Whilst KL2 did have the most isolates carrying hypervirulence genes of any capsular type, there were also 13/35 KL2 isolates with no hypervirulence loci (Fig.3). 36/37 KL1/2 isolates had O1v1 or O1v2 antigens (the other had O1/O2v2) meaning that a quadrivalent O-antigen-targeted vaccine would theoretically provide good coverage against this important group of isolates. Whilst there were more CA than HA isolates for the KL2 capsular type (19 vs 15 respectively), this was not the only K-type where this was true and the clinically hypervirulent phenotype was observed in 74 (80%) capsular types in total.

As the isolates from my study originated from a single region of the UK with a relatively low AMR burden, I compared the serotypes to those from recently published studies in

the UK[41,42,67] and globally[40,52,66,71,140]. Overall, whilst there was some regional variation, findings supported the potential for O-antigen-targeted quadrivalent vaccines to contribute to substantial reductions in overall BSIs, as well as those caused by ESBL and MDR isolates, in both *E. coli* and *K. pneumoniae* (Table 6). For *K. pneumoniae* they would likely provide a good level of protection against carbapenem-resistant and (genetically) hypervirulent isolates.

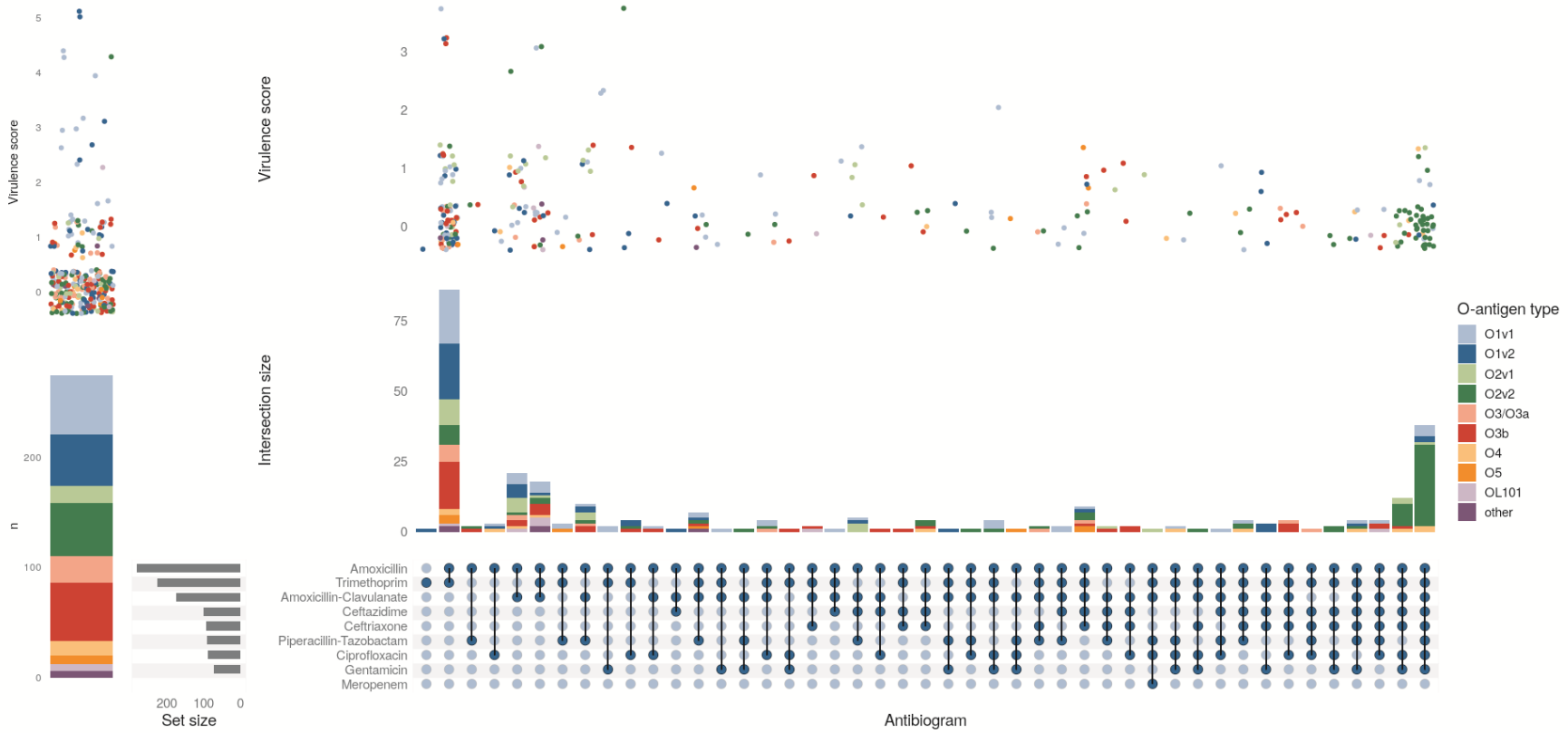


Figure 20 - Distribution of AMR phenotypes by K. pneumoniae O-antigen serotype

Left: Number of K. pneumoniae BSI isolates either fully susceptible or resistant only to amoxicillin shown by O-antigen type. The jitter plot above shows the distribution of virulence scores, points are coloured according to their O-antigen type. Right: Upset plot showing the number of K. pneumoniae isolates with resistance to the classes of antibiotics shown. Stacked bar plot shows the distribution of these by O-antigen type, and the jitter plots above demonstrate the corresponding distributions of virulence scores. Virulence score is as defined by Kleborate (see methods). Abbreviation: BSI, bloodstream infection.

Table 6- Comparison of the Oxfordshire BSIs with those from similar national/regional studies.

For each group I considered the proportion of isolates in each study carrying the four most prevalent antigens in the Oxfordshire dataset to consider the hypothetical coverage a vaccine targeted at these might provide. CUH - Cambridge University Hospitals, EUSCAPE - The European Survey of Carbapenemase-Producing Enterobacteriaceae, SE Asia - South-east Asia. 3479 *E. coli* isolates from Oxfordshire were sequenced of which 3278 passed Ectyper QC. 556 Oxfordshire *K. pneumoniae* isolates were sequenced of which 535 (K-antigen)/549 (O-antigen) passed Kleborate QC. 1510 *E. coli* isolates from published datasets were included of which 1411 passed Ectyper QC. Of the 2168 *K. pneumoniae* isolates included from published datasets, 2029 (K-antigen)/2123 (O-antigen) passed Klerorate QC typing. The ESBL/MDR/carbapenem resistant numbers are based on gene detection from Kleborate, including for Oxfordshire isolates (in contrast to the main text where I report the results of in vitro drug susceptibility testing).

***E. coli* - O antigen**	**Proportion from O1/O2/O6/O25**						
Study	**Oxford**	**CUH**[41,67]	**EUSCAPE**[71,140]	**SE Asia**[66]	**Netherlands**[52]	**Australia**[40]	**Scotland**[12]
Isolate group							
All	1499/3278 (46%)	205/399 (51%)	130/329 (40%)	NA	188/445 (42%)	44/79 (56%)	64/159 (40%)
ESBL	173/256 (68%)	30/36 (83%)	61/154 (40%)	NA	113/244 (46%)	12/15 (80%)	12/15 (80%)
MDR	616/1434 (43%)	95/192 (49%)	87/244 (36%)	NA	113/274 (41%)	26/44 (59%)	27/85 (32%)
Carbapenem resistant	0/2 (0%)	NA	18/86 (21%)	NA	1/2 (50%)	NA	NA
K. pneumoniae - K antigen	**Proportion from KL2/KL242/KL24/KL3**						
All	136/535 (25%)	21/152 (14%)	261/1634 (16%)	56/243 (23%)	NA	NA	NA

ESBL	50/123 (41%)	8/94 (9%)	124/856 (14%)	30/107 (28%)	NA	NA	NA
MDR	65/148 (44%)	12/110 (11%)	181/1221 (15%)	37/130 (28%)	NA	NA	NA
Carbapenem resistant	1/3 (33%)	0/2 (0%)	81/678 (12%)	2/15 (13%)	NA	NA	NA
Hypervirulent	15/28 (54%)	0/1 (0%)	21/55 (38%)	20/72 (28%)	NA	NA	NA
***K. pneumoniae* - O antigen**	**Proportion from O2v2, O1v1, O3b and O1v2**						
All	410/549 (75%)	119/154 (77%)	1173/1651 (71%)	228/318 (72%)	NA	NA	NA
ESBL	106/123 (86%)	84/96 (88%)	592/864 (68%)	116/157 (74%)	NA	NA	NA
MDR	122/148 (82%)	98/112 (88%)	884/1231 (72%)	137/191 (72%)	NA	NA	NA
Carbapenem resistant	0/3 (0%)	2/2 (100%)	521/684 (76%)	34/45 (76%)	NA	NA	NA
Hypervirulent	27/28 (96%)	1/1 (100%)	47/55 (85%)	70/87 (80%)	NA	NA	NA

Conversely for *E. coli*, only 19/90 (21%) isolates carrying carbapenem resistance genes in all studies carried O-antigens covered by the ExPEC4V vaccine, with the O102 (n=18), O25 (n=11), O8 (n=11) and O160 (n=10) antigens predominating in these CPE isolates. This was only slightly improved when considering the antigens in the ExPEC10V vaccine which would cover 35/90 (39%) CPE isolates in all studies. Reflecting this, only 13/90 isolates carrying carbapenem resistance genes belonged to the major group of sequence types (STs) (131/95/73/69) which are responsible for most (predominantly CA) *E. coli* BSIs reported previously[40–42].

3.4 Discussion

A quadrivalent vaccine comprising the four most prevalent O-antigens for *E. coli* and *Klebsiella* spp. could theoretically have a significant impact on the incidence of BSIs and ESBL/MDR infections, as well as offering an alternative approach to help tackle the global epidemic of AMR in *Enterobacteriaceae*. Ten-valent vaccines are at a slightly earlier stage of development for *E. coli* but my data suggests they have the potential to make a significant impact on the incidence of invasive disease and associated antimicrobial resistance. Such a vaccine for *K. pneumoniae* would have covered nearly all isolates (541/549 [99%]) associated with invasive disease over the past decade in Oxfordshire.

Vaccines should also provide some protection against non-BSI infections which may evolve into BSIs, in particular urinary tract infections. To my knowledge there are no studies characterising the serotypes of commensal strains of *E. coli*/*Klebsiella* spp and so it is unknown whether these differ from pathogenic strains. It would be important to understand whether O-antigen vaccinations significantly alter the gut microbiome (which could have intended consequences); metagenomic studies as part of clinical trials and/or

post-marketing surveillance could investigate this and possibly improve the sensitivity for detecting adverse events. As well as reducing morbidity and providing economic benefits, this could also be expected to lower antibiotic selection pressures against *E. coli* and *Klebsiella* spp.. The relatively much higher incidence of these primary infections vs BSIs may increase the economic viability of any prospective vaccine which, particularly for *Klebsiella* spp., may be a significant barrier to development given the difficulty in identifying groups of patients in which vaccination would be cost-effective[146].

I would expect the ExPEC4V/ExPEC10V vaccines to offer less protection against carbapenem-resistant *E. coli* infections. Whilst at first this may appear disappointing, the prevention of ESBL infections would still reduce the carbapenem selection pressure applied by the current practice of treating ESBL-associated BSIs with carbapenems as first-line. Additionally, it was notable that the majority of carbapenem resistance gene carrying *E. coli* isolates were not from the sequence types which have been found to cause most (predominantly CA) BSIs in other studies[40–42]. This might indicate that these isolates do not originate from the community reservoir and that they may be associated with nosocomial transmission as was found to be the case in *K. pneumoniae*[71]. Alternatively these may represent travel imported cases (e.g. from Asia/West Pacific areas previously were carbapenem resistance gene carrying isolates have been shown to be less likely to belong to phylogroup B2)[147]. Widespread dissemination of carbapenem resistance in sequence types associated with global epidemics of CA-MDR disease (e.g. ST131) would be a public health challenge and a vaccine may have a role in preventing this.

There is an implicit assumption in this chapter that O-antigen targeted vaccines are safe, provoke a robust immune response and further that this immune response would actually be protective against invasive disease. There is data from both animal models[148] and early phase human clinical trials[134,149,150] to suggest that the ExPEC4V vaccine (no longer

in active development) was safe and immunogenic. There is also (now rather old) data from animal models suggesting that antibodies against O-antigens confer some protection against death[151,152]. *E. coli* bacteraemia in humans provokes an O-antigen directed antibody response[153,154], though to my knowledge there is no data examination correlating this with clinical outcome. A small phase 1-b trial did however find some evidence of a lower number of "high-count" (>= 10^5 CFU/ml) UTIs amongst participants receiving the ExPEC4V vaccine vs placebo[150]. The ongoing ExPEC9V trial (NCT04899336) should provide a clear answer regarding whether these O-antigen targeted vaccines may have a role in reducing the incidence of *E. coli* BSIs.

The main limitation of my study is that it is from a single region. To mitigate this I analysed data from other published studies; however, in contrast to the Oxfordshire long-term regional surveillance dataset, most of these other studies were either deliberately enriched for MDR isolates or at high risk of bias due to unclear sampling methodologies. Whilst direct comparison with my data is therefore difficult, my overall findings appear to be reflected both nationally and globally. I was unable to characterise the distribution of capsular (K) antigen types for *E. coli* because to my knowledge there is no curated database of sequences currently available, but K-antigens look less promising than O-antigens as vaccine targets in *Klebsiella* spp.. Additionally I was unable to assign an *in silico* O-antigen call to ~5% of isolates in both *E. coli* and *K. pneumoniae*. This may be due to assembly errors or the presence of novel serotypes.

3.5 Conclusions

Overall, O-antigen targeted vaccines for *E. coli* and *Klebsiella* spp. could theoretically have a significant impact on the incidence of BSIs and ESBL/MDR infections, as well as offering an alternative approach to help tackle the global epidemic of AMR in

Enterobacteriaceae. My findings support future development and subsequent efficacy studies for such vaccines, provided they are demonstrated to be safe and immunogenic. The substantial capsular (K) antigen diversity observed for *Klebsiella* spp. BSIs would likely preclude this as a useful vaccine target. Surveillance following the implementation of any vaccine would be essential in identifying any evidence of strain replacement.

4 Bacterial genomic features associated with patient outcome in *E. coli* bloodstream infections

4.1 Introduction

E. coli bloodstream infections are associated with substantial 30-day mortality; reported to be between 10 and 25%[62,155]. In a large 18-year study in Oxfordshire, mortality was stable in all healthcare exposure groups except in the nosocomial group, where there was some evidence that it declined slightly (adjusted IRRy 0.98, 95% CI 0.96-1.00)[62]. Several studies have examined the effect of host/clinical factors on outcome, revealing significantly higher mortality associated with multimorbidity, nosocomial onset, male sex, older age, active cancer and respiratory/gastrointestinal portal of entry[155–157].

Other studies have attempted to incorporate pathogen factors into survival models, though these have largely been restricted to *in vitro* antimicrobial phenotype, population structure (e.g. phylogroup/sequence type) and antimicrobial resistance/virulence genes contained within widely used reference databases. In a study of 162 *E. coli* BSIs in Scotland, there was no evidence of an effect of sequence type on mortality in univariate models, although low sample size undoubtedly limited power[42]. Similarly in a study of 545 patients in Parisian hospitals, there was no evidence of an effect of phylogroup nor most sequence types on mortality, although there was some evidence of increased mortality associated with ST88 (aOR 3.62, 95%CI 1.30-10.09, p=0.014)[155].

More recently a large genome wide association study has been performed on a combined cohort of two studies including 912 BSI patients from Paris[158]. Here the authors were not

able to identify any significant associations between pathogen genomic features and three clinical outcomes (30-day mortality, septic shock and admission to ITU) but did confirm previous findings of the association of the *papG* operon with a urinary portal of entry, also demonstrated in my study in chapter 2. The authors performed some simulations to assess statistical power in their sample and hypothesised that an increase in sample size by one order of magnitude (e.g. ~10,000 isolates) might allow detection of relevant pathogen genomic factors associated with patient outcome.

In this study I therefore sought to utilise the large Oxfordshire *E. coli* BSI WGS dataset linked to detailed clinical data to discover pathogen genomic factors linked to clinical outcomes. Identification of such virulence factors might be useful for future molecular epidemiological studies to better understand the mechanisms for their dissemination and/or predict the risk associated with particular clades. It might also provide insights into the molecular pathogenesis of *E. coli* BSIs and improve clinical risk prediction, which could help clinical teams to prioritise patients most likely to deteriorate. As host factors are known to be a substantial factor in determining patient outcomes, I first created a model of these similar to those that have previously been published. I then tested the association between pathogen genomic factors and 30-day mortality. Finally I combined significant predictors from the host and pathogen models into a unified model.

4.2 Methods

4.2.1 Study description

In this study I included all *E. coli* BSI isolates processed by the Oxford University Hospitals NHS Foundation Trust laboratory at the John Radcliffe Hospital between 2009-2018 (de-duplicated to 90 days per patient episode as in the previous chapters).

4.2.2 Linkage to clinical data

Isolates were linked to clinical data captured in IORD, as in previous chapters. Linkage was performed using isolate accession numbers and collection dates (with a 30-day fuzzy joining rule to account for human data entry error). Isolate accession numbers in our laboratory were recycled approximately every six-months during the study period, and so accession numbers/dates within 30 days would represent unique cases.

4.2.2 Sequencing and bioinformatics

Isolates were sequenced as previously described in chapters 2 and 3 on Illumina platforms[68]. De novo assembly was performed using Shovill[81] (default settings). For the pangenome wide association screen, following annotation with Prokka[82] (default settings), a pangenome was constructed with panaroo (using options –merge_paralogs –clean-mode moderate). AMRFinder[159] was used on default settings with the (-O Escherichia and –plus flags) to annotate ARGs/virulence/stress genes. A phylogeny was created using IQ-tree (generalised time reversable model) with the core genome alignment from Panaroo as input and a similarity matrix subsequently created using the Pyseer phylogeny_distance.py script. An all vs all distance matrix was created using mash (-k31, -s 100,000) and this was represented by multidimensional scaling in R[113].

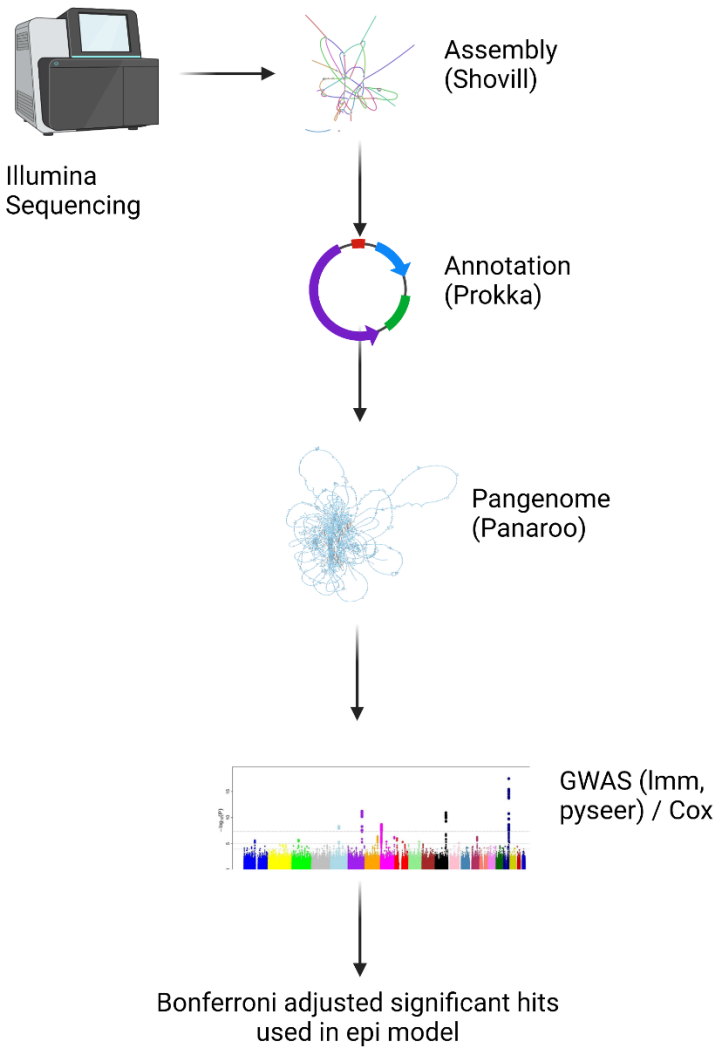


Figure 21 – Sequencing/bioinformatic workflow

4.3.3 Genome wide association study

A GWAS was performed using genes and then unitigs as features using Pyseer[108], with a linear mixed model used to adjust for population structure (using the similarity matrix created above). The binary phenotype was mortality within 30 days of date of positive blood culture collection (coded as 1) or alive at 30 days (0). Features with a Bonferroni adjusted p-value <0.05 were judged to be significantly associated with 30-day mortality.

4.2.4 Patient factors model

I first included all available variables which could possibly confound the relationship between developing an *E. coli* bloodstream infection (exposure) and death (outcome). I then constructed a Cox proportional hazards regression model (Survival library in R) where the outcome was death and patients where censored if still alive at 30 days after their blood culture was collected. Backwards elimination minimising the Akaike information criterion (AIC) was used to construct a final host model incorporating potential confounders at a relatively relaxed significance threshold (AIC corresponding approximately to an exit p value of 0.15). Finally, significant genetic factors from the pathogen model were added into the patient factors model to create the “combined” patient/pathogen model.

4.2.5 Software used for visualisation

Figures were produced using ggplot2[160] in R version 4.1.0[113].

4.3 Results

4.3.1 Characteristics of 3394 patients with *E. coli* BSI

3394 isolates collected between 15/09/2008 and 01/12/2018 had undergone sequencing and could be linked to clinical metadata. Overall 30-day mortality was 517/3394 (15%) and patients that did not survive were on average older, more likely to be male, had more comorbidities and admissions in the previous year and were more likely to be nosocomial onset (

Table 7).

Table 7 - Clinical characteristics of 3394 cases of *E. coli* bloodstream infections.
P values from Kruskal-Wallis/Wilcoxon/chi-squared tests.

Variable		**Censored** N=2877	**Death within 30 days** N=517	**p**
Age	Median (IQR)	72.9 (57.8 to 82.5)	80.2 (69.5 to 86.8)	<0.001
Sex	F	1469 (51.1)	238 (46.0)	0.040
	M	1408 (48.9)	279 (54.0)	
Charlson score	Median (IQR)	1 (0 to 2)	2.0 (1.0 to 3.0)	<0.001
Admissions in previous year	Median (IQR)	1 (0 to 3)	1.0 (0.0 to 3.0)	<0.001
IMD Score	Median (IQR)	10.3 (6.6 to 16.7)	9.6 (6.6 to 16.2)	0.23
Healthcare exposure	Community	1810 (62.9%)	256 (49.5%)	<0.001
	Healthcare	1067 (37.1%)	261 (50.5%)	
Congestive heart failure	No	2437 (89.7%)	400 (81.0%)	<0.001
	Yes	280 (10.3%)	94 (19.0%)	
Cerebrovascular disease	No	2542 (93.6%)	449 (90.9%)	0.039
	Yes	175 (6.4%)	45 (9.1%)	
Dementia	No	2458 (90.5%)	418 (84.6%)	<0.001
	Yes	259 (9.5%)	76 (15.4%)	
COPD	No	2220 (81.7%)	403 (81.6%)	1.00
	Yes	497 (18.3%)	91 (18.4%)	

Variable		**Censored** N=2877	**Death within 30 days** N=517	**p**
Rheumatological disease	No	2606 (95.9%)	473 (95.7%)	0.96
	Yes	111 (4.1%)	21 (4.3%)	
Peptic ulcer disease	No	2656 (97.8%)	479 (97.0%)	0.37
	Yes	61 (2.2%)	15 (3.0%)	
Diabetes without complications	No	2126 (78.2%)	386 (78.1%)	1.00
	Yes	591 (21.8%)	108 (21.9%)	
Diabetes with complications	No	2642 (97.2%)	479 (97.0%)	0.85
	Yes	75 (2.8%)	15 (3.0%)	
Hemi/paraplegia	No	2642 (97.2%)	479 (97.0%)	0.85
	Yes	75 (2.8%)	15 (3.0%)	
Renal disease	No	2348 (86.4%)	384 (77.7%)	<0.001
	Yes	369 (13.6%)	110 (22.3%)	
Malignancy	No	2055 (75.6%)	308 (62.3%)	<0.001
	Yes	662 (24.4%)	186 (37.7%)	
Mild liver disease	No	2522 (92.8%)	455 (92.1%)	0.64
	Yes	195 (7.2%)	39 (7.9%)	

Variable		**Censored** N=2877	**Death within 30 days** N=517	**p**
Moderate/severe liver Disease	No	2653 (97.6%)	460 (93.1%)	<0.001
	Yes	64 (2.4%)	34 (6.9%)	
Portal of entry	Gastrointestinal	184 (9.5%)	54 (15.7%)	<0.001
	Hepatobiliary	349 (18.1%)	50 (14.6%)	
	Other	78 (4.0%)	2 (0.6%)	
	Respiratory	40 (2.1%)	19 (5.5%)	
	Skin/soft tissue	19 (1.0%)	6 (1.7%)	
	Unclear	319 (16.5%)	94 (27.4%)	
	Urinary tract	944 (48.8%)	118 (34.4%)	

4.3.2 Core patient factors model

I included the 20 available baseline factors in Table 7 above as initial explanatory variables and then selected the final model by backwards elimination based on minimising the AIC. In the multivariable Cox model, age (adjusted hazard ratio, aHR 1.03, 95%CI 1.02-1.04, p<0.001), Charlson comorbidity score (aHR 1.27, 95%CI 1.04-1.54, p=0.017), presence of moderate/severe liver disease (aHR 1.69 (95%CI 1.04-2.76, p=0.035) were associated with higher 30-day mortality whereas urinary (aHR 0.47 95%CI 0.33-0.67, p<0.001), hepatobiliary (aHR 0.47 95%CI 0.31-0.72, p=0.001) and other (aHR 0.16, 95%CI 0.04-0.64, p=0.010) portals of entry were associated with lower 30-day mortality (Table 8).

Table 8 - Factors associated with 30-day mortality in the core patient factors model.
Ref – reference category

Variable		HR (univariable)	HR (multivariable)
Age		1.03 (1.02-1.03, p<0.001)	1.03 (1.02-1.04, p<0.001)
Sex	F		
	Male vs female	1.21 (1.02-1.44, p=0.028)	
Charlson score		1.29 (1.22-1.36, p<0.001)	1.27 (1.04-1.54, p=0.017)
Admissions in previous year		1.01 (1.00-1.01, p=0.116)	
IMD Score		1.00 (0.99-1.01, p=0.382)	
Healthcare Exposure	Healthcare vs. community associated	1.63 (1.37-1.94, p<0.001)	
Cerebrovascular Disease	Yes vs No	1.40 (1.03-1.90, p=0.031)	
Congestive Heart Failure		1.91 (1.52-2.39, p<0.001)	
Dementia		1.60 (1.25-2.05, p<0.001)	-

Variable	HR (univariable)	HR (multivariable)
COPD	0.99 (0.79-1.25, p=0.955)	
Rheumatological Disease	1.05 (0.68-1.63, p=0.812)	
Peptic Ulcer Disease	1.34 (0.80-2.23, p=0.270)	
Mild liver disease	1.11 (0.80-1.54, p=0.539)	
Diabetes without complications	1.01 (0.82-1.26, p=0.895)	
Diabetes with complications	1.10 (0.66-1.85, p=0.704)	
Hemi/paraplegia	1.11 (0.66-1.85, p=0.703)	
Renal disease	1.71 (1.38-2.12, p<0.001)	
Malignancy	1.76 (1.47-2.11, p<0.001)	
Moderate/severe liver disease	2.67 (1.88-3.78, p<0.001)	1.69 (1.04-2.76, p=0.035)
Portal of Entry GI	ref	ref

Variable		HR (univariable)	HR (multivariable)
	HPB	0.52 (0.35-0.76, p=0.001)	0.47 (0.31-0.72, p=0.001)
	Other	0.10 (0.02-0.40, p=0.001)	0.16 (0.04-0.64, p=0.010)
	Respiratory	1.58 (0.94-2.67, p=0.084)	1.31 (0.73-2.33, p=0.362)
	Skin/Soft Tissue	1.07 (0.46-2.49, p=0.873)	1.16 (0.49-2.73, p=0.735)
	Unclear	1.05 (0.75-1.46, p=0.789)	0.95 (0.66-1.38, p=0.801)
	Urinary Tract	0.45 (0.33-0.63, p<0.001)	0.47 (0.33-0.67, p<0.001)

4.3.3 Association of population structure with patient outcome

Visual inspection of mortality events on a phylogenetic tree (Figure 22) suggested that these appeared to be randomly distributed and there was no clear concentration in a particular phylogroup. In keeping with this, was no evidence of association between any of the phylogroups and mortality in a univariable or multivariable Cox regression analysis (Table 9). In the univariable analysis there was some evidence of lower mortality for ST95 (compared to ST131, HR 0.61 95%CI 0.41-0.91, p=0.02)) but this association appeared to be confounded by patient/clinical factors (aHR 95%CI 0.73 0.43-1.22, p=0.22).

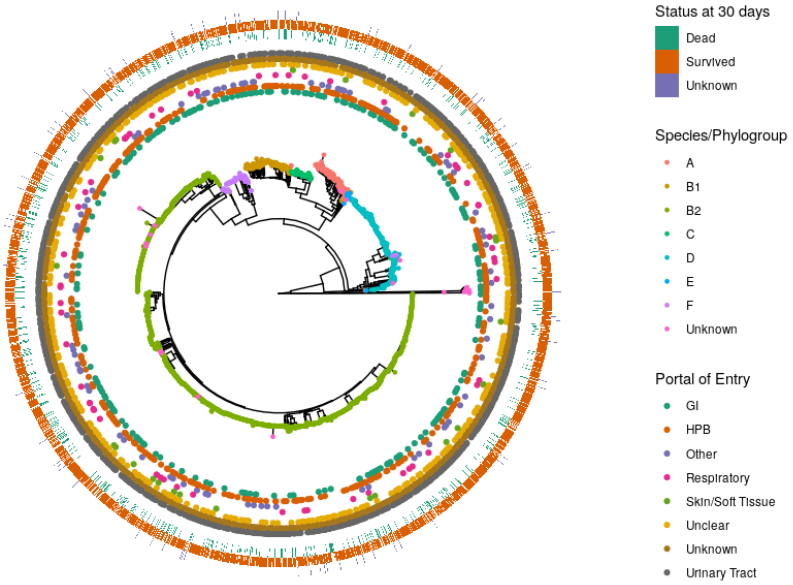


Figure 22 – Phylogenetic tree of isolates included in the study

Tip colours show phylogroup, the inner ring shows portal of entry and survival status at 30 days is shown on the outer ring

Table 9 – Survival models incorporating pathogen population structure

	HR	aHR (phylogroup model)	aHR (MLST model)
Age		1.03 (1.02-1.04, p <0.001)	1.03 (1.02-1.04, p<0.001)
Charlson score		1.21 (1.12-1.29, p <0.001)	1.20 (1.12-1.29, p<0.001)
Moderate/severe liver disease		1.72 (1.05-2.82, p=0.03)	1.70 (1.03-2.78, p=0.03)
Gastrointestinal source		ref	ref
Hepatobiliary source		0.37 (0.25-0.55, p<0.001)	0.35 (0.23-0.53, p<0.001)
Other source		0.14 (0.03-0.58, p<0.001)	0.14 (0.03-0.58, p<0.001)
Respiratory source		1.21 (0.71-2.07, p=0.48)	1.22 (0.71-2.09, p=0.46)
Skin/Soft Tissue source		0.97 (0.41-2.26, p=0.94)	0.97 (0.41-2.28, p=0.95)
Unclear source		0.85 (0.60-1.21, p=0.37)	0.87 (0.61-1.22, p=0.41)
Urinary tract source		0.38 (0.27-0.53, p<0.001	0.39 (0.28-0.55, p<0.001)
Phylogroup			
B2	ref	ref	
A	0.73 (0.47-1.13, p=0.16)	0.58 (0.30-1.14, p=0.11)	
B1	0.88 (0.59-1.31, p=0.53)	0.82 (0.49-1.39, p=0.46)	
C	1.28 (0.81-2.03, p=0.30)	0.98 (0.53-1.79, p=0.94)	
F	0.89 (0.69-1.16, p=0.40)	0.92 (0.66-1.27, p=0.61)	
F	1.18 (0.78-1.80, p=0.44)	1.35 (0.81-2.25, p=0.24)	
Other	1.15 (0.63-2.09, p=0.66)	1.24 (0.58-2.64, p=0.58)	
MLST			
73	ref		ref
10	0.84 (0.39-1.81, p=0.65)		1.21 (0.43-3.38, p=0.72)

12	0.98 (0.62-1.54, p=1.00)	1.09 (0.62-1.91, p=0.76)
127	1.00 (0.60-1.65, p=0.99)	1.50 (0.80-2.83, p=0.20)
131	0.98 (0.72-1.65, p=0.53)	1.22 (0.83-1.79, p=0.31)
69	0.82 (0.57-1.18, p=0.28)	1.05 (0.66-1.67, p=0.83)
95	0.61 (0.41-0.91, p=0.02)	0.73 (0.43-1.22, p=0.22)
Other	0.95 (0.75-1.22, p=0.73)	1.10 (0.80-1.50, p=0.56)

4.3.4 Curated association study

The NCBI AMRFinder database comprises a highly curated set of genes known to confer antimicrobial resistance or be involved in virulence or stress response and therefore might be most likely to be associated with patient outcome. I first screened all 138 elements in this database present in at least 1% of isolates for their association with death in a univariate Cox model. After adjusting for multiple comparisons using the Bonferroni method (N=138 comparisons), only the gene *papGII* remained significant (HR 0.60, 95%CI 0.41-0.80, $p=2.42x10^{-7}$). The *iuc* (aerobactin: *iucA*, *iucB*, *iucC*, *iucD*, *iutA* and *shiF*) operon has previously shown to be associated with invasive UPEC compared to simple cystitis[161], however I found no association with patient outcome in my univariable analysis (*iucA* HR 0.98, 95%CI 0.83-1.16, p=0.82, *iucB* HR, 0.97 95%CI 0.82-1.15, p=0.75, *iucC* HR 0.96, 95% 0.81-1.14, p=0.63, *iucD* HR 0.94, 95% CI 0.79-1.11, p=0.45, *iutA* HR 0.95, 95%CI 0.80-1.13, p=0.57).

4.3.5 Agnostic genome wide association study

Next I explored the association between all genes in the pangenome of the 3394 isolates in my dataset occurring with at least 1% population frequency (N=10,068) with 30-day mortality. To control for population structure, I added ten dimensions of multidimensional scaling of mash distances to the Cox model as covariates. Again the only gene associated with 30-day mortality after adjusting for multiple comparisons (N=10,068) using the Bonferroni method was *papGII.* I repeated this analysis using the Pyseer package which uses Firth logistic regression. This may improve power in cases where the outcome is relatively rare and groups of categorical predictors all have the same phenotype (known as “separation”) which can cause very large standard errors. Pyseer

also uses a linear mixed model incorporating random effects to control for population structure. There were three significant hits in this analysis after adjusting for multiple testing, a recombinase family protein hin (OR 1.23, 95%CI 1.15-1.33, $p=3.17x10^{-6}$), *papGII* (OR 0.94, 95%CI 0.91-0.96. $p=5.22x10^{-7}$) and a hypothetical protein of unclear function (OR 1.34. 95%CI 1.23-1.46, $p=5.22x10^{-7}$). To try to gain some insight into the putative role of the hypothetical protein I used Google Research's ProteInfer tool[162] (accessed 10/5/22) which uses a deep neural network to try to infer protein function. This suggested with high confidence that this protein is an integral component of the cell membrane and/or with less confidence that it may be an organelle membrane protein (Figure 23).

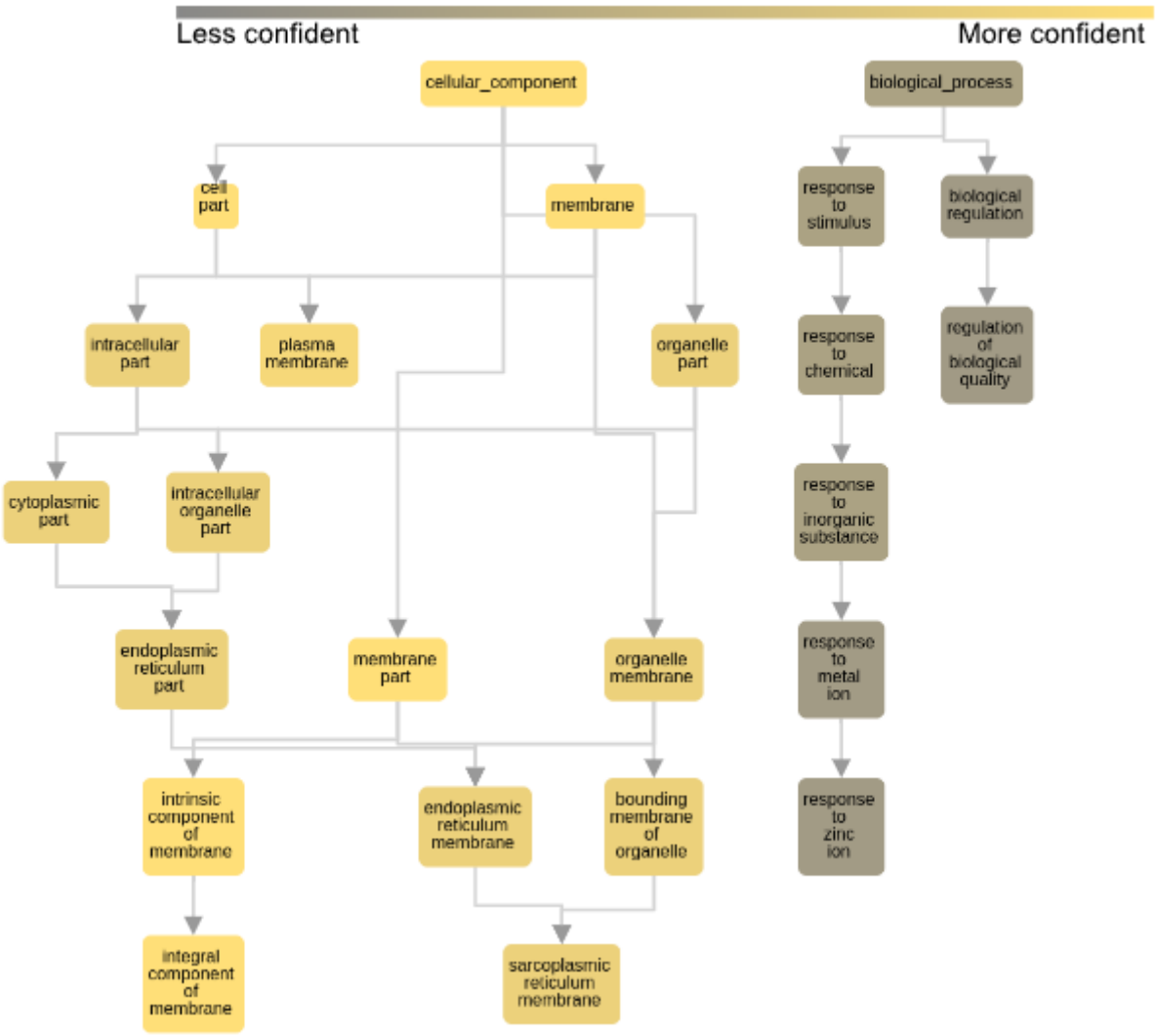


Figure 23 - ProteInfer prediction of the possible function of the hypothetical protein found to be significantly associated with increased 30-day mortality in the Pyseer GWAS.

4.3.6 Combined host/pathogen model

To investigate the extent to which the effects observed for the three genes found to be significantly associated with mortality in the univariable Pyseer GWAS after adjustment for multiple comparisons above are confounded by host factors, I constructed a “combined host/pathogen model”. Firstly I explored the distribution of isolates carrying the three genes identified in the GWAS above by portal of entry, confirming previous observations that papGII is over-represented in association with a urinary source (Table 10). Next I

added all three to the host model ("Combined multivariable model" in Table 11). Whilst there was some evidence that the effect of *papGII* is confounded by patient factors (aHR 0.66, 95%CI 0.51-0.84, p<0.001 vs HR 0.60, 95%CI 0.49-0.73, p<0.001, Table 11), it remained independently associated with reduced mortality in the multivariable model. The *hin2*/the hypothetical protein genes on the other hand appeared to be confounded to a greater extent by patient factors/each other, although the effect estimate was still (non-significantly) in favour of worse outcome. I hypothesised that one explanation of this might be that they were involved in the same pathway/had related functions. To further investigate this hypothesis I modelled the count of these two genes as a categorical variable (i.e. a variable that has value 0 if neither gene is present, 1 if one is and 2 if both are – combined multivariable model 2 in Table 11) and the presence or absence of one or both (i.e. a variable with value 0 if neither are present and 1 if one or both are – combined multivariable model 3 in Table 11). This analysis provided some support for this hypothesis and suggested similar effects associated with the presence of one or both of these genes

Table 10 - Distribution of significant hits from GWAS by portal of bloodstream infection entry
Numbers in Brackets represent percentages.

Gene	**Present**	**Gastrointestinal** (n=238)	**Hepatobiliary** (n=399)	**Missing** (n=1197)	**Other** (n=80)	**Respiratory** (n=59)	**Skin/Soft Tissue** (n=25)	**Unclear** (n=413)	**Urinary Tract** (n=1062)
papGII	no	180 (75.6)	318 (79.7)	740 (61.8)	54 (67.5)	45 (76.3)	16 (64.0)	278 (67.3)	504 (47.5)
	yes	58 (24.4)	81 (20.3)	457 (38.2)	26 (32.5)	14 (23.7)	9 (36.0)	135 (32.7)	558 (52.5)
hin2	no	233 (97.9)	389 (97.5)	1168 (97.6)	76 (95.0)	58 (98.3)	25 (100.0)	408 (98.8)	1053 (99.2)
	yes	5 (2.1)	10 (2.5)	29 (2.4)	4 (5.0)	1 (1.7)	0 (0.0)	5 (1.2)	9 (0.8)
hypothetical protein	no	236 (99.2)	393 (98.5)	1180 (98.6)	79 (98.8)	58 (98.3)	25 (100.0)	407 (98.5)	1057 (99.5)
	yes	2 (0.8)	6 (1.5)	17 (1.4)	1 (1.2)	1 (1.7)	0 (0.0)	6 (1.5)	5 (0.5)

Table 11 – Combined host/pathogen multivariable models of patient outcome in *E. coli* bloodstream infection

Note - the most common phylogroup (B2) and ST (73) are used as the reference (ref) categories. P values <0.05 are highlighted in bold. Combined model 1 includes all genomic factors found to be significantly associated with outcome in the univariable GWAS as well as all variables in the multivariable patient factors model. In combined model 2, the variables representing the presence/absence of the hin2/hypothetical protein genes are replaced with a categorical variable which represents the count of these and in model 3 by a binary variable which has value 1 if either gene is present or 0 if neither are.

	Gene univariable models	**Separate multivariable models** **Host factor only**	**Combined multivariable model 1**	**Combined multivariable model 2**	**Combined multivariable model 3**
	aHR (95% CI)	aHR (95% CI)	aHR (95% CI)	aHR (95% CI)	aHR (95% CI)
Age		1.03 (0.97-1.02, p<0.001)	1.02 (1.02-1.04, <0.001)	1.03 (1.02-1.04, p<0.001)	1.03 (1.02-1.04, p<0.001)
Charlson score		1.21 (0.83-1.13, p<0.001)	1.21 (1.13-1.30, <0.001)	1.22 (1.13-1.31, p<0.001)	1.21 (1.13-1.31, p<0.001)
Moderate/severe liver disease		1.69 (0.59-1.04, p=0.035)	1.74 (1.06-2.84, p=0.028)	1.72 (1.05-2.81, p=0.032)	1.71 (1.05-2.80, p=0.03)
Gastrointestinal source		ref	ref	ref	ref
Hepatobiliary source		0.36 (0.24-0.54, p<0.001)	0.34 (0.23-0.51, p<0.001)	0.34 (0.23-0.51, p<0.001)	0.34 (0.23-0.51, p<0.001)
Other source		0.14 (0.03-0.58, p=0.0067)	0.14 (0.03-0.59, p=0.0070)	0.14 (0.03-0.57, p=0.0062)	0.14 (0.03-0.57, p<0.001)
Respiratory source		1.23 (0.72-2.10, p=0.44)	1.24 (0.73-2.12, p=0.43)	1.25 (0.73-2.13, p=0.41)	1.25 (0.73-2.13, p=0.41)
Skin/Soft Tissue source		0.97 (0.41-2.26, p=0.94)	1.00 (0.42-2.34, p=1.00)	1.01 (0.43-2.35, p=0.99)	1.01 (0.43-2.35, p=0.99)
Unclear source		0.86 (0.61-1.21, p=0.38)	0.88 (0.62-1.25, p=0.48)	0.89 (0.63-1.26, p=0.5)	0.89 (0.63-1.26, p=0.50)

Urinary tract source		0.39 (0.27-0.54, p<0.001)	0.43 (0.31-0.61, p<0.001)	0.43 (0.31-0.61, p<0.001)	0.43 (0.31-0.61, p<0.001)
		Gene only			
papGII	0.60 (0.49-0.73, p<0.001)	0.61 (0.50-0.74, p<0.001)	0.66 (0.51-0.84, p<0.001)	0.65 (0.51-0.84, p<0.001)	0.65 (0.51-0.84, p<0.001)
hin2	2.63 (1.70-4.08, p<0.001)	1.69 (0.78-3.66, p=0.18)	1.48 (0.43-6.17, p=0.53)	-	
Hypothetical protein	3.37 (2.02-5.63, p<0.001)	1.78 (0.72-4.43, p=0.21)	1.86 (0.46-7.62, p=0.39)	-	
Count of *hin2* + hypothetical protein					
0	ref			ref	
1	2.11 (1.05-4.25, p=0.036)			2.42 (0.90-6.56, p=0.081)	
2	3.26 (1.92-5.55, p<0.001)			2.48 (1.10-5.60, p=0.029)	
hin2 and/or hypothetical					2.46 (1.30-4.64, p=0.0057)

4.4 Discussion

In this study I have unified previous findings demonstrating lower mortality associated with a urinary portal of entry in *E. coli* bloodstream infections and the association of *papGII* with the urinary portal of entry. Here, for the first time, I demonstrate an independent association between *papGII* with decreased 30-day mortality even after adjusting for patient factors. My findings validate previous simulation work suggesting that larger sample sizes may provide significant hits in genome wide association studies between *E. coli* BSI and death. Whilst there are now a huge number of publicly available *E. coli* genomes from clinical isolates, the associated available metadata is very poor meaning that such datasets are unhelpful for this sort of analysis and improvements in this area are vital if this ambition is to be achieved.

Targeted PCR-based studies[163,164] and more recently pangenome GWAS[158,161] have previously associated the *papGII* (fimbial adhesin protein) locus with pyelonephritis and BSI of urinary origin; also a finding of this thesis in chapter 2. To cause infection, *E. coli* cells have to adhere to the uroepithelium which allows them to effectively colonise the urinary tract as well as preventing their clearance during urination. *PapG* facilitates the adherence of P fimbriae to the uroepithelium and is therefore a key virulence factor for *E. coli* entering via the urinary tract. Its association with better outcome could be simply due to the fact that it may be a better proxy than mandatory reporting records for urinary portal of entry. *E. coli* BSIs which originate from UTIs may be easier to diagnose, treat and gain effective source control (e.g. by replacing a urinary catheter) compared to those from other portals of entry. However, the fact that the protective effect of *papGII* does not disappear after adjusting for portal of entry may hint that the association is partly caused by another independent mechanism (perhaps provocation of an immune response).

In addition to the *papGII* locus (associated with better outcome), my Pyseer pangenome analysis found significant associations with worse outcome for the DNA invertase *hin* gene and a hypothetical protein of unknown function. The putative role of these in virulence is much less clear though previous studies in *Salmonella* spp. have suggested that *hin* plays a role flagellar phase variation[165], so it is possible that this gene might be involved in immune evasion and persistence. My exploratory analysis hints at the possibility that these two genes could be involved in the same pathway/biological process. The two novel hits I identify in this study would be potential candidates for further evaluation with e.g. knock-out studies in animal models in future work if their impact is replicated in independent datasets.

Interestingly none of my analyses found an association between 30-day mortality and antibiotic resistance genes. This observation is limited by the fact that antimicrobial prescribing data was not available, although the majority of patients would have received empirical co-amoxiclav +/- gentamicin. Nevertheless this limitation may have obscured any signal. It is also possible that resistance to these drugs (particularly co-amoxiclav) is a multi-gene phenomenon which we were underpowered to detect here. This finding is however consistent with previous targeted genomic studies and recent epidemiological data demonstrating no difference in mortality for patients given active vs inactive empirical treatment except for isolates with an MIC >32/2 mg/L[166].

In summary, here I present the largest GWAS between patient outcome and pathogen genetics in *E. coli*. I demonstrate that the fimbrial adhesin protein *papGII* is associated with reduced 30-day mortality, even after adjusting for patient factors. Whilst my study is c. 3.7x larger than the recent comparable study of Denamur and colleagues[158], it is still far smaller than the log-order increase in magnitude predicted to be required to power the detection of pathogen genetic features associated with patient outcome, suggesting that pooling my dataset with others that have high quality linked clinical metadata would be a useful exercise which may uncover further hits, particularly those with lower heritability.

5. The Mobilome associated with Gram-negative bloodstream infections: A large-scale observational hybrid sequencing based study

Note: This chapter has previously been published as a pre-print[167] and the methods development section as a peer-reviewed manuscript[168]. I authored the text and figures below (which are reproduced here largely unaltered from the published work) with input from supervisors commensurate with the amount of input/advice that would be considered appropriate for a DPhil thesis. All authors have agreed to the inclusion of this published work in my thesis. Sections 5.2.1 and 5.3.5 also describe development and use of the Flanker[169] tool which was produced in work performed in parallel to this thesis in collaboration with Will Matlock.

5.1 Introduction

As in previous chapters, Gram-negative bloodstream infections (BSI) are associated with substantial morbidity and mortality; their incidence continues to increase both in the UK and globally[62,63]. Multidrug-resistant and hypervirulent phenotypes are a particular concern, especially since genes conferring these characteristics (and others which may have either positive or negative fitness effects) are carried on plasmids, frequently in association with other smaller mobile genetic elements[6,66]. Plasmids are thought to facilitate the rapid dissemination of these genes within and between species. A detailed understanding of their biology and epidemiology is therefore likely to be critical in tackling the global threat of antimicrobial resistance (AMR).

Complete and accurate genome assemblies, such as those produced by "hybrid" assemblies of short and long-read sequencing data, are crucial for the study of plasmid epidemiology. Until recently however, these have been prohibitively expensive for use in large genomic epidemiology studies. Thus whilst this approach has recently been used at scale to evaluate the plasmidome of environmental/agricultural isolates[127], to my knowledge its application to human-associated isolates has been mostly restricted to relatively small numbers of isolates selected based on AMR phenotype[6,170]. This phenotype driven selection strategy has identified several plasmid types associated with the dissemination of key ARGs but it is currently not known whether similar plasmids are also found in susceptible populations. Recently, two studies have demonstrated the utility of network-based approaches to classify plasmid assemblies from public databases, offering insights into the host range of these plasmids, though such analyses suffer from sampling bias as well as a lack of clinical context and accurate metadata[171,172]. Therefore, the plasmidome associated with Gram-negative isolates causing both antimicrobial susceptible and resistant clinical infections remains largely uncharacterised.

In this study, I generated a large collection of complete *E. coli*/*Klebsiella* spp. genomes from all BSIs collected in 2009 and 2018 in Oxfordshire, UK, as well as a representative sample from intervening years. Using this unprecedented dataset, I first sought to investigate the extent to which plasmids are shared and contribute to overlaps in the pangenome within and between species, and then compared plasmids associated with ARG carriage to those that are not. Subsequently I investigated the dissemination dynamics of the most prevalent ESBL gene in the population, *bla*$_{CTX\text{-}M\text{-}15}$, highlighting complex nested mobilisation that can only be unravelled using hybrid assembly. Finally I contextualised my findings by comparing my plasmid dataset to a large global collection and investigated whether features of "successful" plasmids and those with the potential for ARG carriage are predictable.

5.2 Methods

5.2.1 Methods development

As mentioned above, use of long read sequencing with Oxford Nanopore flowcells has been historically limited by the associated cost. I overcame this challenge by utilising flowcell wash kits and the validation and optimisation of this step formed part of the preparatory work performed in the course of this thesis which has now been published in Microbial Genomics[168]. In brief I demonstrated that after 24 hours of sequencing, most assemblies were complete and there was no benefit associated with longer sequencing times. Inter-library contamination (resulting from using the same barcodes for closely related samples) did not appear to affect subsequent hybrid assemblies. Sequencing three libraries of 12 isolates on a single flowcell reduces the cost from ~£52.20/isolate to ~£38.20/isolate. This initial work allowed me to greatly increase the number of isolates selected for hybrid sequencing. Since I completed this work, my finding that uneven and relatively low coverage Nanopore long-reads can often provide good quality assemblies has been replicated[173] and the introduction of the new 96-plex rapid barcoding kit means that this approach can likely be optimised even more in the future.

Another challenge was finding a method to analyse gene flanking regions (the genetic sequence up- and downstream of genes of interest). I was not able to identify a suitable existing method to extract these regions from FASTA files and subsequently cluster them into groups based on sequence homology. In conjunction with Will Matlock (another DPhil student in the group), we therefore wrote a new python script which we called “Flanker”[169]. Flanker uses Biopython to extract regions (flexibly specified by the user) from FASTA files around genes identified by Abricate. It additionally handles strand differences and circular contigs. Users can opt to write the flanking regions +/- gene of interest to a new FASTA

file for a range of windows (e.g. 100/200/300bp upstream, downstream or in both directions from the gene). Finally, Flanker provides functionality to cluster these flanking regions based on single-linkage of Mash similarity. In this chapter I have used Flanker to explore the genetic context of *bla*$_{CTX\text{-}M\text{-}15}$ (though I have modified the clustering algorithm slightly to alleviate the need to choose an arbitrary window size) and in work performed in parallel to this thesis, not included here, Will and I have demonstrated its application to *bla*$_{KPC\text{-}2/3}$ and *bla*$_{OXA\text{-}48}$ gene flanking regions to show geographical partitioning and the effect of different promotors on antimicrobial susceptibility phenotype, respectively[169].

5.2.2 Isolate selection

In chapter 2, I reported analyses of short read sequencing data from *E. coli* and *Klebsiella* spp. bloodstream infection isolates in Oxfordshire between 2009 and 2018[68,129]. In this study, I additionally sequenced all *E. coli* and *Klebsiella* spp. isolates from 2009 and 2018 using Oxford Nanopore Technologies using the flowcell washing protocol described above. I also sequenced a subset of isolates from intervening years, selected using stratified random sampling based on analysis of short-read data to capture maximum plasmidome diversity. I also selected isolates from clinically important local AMR-associated outbreaks and representatives of other Gram-negative species cultured from BSIs with apparently similar plasmidomes in the same setting. Details of successfully sequenced isolates, those excluded and the stratification and selection methods are shown in Figure 24.

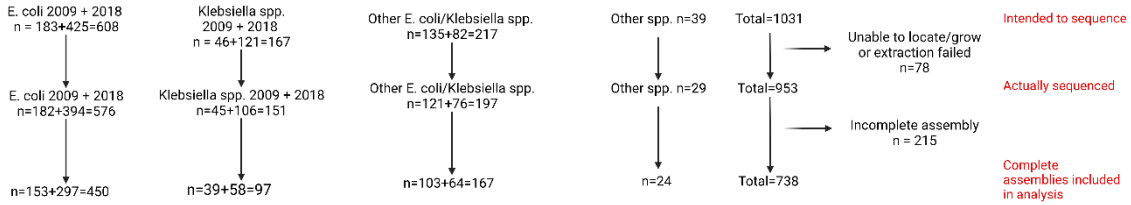


Figure 24 – Details of isolates included/excluded from the chapter 5 study

5.2.3 Sequencing

DNA for long-read sequencing was extracted either using Qiagen Genomic Tip/100G according to the manufacturer's instructions, or with the BioMerieux Easymag using the manufacturer's generic short protocol with a final elution volume of 50µL. The Qubit 2.0 Flourometer was used to quantify DNA. Sequencing libraries were prepared using the Oxford Nanopore Technologies Native (n=23) and Rapid (all other) barcoding kits, according to the manufacturer's instructions. Sequencing was performed on ONT GridIons with R9.4 flowcells, which were reused multiple times utilising the ONT Flow Cell Wash kit and our previously validated protocol[168]. Illumina (short-read) data was created as previously described[68].

5.2.4 Bioinformatics

Reads were first base-called and demultiplexed using Guppy (v3.1.5, Oxford Nanopore Technologies) with Deepbinner[174] (v0.2.0) subsequently used to recover additional unclassified reads as previously described[168]. My strategy for hybrid assembly is depicted in Figure 25. I first assembled all isolates using Unicycler[175] (–mode bold) with the raw Illumina and ONT reads as input. In parallel I performed another assembly where

Unicycler was given an assembly graph from Flye[176] (run with –plasmids –meta and reads which had been polished using Ratatosk[177]) and short reads pre-processed by Shovill[81]. The most contiguous assembly of these was used (or the latter if both were complete). If neither hybrid assembly completed then I used the Flye assembly (with four subsequent rounds of Pilon[178] polishing) if this was complete. Incomplete assemblies (where ≥1 replicon [i.e. either plasmids or the chromosome] had >1 contig) were excluded from further analysis (n=215).

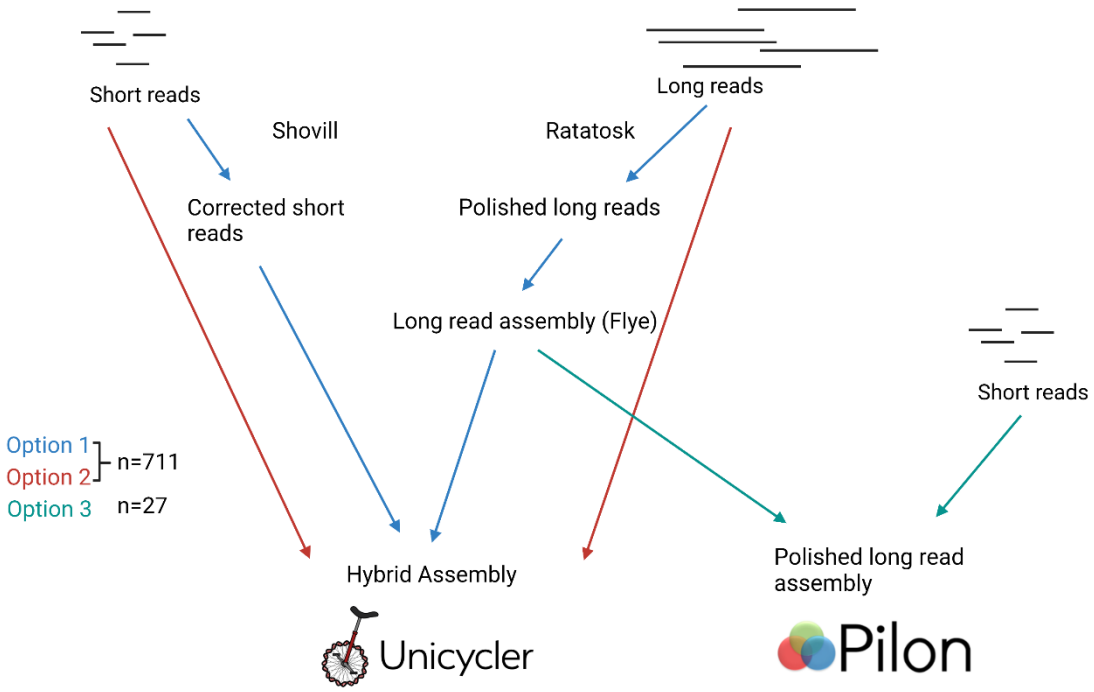


Figure 25 – Assembly methods used

Rarefaction analysis was performed using the R library Micropan[179]. Annotation of genes was performed using AMRFinder Plus[159], ABRicate[87], TADB 2.0[180] and Prokka[82]; custom/manually augmented databases (available at: www.github.com/samlipworth/GN_BSI_Hybrid) were used for the latter two to attempt to improve the proportion of annotatable toxin-antitoxin systems/plasmid associated genes respectively. GC content and predicted mobility were extracted from Mobsuite output. GC-

gap was defined as *GC content of plasmid – GC content of chromosome*. Pangenomes were analysed using Panaroo[98] (v1.2.8 – clean-mode sensitive) and visualised with a Umap projection created using the R package Umap[181]. Variance in the pangenome explained by e.g. AMR content/year/species was examined using a permanova perfomed in the R package vegan[107]. Gene flanking regions were analysed using Flanker (v1.0, –w 0 –wstop 5000 –wstep 100)[169]. The Reder package[182] was used to cluster a weighted graph created from a matrix in which distances were determined to be the greatest distance from the gene (in both upstream and downstream directions) in pairs of isolates which were in the same Flanker cluster; this analysis was repeated in an all vs all fashion for all isolates. Flanking regions were annotated using the Galileo AMR software (Arc Bio, Cambridge MA USA). Copy number was estimated using Unicycler.

5.2.5 Plasmid clustering

Robust taxonomic classification of plasmids remains a challenge[183]. I therefore used two established methods that have been applied to large-scale short read sequencing datasets, Replicon typing using PlasmidFinder[86] and relaxase typing with MOB-suite[184]. I also typed all plasmids using the recently described Plasmid Taxonomic Unit nomenclature[172] (using COPLA[185]). As a substantial number of plasmids remained unclassified by all these methods, I additionally utilized a recently described graph-based classification system[186]. Mash(v2.3)[92] (-s 1000, -k 21) was used to create an all vs. all distance matrix of plasmid assemblies where the distance was taken to be 1 – the proportion of shared kmers between the plasmid of interest and plasmids in the sketch sequences, where plasmids with a distance of 0 share all kmers in the sketch space whereas those with a score of 1 share no common kmers. This was used to create a weighted graph using the R package Igraph[110] where vertices represent plasmids and edges between these are weighted by the distance described above. Community

detection on this graph was performed using the Louvain algorithm which seeks to maximise the density of edges within vs between communities. I optimised performance of this algorithm as described previously[186] by sparsifying the graph, removing edges with a weight ≤ a threshold which was selected by iteration. This approach performed optimally (i.e. assigned the maximum number of isolates to larger [n≥10 isolate] clusters) when the graph was sparsified at an edge weight of ≤0.551 prior to community detection (Figure 26); this parameter was used for all subsequent analysis. The final sparsification threshold was selected to optimise the number of plasmids assigned to large (n≥10) clusters. I compared the classifications given by this approach to other methods using the Normalized Mutual Information index in the R package NMI[187], which demonstrated good agreement with previously described classification methods using normalised mutual information (NMI; see Methods): replicon-typing NMI=0.81, relaxase-typing NMI=0.93, plasmid taxonomic unit (PTU) NMI=0.81.

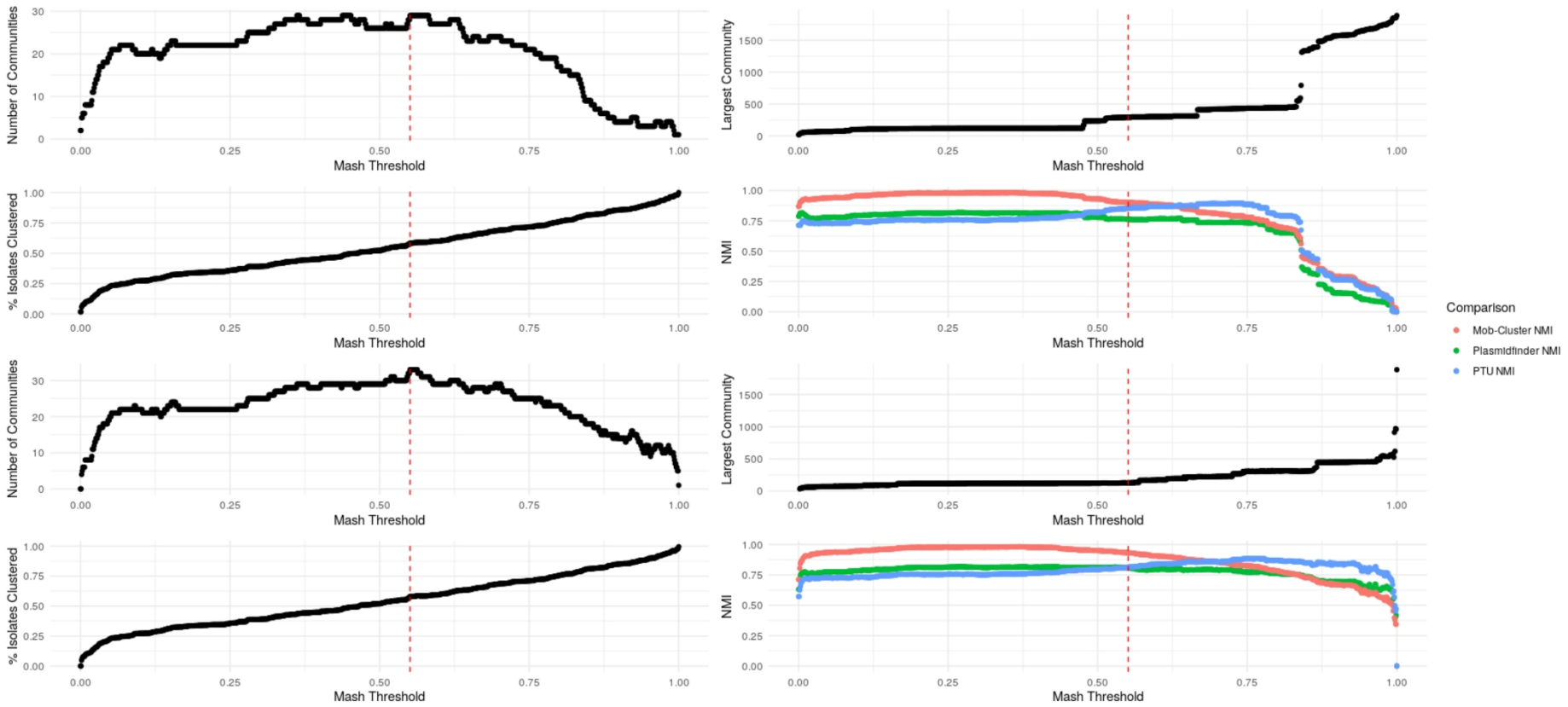


Figure 26 - Optimisation of threshold and method using to assign plasmids to plasmid groups.

The top two rows show the results of a single linkage methodology whereas the bottom two show results obtained using the Louvain clustering algorithm used in my analysis. X-axis shows the Mash distance used to sparsify the group. Number of communities represents the number of plasmid groups n≥10, whereas the largest community represents the largest plasmid group formed at a given threshold. NMI – Normalised mutual information, PTU – plasmid taxonomic unit. The red hashed line shows the Mash distance threshold of 0.551 which was used in the final analysis.

5.2.6 Comparison with existing plasmid sequencing data

To place my plasmid sequencing data in a global context, I downloaded all available plasmids (n=10,159) from a recently curated plasmid collection[171] for comparison. I refer to the Oxfordshire isolates as the “Oxfordshire dataset” and the combined collection as the “Global dataset”. I computed a pairwise distance matrix and performed Louvain-based clustering as described above, sparsifying the graph using the same threshold (0.551) as in the main analysis.

5.2.7 Statistical analysis

To investigate factors associated with geographical dissemination of plasmid groups I considered only Oxfordshire isolates and those from NCBI not from the UK and where the location of isolation was known. I further included only plasmid clusters observed at least once in Oxfordshire. Isolation in more than one country was used as the binary dependent variable in a logistic regression with other plasmid group features (e.g. ARG/virulence/GC content) as independent variables. Multivariable associations between all available plasmid group metrics (independent variables) and plasmid group frequency in the dataset (dependent variable) were estimated using Poisson regression in exploratory analyses. Comparisons of continuous variables and proportions between groups used Kruskal-Wallis/Wilcoxon rank-sum and Fisher/chi-squared tests respectively in R version 4.1[113].

To search for non-AMR plasmid-borne genes associated with carriage of ARGs, I performed logistic regression with membership of an ARG-associated plasmid group as the dependent variable and each gene in the plasmid pangenome as the independent

variable, adjusting for population structure using multi-dimensional scaling (MDS) of mash distances (R package CMD scale), represented in 10 dimensions[188]. I additionally adjusted for plasmid size using three categories ('large' ≥100,000bp, medium ≥10,000-<100,000bp and small <10,000bp). P-values were adjusted for multiple comparisons using the Bonferroni method after removing genes with <1% population frequency (n=11474 comparisons). This pangenome-wide association study was performed using only the Oxfordshire plasmid dataset.

I then tested the predictive value of the genes (n=178) identified in the above logistic regression analysis to identify plasmids not carrying ARGs (ARG-negative plasmids) which were found in ARG-associated groups in the Oxfordshire dataset (n=1447 plasmids of which 609 where in ARG-associated groups). I hypothesised that such ARG-negative plasmids might have the potential to readily acquire ARGs and therefore represent a potential risk. I evaluated the performance of nine models (logistic regression, linear discriminant analysis, K neighbours classifer, decision tree classifier, gaussian naive Bayes, random forest classifier and gradient boosting classifier and a voting classifier combining all of these) using 10-fold cross validation, which was repeated 100 times. I further evaluated the performance of the best performing model (random forest) using ARG-negative *E. coli* and *Klebsiella* spp plasmids in the global dataset with the same features as before (significant gene hits from the pangenome GWAS above). I split the dataset into plasmids collected prior to 2018 (global training set on which the model was re-trained; n=656 plasmids of which 221 where in ARG associated groups) and those collected subsequently (held-out global testing set on which final metrics were reported; n=306 plasmids of which 99 where in ARG-associated groups). This analysis was performed using the SciKitLearn[189] package in Python version 3.7.7.

5.2.8 Data visualisation

Data were visualised using the ggplot2[160] and gggenes[190] packages in R, Clinker[191], Cytoscape[192] and Biorender (www.biorender.com).

5.3 Results

I successfully sequenced and assembled n=738 isolates of which 75% (553/738) were *E. coli* (n=153, 297, 103 in 2009, 2018, intervening years, respectively), 22% (161/738) *Klebsiella* spp. (n=39, 58, 64 in 2009, 2018, intervening years, respectively) and 3% (24/168) other Enterobacterales spp. (details in Figure 24).

In total, these 738 isolates carried 1,880 complete plasmids with a median of 2 plasmids per isolate (interquartile range (IQR) 1-3). 10% (77/738) isolates carried no plasmids, 29% (211/738) carried one and 61% (450/738) more than one (Figure 27A). Of the n=661 isolates with at least one plasmid, 77% (508/661) carried at least one large plasmid (i.e., sequence length >100,000bp), and 94% 621/661) at least one large or medium plasmid (i.e., sequence length >10,000bp); of these 53% (329/621) also carried at least one small plasmid (i.e. sequence length <10,000bp). Carriage of one or more small plasmids in the absence of any medium or large plasmid was relatively rare at (6% [40/661]). Rarefaction analysis suggested that a substantial number of plasmid groups remain unsampled and that there is a significantly greater diversity amongst groups containing smaller (<100,000bp) vs larger (≥100,000bp) plasmids (Figure 27B). There was some evidence that *Klebsiella* spp. isolates tended to carry slightly more plasmids than *E. coli*: median 2 (IQR 1-5) vs (median 2 (1-3) plasmids respectively (Kruskal-Wallis, p-value=0.03; Figure 27A), as did multi-drug resistant (MDR i.e., carriage of ≥3 ARG classes) vs. non-MDR

isolates: (n=317/738 vs. n=421/738 isolates; median 3 (IQR 2-4) vs. median 2 (1-3) Kruskall-Wallis, p-value<0.001]).

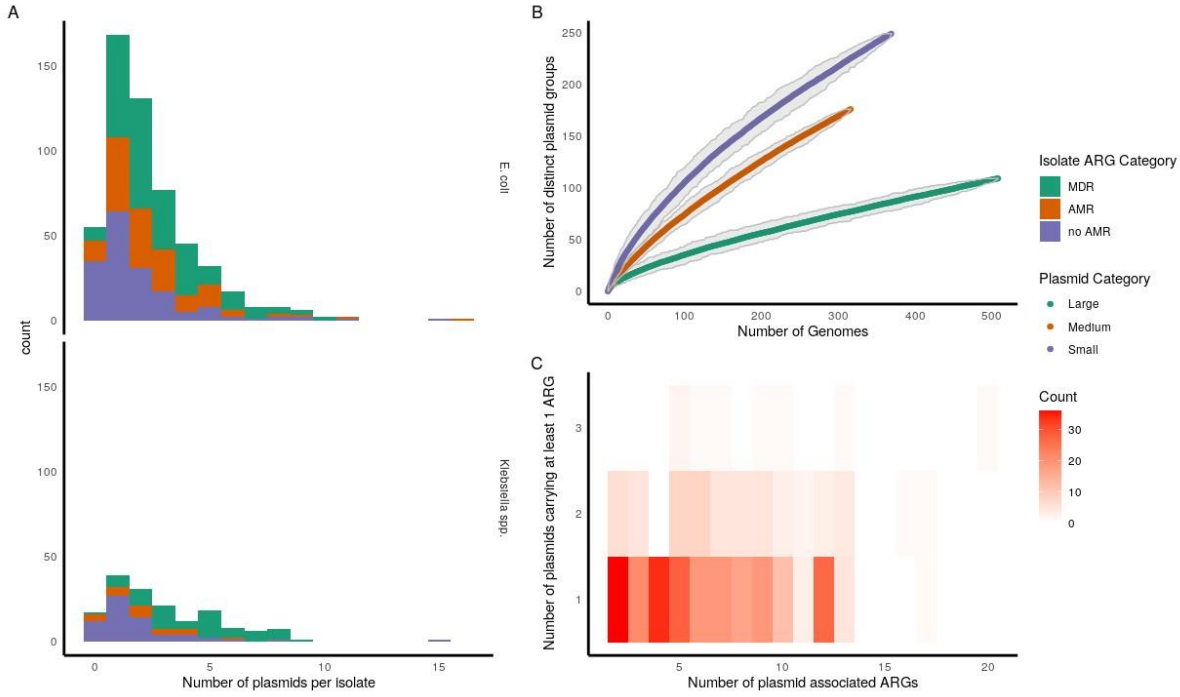


Figure 27 – Distribution of plasmids by species and ARG carriage
A – number of plasmids per isolate for *E. coli* (top panel) and Klebsiella spp. (bottom panel), coloured by the number of ARG classes per isolates where MDR is ≥3 and AMR 1-2. B - Rarefaction curve of the number of novel plasmid groups per new plasmid sequenced stratified by size (large ≥100,000bp, medium ≥10,000 <100,000bp, small < 10,000bp. C – Number of plasmid-associated ARGs per isolate vs number of plasmids carrying at least one ARG. Isolates with only one plasmid associated ARG (by definition carried on one plasmid) are excluded.

Despite comprising a relatively small proportion of the total genome (median=2.79%, IQR=1.97-3.97%), plasmids carried 39% (2069/5311) ARGs, 12% (987/8315) virulence genes and 60% (2836/4735) stress response genes. 50% (368/738) isolates carried at least one plasmid-borne ARG and 306 at least 2; of these, 79% (242/306) carried all annotated ARGs on a single plasmid (Figure 27C). In isolates with a medium or large plasmid, co-carriage of a small plasmid was significantly more common in isolates harbouring plasmid-borne ARGs (58% (210/361) vs. 46% (119/260) without, Fisher's exact p-value=0.003).

5.3.1 Most BSI isolates carry a large (>100,000bp) plasmid from a small number of common plasmid groups

I first attempted to classify plasmids using existing tools; 17% (317/1880) plasmids could not be assigned a replicon type and 33% (622/1880) had no identifiable relaxase type. Similarly 26% (487/1880) plasmids were not typable using the recently described Plasmid Taxonomic Unit (PTU) scheme[185]; 7% (128/1880) were not typable by any method. Subsequently, I therefore opted to use a previously described classification approach, utilizing a graph-based Louvain community detection algorithm[186] (see methods) which has the advantage of not being reliant on reference databases for group assignation and is thus able to classify all plasmids into groups. This approach yielded 513 groups from 1880 plasmids, of which 164 (32%) contained >1 plasmid, but only 33 (6%) contained ≥10 plasmids, and most were singletons (349/513 (68%)). 322/553 (58%) *E. coli* isolates carried a plasmid from one of the four most common, predominantly *E. coli*-associated, large (≥100,000bp) plasmid groups (4/6/7/8 in Figure 28) and similarly 76/161 (47%) *Klebsiella* spp. isolates contained a plasmid from one of the three most common, predominantly *Klebsiella* spp. associated, large plasmid groups (1/2/5 in Figure 28).

5.3.2 Plasmid groups are structured by host phylogeny but there is evidence of intra- and inter-species transfer events

I found strong evidence that the plasmidome of BSI isolates was structured by host phylogeny, although there was also vast and persistent background diversity. Overall, 141/513 (27%) groups were found in >1 MLST and 22/513 (4%) were found in more than one species; multi-species groups had ≥10 members significantly more commonly (8/22 (36%) vs 25/491 (5%), p<0.001) (Figure 28). Sequence type and host species explained

8% and 7% (Adonis p=0.001 for both) of the observed plasmidome variance respectively. ARG content explained a comparatively small amount of variance (R^2=2%, p=0.001) as did year of isolation (0.03%, p=0.005) and source attribution (R^2=1.2%, p=0.99, i.e. suspected focus of infection, only available for a small subset of isolates [198/738]) (Figure 29, panels a, b, c and d respectively).

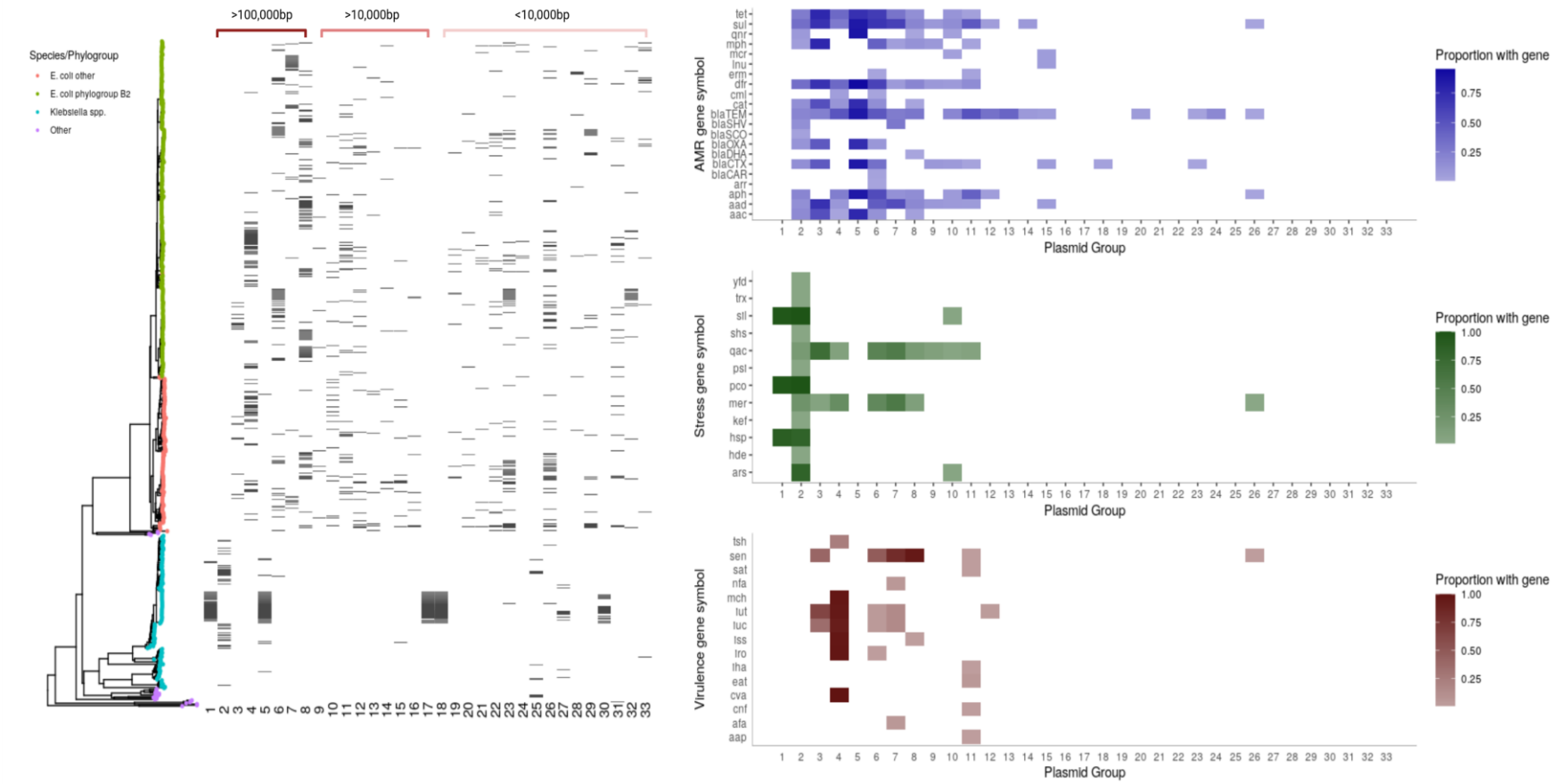


Figure 28 - Phylogenetic distribution of the most common (n≥10 members) plasmid groups (n=33 groups) and the gene content of these

The tree is a neighbour joining tree built on Mash distances between chromosomes. Tip colours represent species/phylogroup. The black bars represent the presence or absence of plasmid groups (shown along the bottom x-axis) for each isolate in the tree. The right panel shows the percentage of isolates within each of these 33 plasmid groups carrying the genes indicated (darker colours denote higher proportion of isolates carrying gene). To improve readability, gene groups have been clustered together.

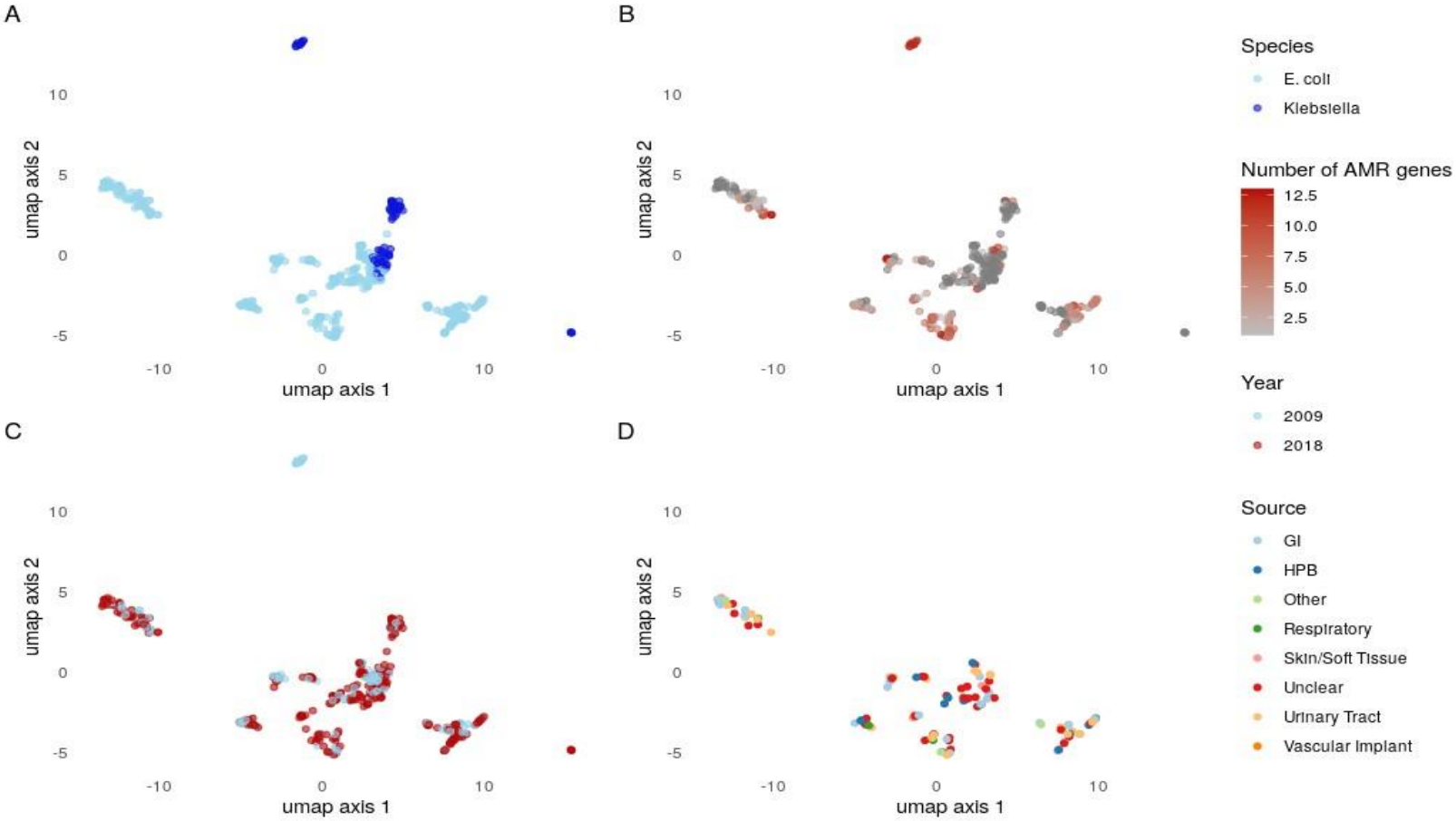


Figure 29 – Plasmidomes are constrained by species but not ARG content, year of isolation or infection source (portal of entry).

A Umap projection of the mash distances between the plasmidomes of isolates (each point represents the plasmidome, i.e. all plasmid sequences of a single isolate), coloured to show the variability explained by species (A)/ARG carriage (B)/year (C) and infection source (D).

5.3.3 Common plasmid groups share genes with each other; gene sharing with chromosomes is also frequent

Whilst I observed that only 4% (22/5143) plasmid groups were shared between species, I hypothesised that this might greatly under-represent the true extent of plasmid-mediated horizontal gene transfer given the role of smaller mobile genetic elements and the fact that BSIs represent a tiny fraction of the overall ecological landscape. I therefore investigated evidence of overlap in the pangenome between different plasmid groups as well as between these and host chromosomes. Most genes in the pangenomes of common (i.e. containing n≥10 plasmids) plasmid groups of *E. coli* and *Klebsiella* spp. were non-unique to their group (median % non-unique genes 88%, IQR 67-98%). Most overlap occurred amongst genes found in the plasmid pangenome from the same species (median 86% (IQR 50-95%) shared genes vs 31% (8-43%) from different species, $p<0.001$). There was also substantial overlap between plasmid group pangenomes and the chromosome pangenome, although there was some evidence of convergence in the chromosomally integrated mobilome between species, evidenced by less difference in the proportion of genes shared with the chromosome for same vs different species (Figure 30, median 33% (IQR 0-45%) vs 21% (0-35%), respectively $p=0.34$).

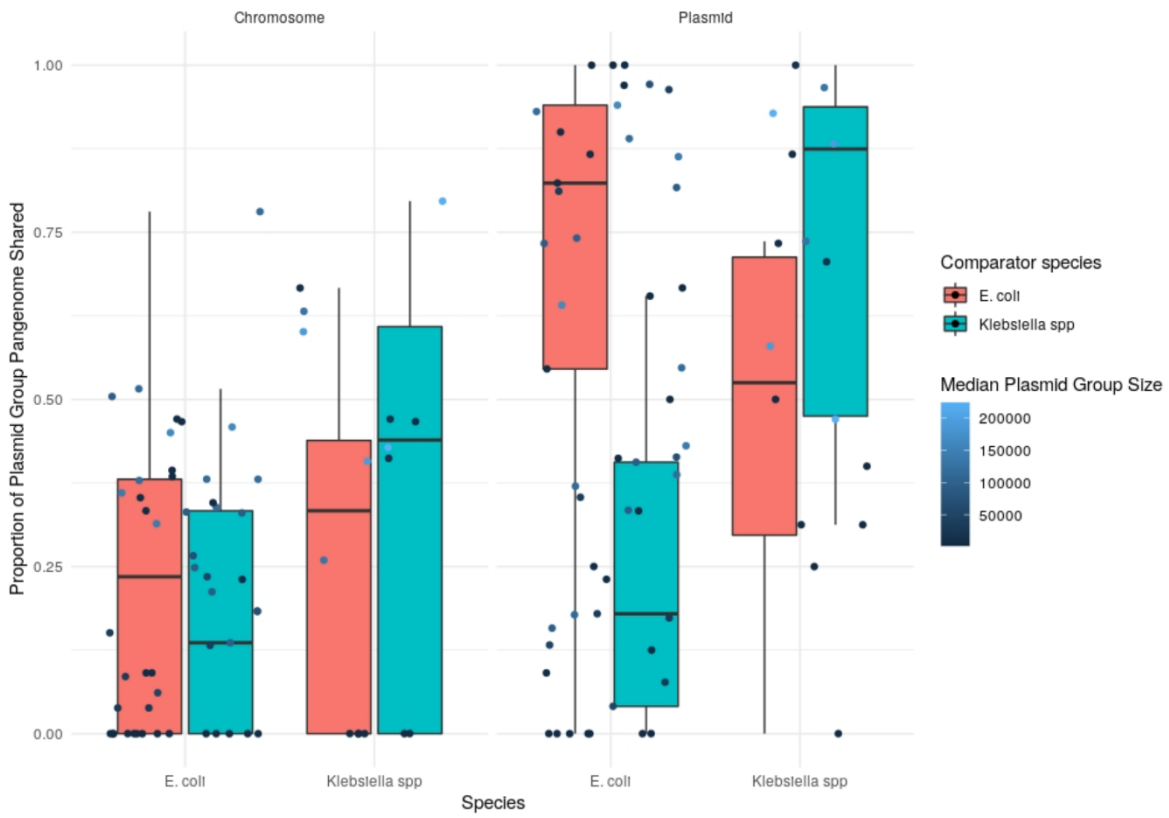


Figure 30 - Overlap between plasmid and chromosomal pangenomes.
Each dot represents a single plasmid group (only the largest [n≥10] plasmid groups are shown). The x-axis denotes the predominant host species in which the plasmid group was isolated and the fill colours denote the comparator species. The y-axis position denotes the proportion of genes in the pangenome (as annotated by Prokka/Panaroo) shared between genetic contexts (i.e. plasmid versus chromosome) as indicated. The colour of the dots shows the median plasmid group size.

5.3.4 Plasmids associated with ARG carriage are often highly similar to those with no such genes

The 439 plasmids carrying at least one ARG were predominantly large (≥100,000bp, 277/439, 63%), low copy number (median 1.80 IQR 1.63-2.37) and conjugative (347/439, 79%). Most plasmid-borne ARGs were carried by plasmids clustering in a small number of groups (81% 1674/2069 ARGs were carried by 8 plasmid groups), but 36% (170/474) plasmids in these groups did not carry an ARG and all groups had at least one such member, highlighting that acquisition of ARGs in ARG-negative plasmid backbones represents a common risk across genetically divergent plasmid groups (Figure 31). I

repeated this analysis using group assignations given by COPLA (Plasmid Taxonomic Units) and Plasmidfinder (replicon typing) and found similar results (Table 12/Table 13), suggesting that this finding is robust to the choice of clustering method.

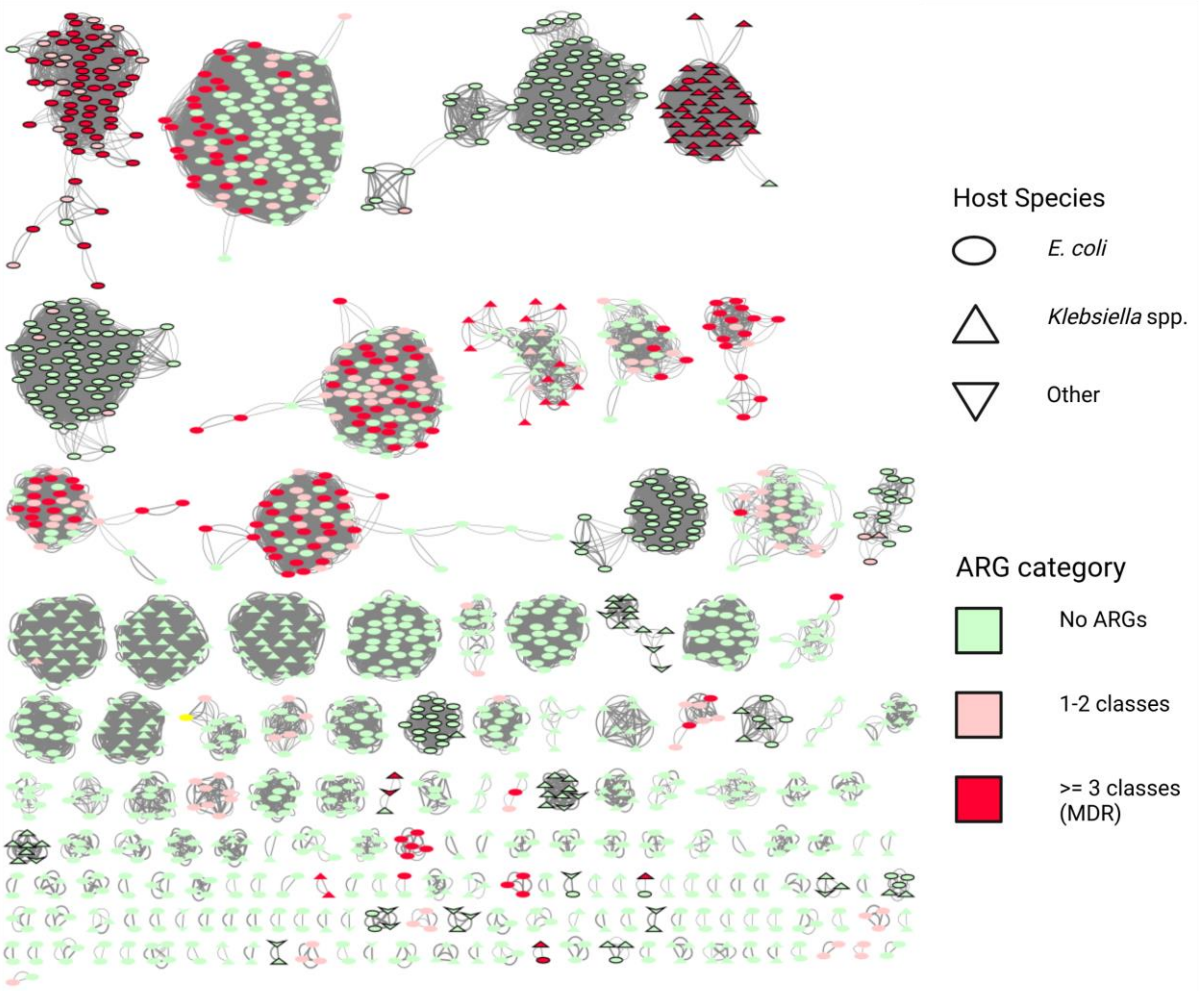


Figure 31 – Antibiotic resistance gene (ARG) carrying plasmids are not a distinct entity
Plasmid network where plasmids (nodes) are connected by edges if they cluster in the same group using the Louvain-based methodology and coloured according to the number of classes of ARGs that they carry. Edge thickness is drawn proportional to the Jaccard distance (see methods) between plasmids. Multi-species clusters are donated by black outlined shapes. Only plasmids groups with ≥2 members are shown.

5.3.5 Hybrid assembly reveals complex nested diversity associated with key AMR genes, significant chromosomal integration of ARGs and presence of multiple copies in different contexts

Chromosomal integration of ARGs was common: for example, in *E. coli*, 56% (23/41) *bla*$_{CTX\text{-}M\text{-}15}$, 9% (2/22) *bla*$_{CTX\text{-}M\text{-}27}$, 14% (42/293) *bla*$_{TEM\text{-}1}$, 42% (14/33) *bla*$_{OXA\text{-}1}$, 39% (7/18) *aac(3)-IIa* and 5% (3/65) *dfrA17* were chromosomally integrated. There was significantly more chromosomal integration of ARGs also seen at least once in a plasmid in my study in *E. coli* vs *Klebsiella* spp. (restricting to 2009 and 2018 only 15% [324/2103] vs 8% 39/478 [8%], chi-squared p<0.001). For *E. coli*, there was also significantly more chromosomal integration in 2018 vs 2009 (19% [285/1485] vs 6% [39/618], chi-squared p<0.001) but there was no evidence of this for *Klebsiella* spp. (7% [13/190] vs 6% [17/279], chi-squared p=0.89). For most of these ARGs, there were multiple instances of isolates carrying two (and occasionally more) copies (9 such examples for *bla*$_{CTX\text{-}M\text{-}15}$ (Figure 35), 1 *bla*$_{CTX\text{-}M\text{-}27}$, 29 *bla*$_{TEM\text{-}1}$, 2 *aac(3)-IIe* and 1 *dfrA7*).

Given the global importance of the ESBL gene *bla*$_{CTX\text{-}M\text{-}15}$ conferring third generation cephalosporin resistance, I focused on its genetic background and putative dissemination mechanisms. As above, plasmid groups carrying this gene in my dataset were generally species-constrained. However, within a single species, considering phylogroup, sequence type and even plasmid group, *bla*$_{CTX\text{-}M\text{-}15}$ was found in a variety of genetic contexts (Figure 35). For example, in *E. coli* ST131 it was found in five plasmid groups and was chromosomally integrated in 41% (17/41) isolates. Within ST131 sub-clades, there was some evidence of vertical transmission, as well as numerous independent integration events. In many cases, several unique gene flanking regions were found in association with *bla*$_{CTX\text{-}M\text{-}15}$ within a single plasmid group, or identical flanking regions were shared

across plasmid groups and between plasmid groups and chromosomes. Visual inspection of gene flanking regions and hierarchical clustering of a weighted graph (as described in methods) suggested that whilst there was substantial diversity, these flanking regions appear to have evolved in a stepwise manner with bilateral association of *bla*$_{CTX\text{-}M\text{-}15}$ and Tn*2* in flanking groups 2, 3 and 6 compared to the presence of Kpn*14* (groups 1 and 5) and IS*26* (group 4) (Figure 32). Inspection of core-genome phylogenies of the two largest *bla*$_{CTX\text{-}M\text{-}15}$ carrying plasmid groups (plasmid groups 2 and 3 in Figure 28) demonstrated multiple probable independent horizontal acquisition events of this gene (and other ARG cassettes as shown in Figure 33/Figure 34), suggesting that a flexible capacity to acquire ARGs through diverse mobile genetic elements, rather than a fixed association with them, might be important factors for the successful dissemination of the host plasmid.

Table 12 - Counts of plasmids with/without ARGs in the largest PTU groups.

Assigned PTU	Not classified	PTU-Novel	PTU-B/O/K/Z	PTU-FE	PTU-FK	PTU-I1	PTU-X4	PTU-Y
Non-ARG containing	439	40	21	174	24	1	14	0
ARG-containing	48	27	37	206	58	3	11	4

Table 13 - Counts of plasmids with/without ARGs in the largest PlasmidFinder groups.

Assigned Rep-type	Col156	IncFIA	IncFIB(AP001918)	IncFIC(FII)	IncFII	IncFII_pKP91	IncFII(29)_pUTI89	IncFII_pRSB107
Non-ARG-Containing	220	12	154	52	16	37	81	14
ARG-containing	107	97	206	74	47	48	41	45

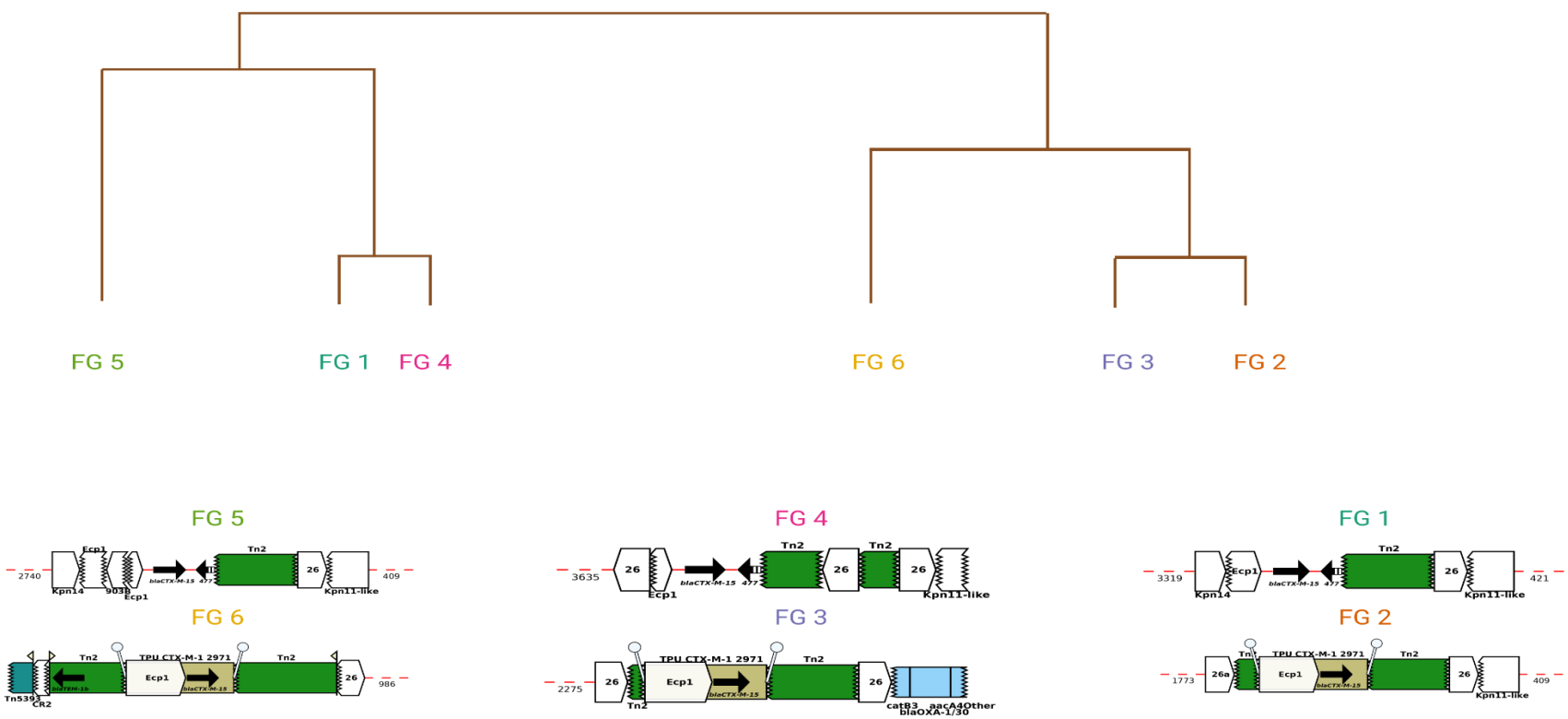


Figure 32 – Similarity of *bla*$_{\text{CTX-M-15}}$ gene flanking regions

The top part of the figure represents hierarchical clustering of a weighted graph formed from a distance matrix where the distance between each pair of plasmids was the maximum window length at which flanker placed them in the same cluster by single linkage. The resulting dendrograph was cut at 0.5 and groups (and nested groups) are shown as identified by RedeR. The lowest levels of these have been renamed sequentially (FG – flank group) and colour coded to match the colours used in Figure 34. The annotations as given by Galileo AMR are shown below for each flank group.

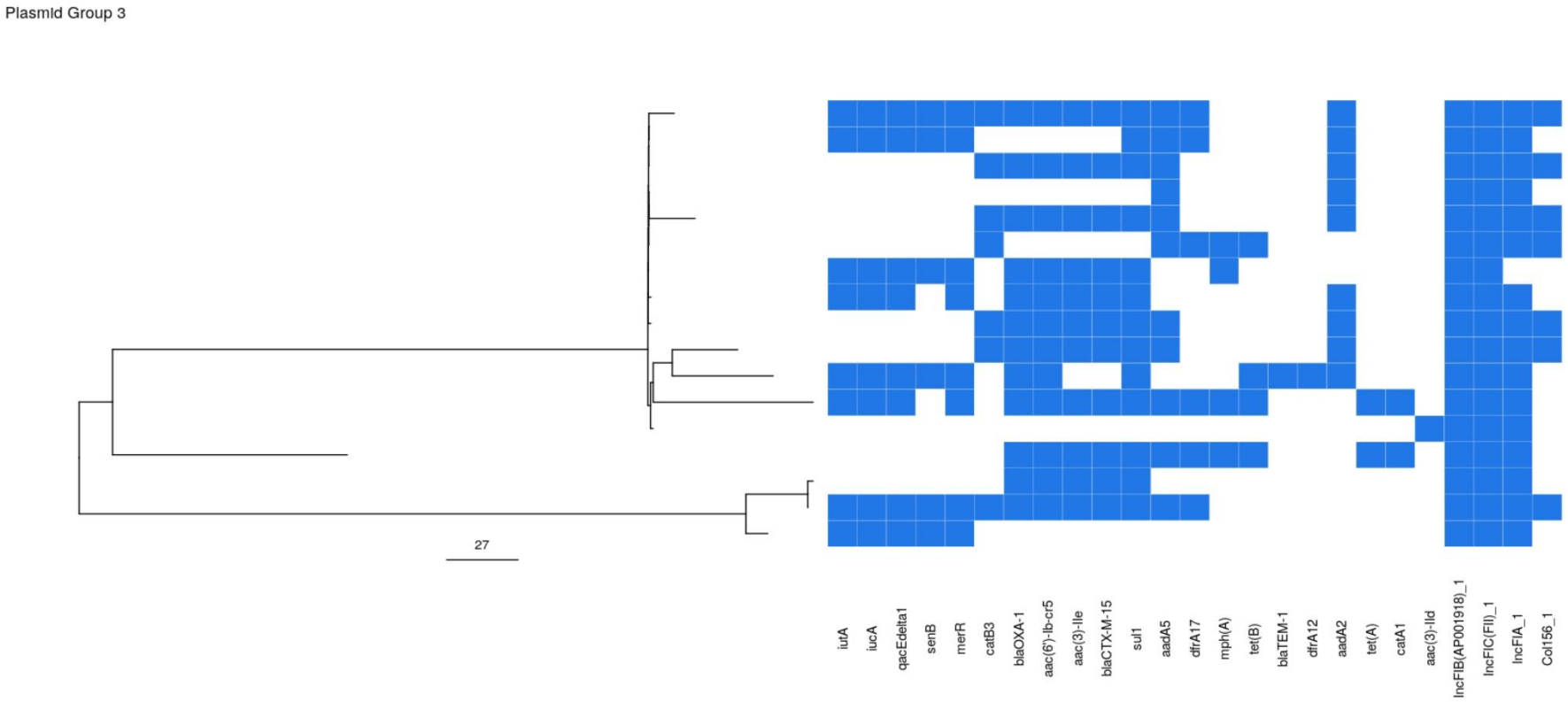


Figure 33 - Core genome phylogeny and gene presence/absence heatmap (for genes in the AMRFinder/PlasmidFinder databases) for my group 3 plasmids.

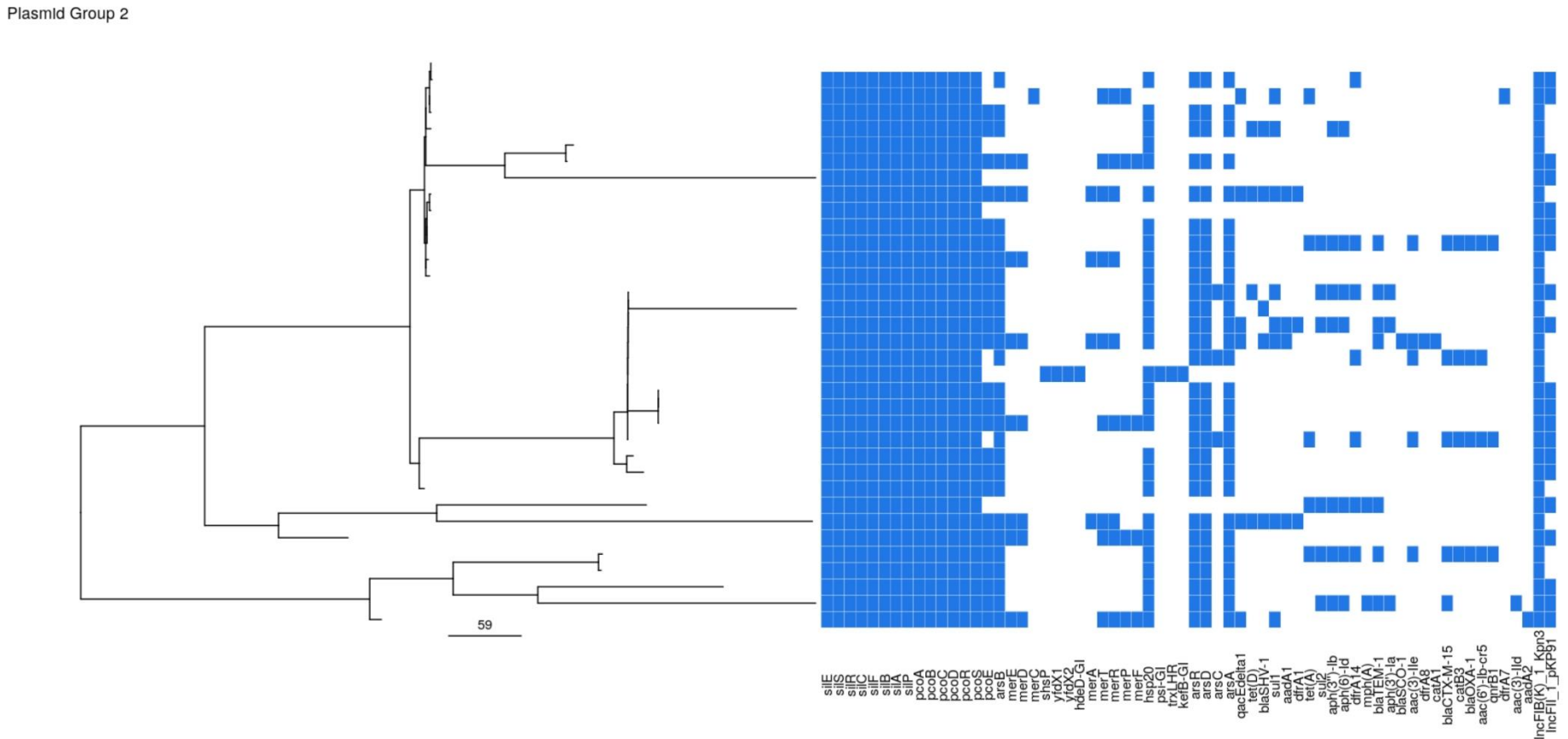


Figure 34 - Core genome phylogeny and gene presence/absence heatmap (for genes in the AMRFinder/PlasmidFinder databases) for my group 2 plasmids.

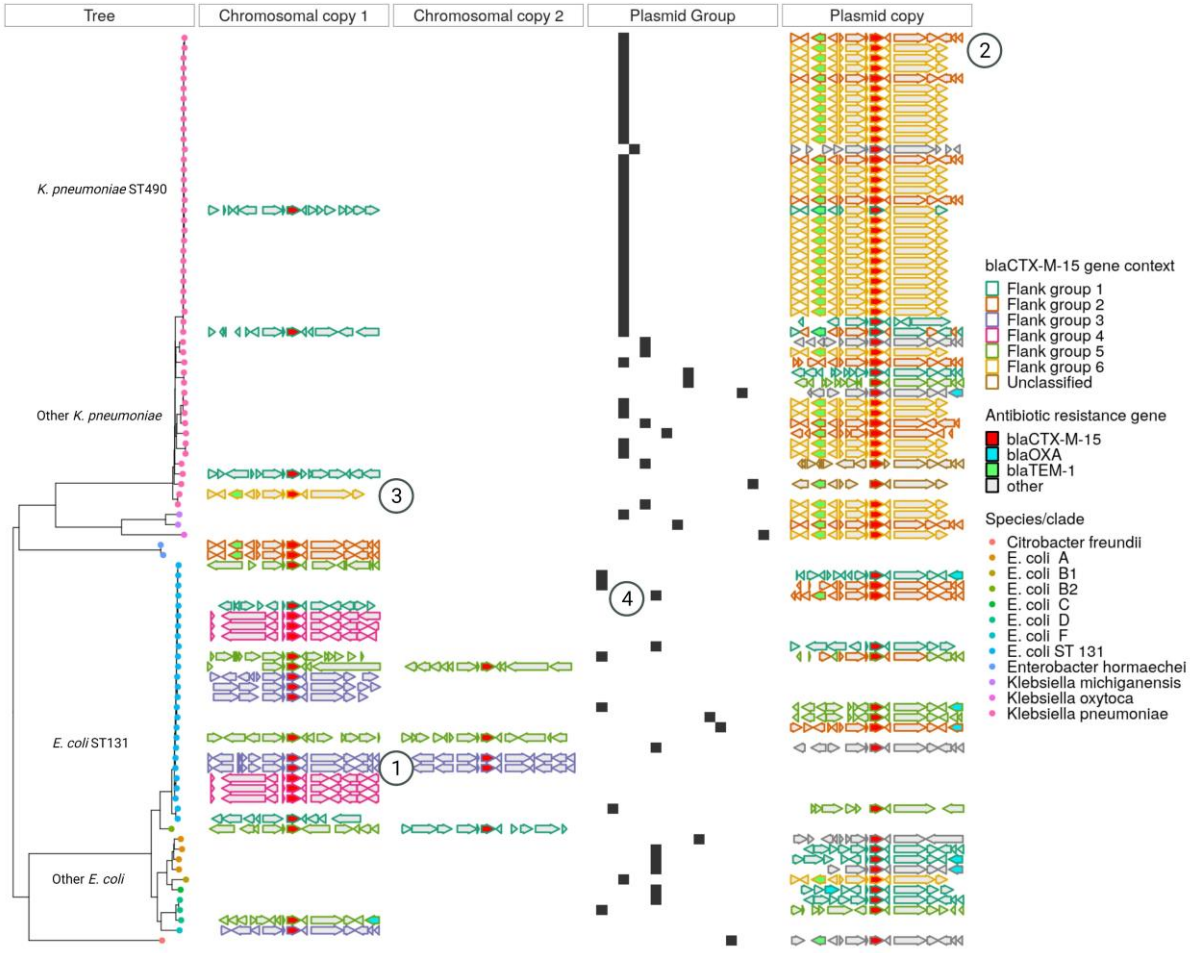


Figure 35 - Nested genetic complexity associated with $bla_{CTX\text{-}M\text{-}15}$ mobilisation.

The 'Tree' panel shows a neighbour-joining tree of Mash distances between chromosomes for isolates carrying $bla_{CTX\text{-}M\text{-}15}$. Tip colours represent species/ST/phylogroup. The chromosomal copy 1 and 2 panels show the genetic context 5000bp up- and downstream from chromosomal copies of *$bla_{CTX\text{-}M\text{-}15}$* (shown in red); the plasmid copy panel shows this equivalent information for isolates carrying a plasmid-borne copy of this gene. The outlining colour in these panels shows the hierarchical cluster assignment of these flanking groups. The plasmid group panel shows group membership of plasmids carrying $bla_{CTX\text{-}M\text{-}15}$ with each x-axis position representing a distinct group and black bars showing the presence or absence of these for isolates in the tree. The encircled numbers denote: 1 - different flanking sequences in the same ST, 2 - different flanking sequences in the same plasmid group, 3 - the same flanking group found in both chromosomal and plasmid contexts and 4 - different plasmid groups harbouring the gene found within the same ST.

5.3.6 Comparison with wider plasmid datasets highlights undersampled plasmid diversity, more widespread interspecies and inter-niche plasmid sharing, and the potential for carbapenemase dissemination amongst “high-risk” plasmid groups

I repeated my graph-based plasmid clustering method on a combined dataset of Oxfordshire plasmids (N=1880, hereby referred to as the ‘Oxfordshire dataset’) and the global collection of plasmids deposited in the NCBI from countries other than the UK (N=10,159, denoted the ‘global dataset’) using the same sparsifying threshold (≤0.551). This yielded 5913 groups of which 484 contained at least one plasmid from the ‘Oxfordshire dataset’; of these, 326 groups (67%) containing 536 plasmids appeared to be unique to Oxfordshire, highlighting significant additional plasmid diversity in this single region not represented in the global database. 79/484 (16%) of groups containing Oxfordshire plasmids were found in more than one species in the full dataset; of these 57 (72%) occurred in only a single species in the Oxfordshire dataset, highlighting the substantial underestimation of wider between-species dissemination by investigating only a single region and single source (i.e. bloodstream infections).

A striking feature of the global network was that plasmids carrying carbapenemase genes clustered with those that did not (Figure 36). Of 122 plasmid groups with at least one member carrying a carbapenemase gene, 19 (16%) contained at least one Oxfordshire plasmid. These included representatives from the *K. pneumoniae* MDR-associated Oxfordshire BSI dataset groups 2 and 5 (Figure 28), three large groups (Figure 28 groups 3/6/8) widely distributed amongst *E. coli* isolates, and two groups of smaller plasmids (<100,000bp, Figure 28 groups 10 and 12), also widely distributed in Oxfordshire *E. coli*. Although only 2% (7/414) Oxfordshire plasmids falling into these groups actually carried a

carbapenemase ARG, this suggests the potential for carbapenemase acquisition and dissemination amongst widespread “high-risk” plasmid backbones.

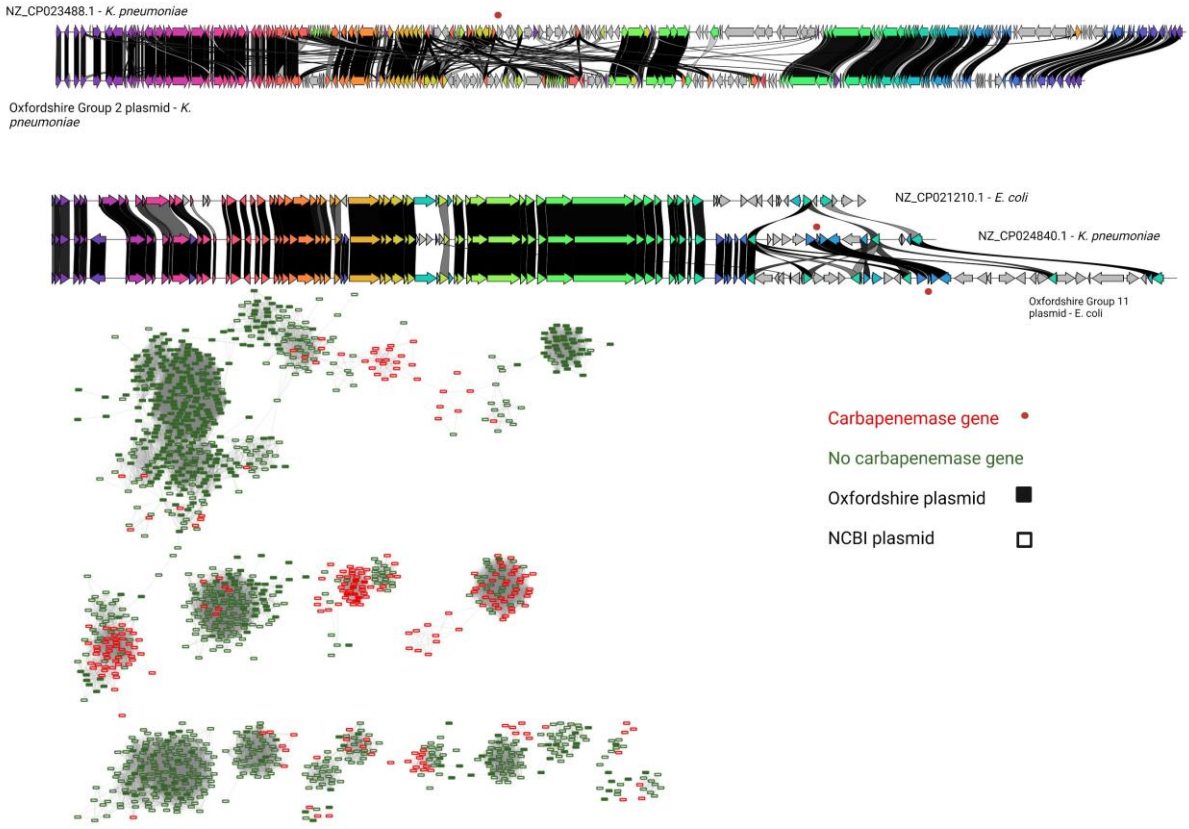


Figure 36 - Plasmids carrying carbapenemase genes are highly similar to plasmids without these genes found in Oxfordshire BSIs.
A) Each horizontal bar represents a plasmid assembly either from Oxfordshire ‘Group 2/11’ or the NCBI global dataset. Common genes are shown in colour (with blast identity between these shown from light grey to black (where the latter represents a perfect nucleotide sequence match) whereas genes unique to a given plasmid are shown in grey. B) A network plot of plasmids which cluster with carbapenemase carrying plasmids in the global network analysis. Plasmids (nodes) are connected with an edge where the edge weight is ≤0.551 (see methods). The thickness of edges is displayed so that it is proportional to the edge weight.

5.3.7 Factors predictive of plasmid group “success”

Having demonstrated that most isolates carry a plasmid from a relatively small number of plasmid groups, I next sought to understand what factors might be driving the widespread dissemination of these amongst BSI isolates. Multivariable Poisson regression analysis revealed that plasmid group frequency (a subjective marker of evolutionary “success”) was associated with isolation in multiple species (adjusted rate ratio aRR 4.89, 95%CI

4.29-5.57, p<0.001), capacity to conjugate (aRR 1.73, 95%CI 1.47-2.04) or mobilise (aRR 1.29, 95%CI 1.13-1.48) (i.e. containing either a relaxase or *oriT* but missing a mate-pair formation marker), carriage of multiple ARGs (aRR 1.23, 95% CI 1.19-1.27) or virulence (aRR 1.44, 95%CI 1.36-1.53) or toxin-antitoxin genes (aRR 1.32, 95%CI 1.18-1.47) and a higher GC content (aRR 1.01 per percentage higher, 95%CI 1.00-1.03) (Table S2). Carriage of ARGs (adjusted Odds Ratio, aOR=2.88, 95%CI 1.53–5.41, p<0.001) and isolation in multiple species (aOR=7.79, 95%CI 3.07-22.90, p<0.001) were independently associated with a higher probability of plasmid groups being observed internationally (Table 14).

Table 14 - Features associated with plasmid groups being unique to Oxfordshire (N=326 groups) or also found in the global plasmid dataset (N=158 groups).
There were 5429 plasmid groups only found in the global plasmid dataset. In this analysis plasmids from NCBI were excluded if they were isolated in the UK or if the country of origin was unknown. Factors with p<0.05 in the multivariable analysis are highlighted in bold.

Variable	Odds Ratio (Univariable)	Lower 95% confidence interval	Upper 95% confidence interval	p value	Odds Ratio (Multivariable)	Lower 95% confidence interval	Upper 95% confidence interval	p value
Antimicrobial resistance gene present (ARG)	3.46	1.98	6.02	<0.001	2.88	1.53	5.41	<0.001
Virulence gene present	1.37	0.56	3.08	0.46	1.63	0.57	4.35	0.34
Stress response gene present	1.77	1.00	3.07	0.04	1.58	0.68	3.66	0.29
Medium vs small size (10,000-99,999bp)	0.92	0.54	1.53	0.75	0.55	0.26	1.11	0.11
Large vs small size (≥ 100,000bp)	1.41	0.82	2.41	0.21	0.61	0.22	1.62	0.34
Conjugative	2.20	1.25	3.90	0.01	1.71	0.84	3.57	0.15
Mobilisable	1.68	0.99	2.91	0.06	1.24	0.69	2.25	0.48
GC content (per 1% higher)	1.06	1.01	1.12	0.01	1.04	0.99	1.10	0.14
Number of species*	1.06	1.01	1.12	0.01	7.79	3.07	22.90	<0.001

*Number of species in which the plasmid was found in the Oxfordshire dataset

5.3.8 Machine learning allows risk stratification of plasmids

Given that I have shown that plasmids carrying ARGs are often very similar to those with no such genes ("ARG-negative plasmids"), I hypothesised that it might be possible to predict whether ARG-negative plasmids pose a risk for eventual association with ARGs. To do this I first performed a genome-wide association study (corrected for population structure and plasmid size, using the Oxfordshire dataset only) to identify genes (excluding known ARGs) significantly more or less likely to be carried by plasmids in ARG-associated groups (i.e. plasmid groups where at least one member carries at least one ARG). This revealed significant associations between ARG-associated plasmid groups and presence of insertion/transposon sequences, various virulence factors, toxin/anti-toxin systems and heavy metal resistance genes (Table 15).

Table 15 - Genes/gene clusters associated with ARG-carrying plasmids.
Univariable logistic regression adjusted for plasmid size (categorical large/medium/small) and population structure (10 dimensions of multidimensional scaling of Mash distances between host chromosomes). P values have been adjusted by the Bonferroni method (n=11474 comparisons); only significant (p<0.05) results are shown for genes annotated as something other than “hypothetical protein”. Annotations are as given by Prokka/Panaroo. Where Panaroo clustered together multiple Prokka annotations, only the first of these is shown. Two plasmids (sizes 1570 and 1308) were excluded from the analysis because they had no coding regions identified. Results are ordered by p value.

Gene	**N in ARG-carrying plasmids**	**N in non-ARG-carrying plasmids**	**Odds Ratio**	**P value**
psiB	437	51	17.52	<0.001
iucD	105	8	21.57	<0.001
insB_1	327	178	18.69	<0.001
shiF	97	8	20.95	<0.001
iucB	97	8	20.95	<0.001
tnpA	272	53	6.47	<0.001
sitD	96	3	53.63	<0.001

Gene	N in ARG-carrying plasmids	N in non-ARG-carrying plasmids	Odds Ratio	P value
sitC	96	3	53.63	<0.001
sitB	95	3	53.61	<0.001
sitA	93	3	53.56	<0.001
ntdC	83	2	91.82	<0.001
orf156_1	331	22	6.28	<0.001
orf157_1	304	21	6.60	<0.001
cvaA	86	2	91.07	<0.001
KF37082	430	53	3.99	<0.001
repA_1	269	56	195.73	<0.001
areA	441	49	8.04	<0.001
hha	32	20	75.45	<0.001
pemK_2	185	4	33.70	<0.001
pemI	185	4	33.70	<0.001
orf36	34	4	23.24	<0.001
yjcC_3	32	2	62.29	<0.001
ycgB_1	452	54	3.62	<0.001
arsH	29	4	24.13	<0.001
orf34	28	4	24.11	<0.001
repI	342	32	5.21	<0.001
yuaT	92	2	59.47	<0.001
orf05	358	93	4.20	<0.001
parM_1	45	3	47.37	<0.001
tnpA	78	4	21.27	<0.001

Gene	N in ARG-carrying plasmids	N in non-ARG-carrying plasmids	Odds Ratio	P value
ydfA	292	24	4.57	<0.001
ydeA	340	53	3.51	<0.001
higA_1	73	13	7.79	<0.001
insE	18	33	0.09	<0.001
yehA	310	26	4.84	<0.001
topIII_1	28	17	32.11	<0.001
higB	73	12	7.73	<0.001
traE_1	28	7	11.70	<0.001
tnpA_3	113	143	0.23	<0.001
etsA	85	11	8.88	0.0012
etsB	85	11	8.88	0.0012
etsC	85	11	8.88	0.0012
mob2	17	83	9.85	0.0013
hnh	10	76	0.12	0.0014
tnpA_3	73	3	38.33	0.0018
TF134_00008	95	1	153.05	0.002
hlyF	86	12	8.03	0.002
celG	3	20	0.04	0.0025
repL	12	8	14.98	0.0027
ant	12	8	14.98	0.0027
topB_1	32	29	34.37	0.004
ppfA	13	8	14.39	0.0046
upfB	13	8	14.39	0.0046

Gene	N in ARG-carrying plasmids	N in non-ARG-carrying plasmids	Odds Ratio	P value
ydcB	463	83	2.97	0.0056
pdcB	58	10	10.48	0.0057
pDCA	58	10	10.48	0.0057
mobA	72	29	9.45	0.0058
doc	58	11	9.53	0.008
ybaA	258	13	5.05	0.0083
arsD	29	12	6.75	0.0084
arsA_2	29	12	6.75	0.0084
cra	13	8	13.09	0.0091
int	379	113	3.97	0.0092
int2	99	25	4.29	0.01
pmgS	13	8	12.52	0.013
hdmD	13	8	12.52	0.013
yacA	130	21	4.60	0.013
toxin_1	9	34	0.11	0.015
traY	124	11	5.98	0.017
arsA	30	13	6.05	0.018
traJ	123	11	5.96	0.018
ydiL	190	10	6.77	0.024
dnaQ	91	7	7.23	0.038
mat	13	8	11.44	0.039
traM_2	511	114	2.94	0.042

I then tested the predictive value of these elements to identify ARG-negative plasmids belonging to ARG-associated groups using a variety of machine learning models on the Oxfordshire dataset (see methods) with stratified 10-fold cross-validation to estimate out of sample performance. The best performing model (random forest) had a mean accuracy of 90.3% (standard deviation [SD] 2.4%), mean area under the receiver operator curve [AUC] 0.90 (SD 0.02), mean sensitivity 86.0% (SD 4.3%) and mean specificity 93.4% (SD 2.9%). I re-trained the random forest model on ARG-negative plasmids in the global dataset using only plasmids sequenced prior to 2018 and subsequently made predictions on the held-out 2018 plasmids. This demonstrated that the model generalised well but was less sensitive on this dataset (accuracy 84.6%, AUC 0.82, sensitivity 73.7% and specificity 89.9%).

5.4 Discussion

In this study, I fully reconstructed 738 isolates (1880 plasmids) to conduct the largest, most unselected and comprehensive evaluation of the epidemiology and function of plasmids associated with Gram-negative isolates causing bloodstream infections to date. Most isolates in this study carried a large plasmid from a small number of plasmid groups; these were frequently, but not invariably, associated with carriage of multiple antibiotic resistance, virulence and heavy metal resistance genes, potentially providing survival and fitness benefits to the host bacterium. The fact that most isolates with multiple plasmid-borne ARGs (often from several different classes) carried all of these on a single plasmid reinforces the importance of good antimicrobial stewardship and avoiding unnecessary exposure to all classes of antibiotics to control co-selection as much as possible. Crucially I also found that plasmids carrying ARGs frequently clustered in large, widely

disseminated groups including plasmids without these genes, representing a potential set of “high-risk” backbones for ARG acquisition and horizontal spread.

To date, most similar sequencing studies have focused on plasmids carrying particular ARGs (particularly those with ESBL/carbapenem resistance genes) and have therefore not considered how these might be related to plasmids without such genes. I hypothesise that plasmid adaptation to coexist with successful lineages often occurs prior to the acquisition of high-risk ARGs, presenting a potential window of opportunity for intervention which is lost if one is solely focused on the presence of these genes. My findings should therefore motivate a shift away from studies focusing on a single phenotype or gene of interest towards efforts to identify and track high-risk plasmid groups or other smaller high-risk mobile genetic elements relevant to ARG dissemination. My work clearly illustrates that such a new surveillance framework must incorporate unselected sampling frames, i.e. not only selecting isolates with particular AMR phenotypes for sequencing. Notably I found an association between small plasmids and medium/large ARG-associated plasmids, suggesting that they may play an important helper role in ARG plasmid persistence/spread, and a more detailed understanding of this possible synergy could be valuable[193].

Whilst plasmid populations were structured, and plasmid groups were mostly constrained to a single species and in some cases species lineages in Oxfordshire, there was also clear evidence of exchange between lineages of a species and different species. Enterobacterales are widely distributed as commensals and in multiple environmental sources; my study sample is thus extremely sparse relative to the whole ecology. Even where I found no evidence that certain plasmids were shared between species in Oxfordshire, these data demonstrated widespread sharing of the plasmid gene repertoire

(including ARGs and their flanking regions) with plasmids and chromosomes in other species. The unexpectedly high proportion of isolates with chromosomally integrated ARGs (and apparent increase across the study period for *E. coli*) may either represent a success of plasmids in conferring survival benefits to their host whilst lowering their own associated fitness cost, or a success of the bacterial host by lowering its dependence on the presence of the plasmid.

A limitation of this study is that it is from a single region, mitigated in part by comparisons with large, publicly available datasets. The lower sensitivity for predicting ARG-group association in the global dataset likely reflects its inherent bias and heterogeneity compared to the Oxfordshire dataset as well as the existence of ARG-associated groups and genes not observed in the model training data from my setting. The inability to sequence and/or assemble all plasmids from the selected cohort is an additional limitation. Sequencing only bloodstream infection isolates may lead to underestimation of how much sharing of plasmids between species truly occurs, given that this represents a highly selected subset of isolates causing severe disease. This is supported by my analysis of the large publicly available dataset, which demonstrated that several groups found only in a single species in my study have previously been seen in other species. My results also highlight the substantial limitations of previous studies using reference database-based approaches for plasmid typing and demonstrate that fully reconstructed genomes (i.e. utilising long-read sequencing data) are essential in order to provide detailed insight. The fact that Illumina and ONT sequencing was performed on different extractions is not ideal and may mean that plasmids sequenced in one extract are subsequently lost during laboratory work (or vice versa).

In conclusion, my study provides the first high-resolution description of the plasmidome associated with *E. coli*/*Klebsiella* spp. bloodstream infections and demonstrates that using long-read data and unselected sampling frames is essential in order to fully appreciate its complexity. Previous studies of plasmid epidemiology in Gram-negatives have primarily focused on MDR/carbapenemase-carrying isolates; my finding that non-ARG carrying plasmids are often highly similar to plasmids isolated in these earlier studies demonstrates the potential for rapid dissemination of ARGs to settings where they are currently rare. I therefore recommend that surveillance is based on unselected sampling frames, long-read sequencing and considers plasmids and smaller mobile genetic elements to develop a representative understanding of the horizontal gene transfer landscape relevant to AMR and to facilitate appropriate intervention.

6. Conclusions and future work

In this thesis I have combined large whole genome sequencing and healthcare informatics datasets to investigate factors driving the increasing incidence of GNBSIs in England. In chapter two, I found evidence of a stable overall population structure in *E. coli*, with continued dominance of the globally successful “major” STs (131/95/73/69). Consistent with previous epidemiological data, I found that an increasing proportion of isolates carry clinically important ARGs; however there appeared to be no substantial change in virulence gene profile.

I was able to refute the hypothesis that *E. coli* has undergone adaptation to the healthcare setting and there was no evidence of geographical clustering which might be consistent with localised transmission. Similarly, I explored whether transmission in nursing homes might be at least partly responsible for incidence increases but found minimal evidence of this. An increasing focus on catheter care is a possible strategy for preventing some GNBSIs and in my analysis I identified a decreasing incidence of infections arising from the urinary tract, possibly representing some success in this area. The increasing proportion of infections of unknown source is concerning, and might reflect the increasing complexity of multi-comorbid patients seen on the acute medical take (or sub-optimal reporting/coding practices) but further compounds the difficulties of identifying potential interventions.

Whilst there does seem to be some association between the increasing incidence of *E. coli* BSIs and ARG carriage, it clearly does not represent the whole story as demonstrated by the success of the relatively drug-susceptible STs 95/73. Indeed, the trends observed

in this work are set against overall reductions in antimicrobial prescribing in primary care, although more detailed work linking individual level data might be useful to identify targeted areas for further antimicrobial stewardship. It would be very interesting to compare the data in this study with similar large international sequencing-based unselected datasets both in other high and lower-middle as well as low income countries. To my knowledge no such datasets exist and whilst there are significant associated challenges it is likely that equitable access to high quality sequencing based surveillance would provide mutual population health benefits.

Prior to the work carried out in this thesis, *Klebsiella* spp. infections had been characterised dichotomously into those caused by largely opportunistic, low virulence and MDR isolates associated with nosocomial infection versus the generally drug-susceptible isolates causing “clinically hypervirulent” disease (previously defined as severe community onset invasive disease in relatively healthy patients). Analysis of this unselected cohort demonstrates that at best this existing dogma is an oversimplification and that existing markers of clinical hypervirulence are specific but have extremely poor sensitivity. More work is required to understand pathogen factors associated with the so-called clinically hypervirulent phenotype, particularly given the convergence and subsequent transmission of known ARG/virulence factors on plasmids which has been seen elsewhere. In stark contrast to my observations in *E. coli*, the most common *Klebsiella* clone in Oxfordshire BSIs (the MDR ST490) represented only a minority of isolates, declined in incidence over the study period and was rarely observed in later years. The majority of infections were caused by sporadic STs rather than dominant clones as seen in *E. coli*. Similarly, the overall proportion of ARG-carrying *Klebsiella* spp. isolates declined for several key antimicrobial classes. The reasons for these decreases are far from clear but might reflect different reservoirs and niches inhabited by these species. Furthermore, these findings might indicate that different strategies might be

required to reduce the incidence of infections caused by *Klebsiella* spp. vs *E. coli*. This study is however limited by the inherent over-representation of the predominantly white, older and affluent population that lives in Oxfordshire (which is itself only a fairly small region of England).

For both *E. coli* and *Klebsiella* spp., identification of candidate interventions remains a significant challenge and it is likely that incidence reductions from any single measure will be marginal. One possible alternative/complementary strategy might be the use of a vaccine. For *E. coli*, there is a candidate currently in clinical trials. In chapter 3, I therefore investigated whether GNBSIs might in theory be a vaccine-preventable disease. Consistent with the highly clonal population structure observed in chapter 2, I found that the investigational quadrivalent/decavalent vaccines (targeted at O antigens, a lipopolysaccharide component) for *E. coli* would be predicted to provide immunity against a high proportion of isolates, both overall as well as many of those carrying ESBL/carbapenemase-producing genes. Despite the more sporadic population structure observed in *Klebsiella* spp. (and the fact that there is currently no comparable investigational product), I also predict that such a strategy might be effective in this species. Whilst the incidence of *Klebsiella* is much lower than *E. coli*, targeted use of a vaccine might be a useful tool in particular high-risk settings and/or to combat the transmission of high-risk ARGs/plasmids. I eagerly await the results of ongoing early-stage clinical trials for *E. coli* vaccines and further work to assess the real-world efficacy of these. Modelling a cost-effective roll-out strategy and monitoring for subsequent serotype-switching (a process where bacteria change their serotype via genetic recombination which can result in vaccine escape[194]) will be important should these trials succeed.

One possible hypothesis for the increasing incidence of GNBSIs is the emergence of a new virulence factor that might allow isolates that would usually be carried as low-virulence commensals to become invasive. Whilst I did to some extent explore this possibility in chapter 2, that analysis was restricted to known genes in existing databases. *E. coli* has a very large and open pangenome with a vast array of associated mobile genetic elements which provide opportunities for horizontal gene transfer to further enrich/diversify this. In chapter 4 I therefore conducted a GWAS to screen for novel virulence genes in *E. coli*. Previous reports had suggested that most variance in patient outcomes in *E. coli* BSIs is explained by patient factors (particularly age, comorbidities and portal of entry where a urinary source is protective). In chapter 2, I demonstrated a strong association between the *papGII* locus and a urinary source of infection, a finding which has since been replicated twice on independent datasets. In this chapter I unified these previous findings, and demonstrated for the first time an independent protective effect of the *papGII* locus on patient outcome even after adjusting for host factors. It is not clear whether there is a direct causal mechanism underlying this association (e.g. provocation of a protective immune response) or whether a urinary portal of entry generally permits better/quicker source control favouring better outcomes. Rather nicely, my findings in this chapter also validate recent simulation work[158] which predicted that larger sample sizes might begin to provide the power needed to uncover pathogen factors associated with patient outcome. The authors predicted that sample sizes a log order greater in magnitude compared to their study (i.e. N ~10,000) might provide the power to allow further associations to be discovered. Further work in this area should therefore concentrate on combining large datasets from the several longitudinal cohorts which now exist and ensuring that these are linked to the necessary clinical metadata. Whilst there are tens or even hundreds of thousands of *E. coli* genomes available in public repositories, availability of such linkable meta-data is almost non-existent which is a significant barrier to the progression of research studies such as this one. It may also be fruitful to characterise the diversity observed in individuals who develop invasive infection

by comparing isolates that remain commensal to those that go on to cause invasive disease (similar to studies previously performed in staphyloccal BSIs[195]). Whilst such studies are likely to be challenging to design and resource intensive, work to further define populations at high risk of developing GNBSIs might make them more feasible.

Whilst in chapter 2 I did not find evidence of significant changes in population-level strain dynamics over time, the use of short-read sequencing meant that I was unable to fully characterise the plasmidome of isolates causing BSIs. In chapter 5 I therefore utilised recent technological advances in long read sequencing (including some validation work performed specifically for this thesis) to create an unprecedented dataset of 738 complete, hybrid sequenced, reference quality genomes. Overall there was evidence of high levels of diversity and an "open" plasmidome, and, in keeping with this, a substantial number of plasmids could not be characterised using existing schemes. I demonstrated that most isolates with multiple plasmid-borne ARGs carry these on a single plasmid, representing the challenge posed by co-selection and emphasising the importance of antimicrobial stewardship for all classes of antibiotic. Knowledge of the genetic environment of ARGs has particular importance for controlling local outbreaks where temporary shifts in prescribing (e.g. avoiding cephalosporins in ESBL outbreaks) might be futile if other ARGs active against replacement drugs are also carried on the same mobile genetic element. In addition, I identified that most ARG-carrying plasmids can be classified into a relatively small number of groups which are largely, but not completely, species restricted. Crucially however, I found that these groups also contain a substantial number of plasmids with no ARGs. Similarly, whilst ARGs conferring carbapenem resistance are currently rare in the Oxfordshire setting, comparison with a large global dataset (N=10,159 plasmid sequences) demonstrated that some plasmids in the Oxfordshire collection are highly similar to those known to have been closely associated with the dissemination of these genes internationally. I therefore hypothesise that the

machinery for their rapid dissemination already exists in Oxfordshire and that this has important consequences for future surveillance efforts.

Most large-scale sequencing-based epidemiology efforts to date have focussed on certain ARGs and their association with particular species lineages, termed “high-risk” clonal groups. In light of my findings, I extend this and introduce the concept of high-risk plasmid groups. I demonstrate that the features of these are predictable and propose that future global sequencing initiatives should focus on their identification and surveillance. Early detection of “high-risk” plasmid groups may afford opportunities to reduce selection pressures and their potential for rapid dissemination of e.g. ESBL/carbapenem resistance conferring genes following introduction events. It might even be useful to extend this concept further and explore other high risk flanking regions/mobile genetic elements. This is particularly pertinent given that I also demonstrate increasing chromosomal integration of ARGs over time and so there may be a window of opportunity to prevent them becoming established in bacterial populations. These findings suggest that studies selecting isolates based on a single phenotype/ARG are limited, and attempts should be made (e.g. via wider sampling or as a minimum by comparison with existing datasets) to consider their findings in a wider ecological context.

As an example of the potential value of agnostic plasmid surveillance, whilst previous studies (using strain level analysis) have largely discounted the importance of one health for gram negative infection in humans[196,197], using data from this thesis we have recently demonstrated evidence of highly genomically similar plasmids found across human and environmental niches[198]. This could have policy repercussions were future studies to demonstrate transmission of “high-risk” plasmid groups between human and animal niches. Further efforts to characterise the plasmidomes of susceptible isolates and

understand the factors driving the dissemination of successful plasmid groups not currently associated with ARGs are likely to be useful. Such efforts might for example identify that successful plasmid groups are more likely acquire one or more genes (e.g. ARGs/heavy metal resistance/toxin-antitoxin or other as yet uncharacterised genes) early in their emergence, potentially providing opportunities to reduce selection pressures. Once high-risk plasmid groups have been identified and the relevant epidemiological exposures identified (via linkage of granular metadata), we then have a chance to trial interventions to disrupt their dissemination over populations.

Finally, I also found that many isolates carry small plasmids (and these often co-occur with ARG plasmids). These have been almost completely neglected in most existing studies because they tend not to carry targets of interest (e.g. ARGs/virulence genes) and yet are highly prevalent and frequently co-occur with large plasmids in my dataset. It could be that they are purely parasites with a very low fitness cost or that there is some synergistic benefit that they confer to the host, or the stability of the plasmidome within the host. Whilst there is still much to learn about the evolution, ecology and life-cycle of all types of plasmids, it would seem prudent not to ignore these in future studies.

In conclusion, in this thesis I have demonstrated how linked large-scale sequencing and healthcare informatics datasets can be used to test hypotheses about the possible mechanisms driving the changing incidence of important infectious diseases. Common themes arising from this work are the importance of large datasets with high-quality linked clinical metadata. Whilst there has been an exponential rise in the availability of sequencing data, the poor availability of linkable high-quality metadata significantly inhibits its usefulness. Work to provide harmonised and secure mechanisms for sharing the latter are crucial if whole genome sequencing is to realise its full potential. Another

common theme is the essential importance of unselected datasets (i.e. not focussing on a single phenotype/sequence type etc). In this thesis there are multiple examples where my large and unselected longitudinal cohort has been able to falsify existing dogma. As sequencing costs continue to fall, careful selection of an unbiased sampling frame should be an important consideration for future studies. Whilst this thesis has failed to identify a single factor that might be amenable to intervention (e.g. infection control in nursing homes/hospitals), this is also an important contribution of molecular epidemiology because it ensures that precious resources can be prioritised in the most effective way possible and future research efforts focussed elsewhere.

Bibliography

1. Ueber die isolirte Färbung der Schizomyceten in Schnitt- und Trockenpräparaten von Dr. Gram, Kopenhagen. — Fortschritte der Medicin 1884 No. 6. Ref. Dr. Becker. *Dtsch. Med. Wochenschr.* **10**, 234–235 (1884).

2. Rogers, D. E. The changing pattern of life-threatening microbial disease. *N. Engl. J. Med.* **261**, 677–683 (1959).

3. Magiorakos, A.-P. *et al.* Multidrug-resistant, extensively drug-resistant and pandrug-resistant bacteria: an international expert proposal for interim standard definitions for acquired resistance. *Clin. Microbiol. Infect.* **18**, 268–281 (2012).

4. Stoesser, N. *et al.* Evolutionary History of the Global Emergence of the Escherichia coli Epidemic Clone ST131. *MBio* **7**, e02162 (2016).

5. Wyres, K. L., Lam, M. M. C. & Holt, K. E. Population genomics of Klebsiella pneumoniae. *Nat. Rev. Microbiol.* **18**, 344–359 (2020).

6. David, S. *et al.* Integrated chromosomal and plasmid sequence analyses reveal diverse modes of carbapenemase gene spread among Klebsiella pneumoniae. *Proc. Natl. Acad. Sci. U. S. A.* (2020) doi:10.1073/pnas.2003407117.

7. Wilson, J. *et al.* Trends among pathogens reported as causing bacteraemia in England, 2004-2008. *Clin. Microbiol. Infect.* **17**, 451–458 (2011).

8. Reacher, M. H. *et al.* Bacteraemia and antibiotic resistance of its pathogens reported in England and Wales between 1990 and 1998: trend analysis. *BMJ* **320**, 213–216 (2000).

9. Aucken, H. M., Ganner, M., Murchan, S., Cookson, B. D. & Johnson, A. P. A new UK strain of epidemic methicillin-resistant Staphylococcus aureus (EMRSA-17) resistant to multiple antibiotics. *J. Antimicrob. Chemother.* **50**, 171–175 (2002).

10. Dingle, K. E. *et al.* Effects of control interventions on Clostridium difficile infection in England: an observational study. *Lancet Infect. Dis.* **17**, 411–421 (2017).

11. Clostridium Difficile Outbreaks at Stoke Mandeville Ho - Hansard - UK Parliament. https://hansard.parliament.uk/commons/2006-07-24/debates/0607246000043/ClostridiumDifficileOutbreaksAtStokeMandevilleHospitalAndPublicationOfDataOnHealthcare.

12. *Annual publication of epidemiological commentary*. https://www.gov.uk/government/statistics/mrsa-mssa-and-e-coli-bacteraemia-and-c-difficile-infection-annual-epidemiological-commentary (2021).

13. Murray, C. J. L. *et al.* Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *Lancet* **399**, 629–655 (2022).

14. WHO publishes list of bacteria for which new antibiotics are urgently needed. https://www.who.int/news/item/27-02-2017-who-publishes-list-of-bacteria-for-which-new-antibiotics-are-urgently-needed.

15. *Preventing healthcare associated Gram-negative bloodstream infections: an improvement resource*. https://www.england.nhs.uk/wp-content/uploads/2020/08/Gram-negative_IPCresource_pack.pdf (2017).

16. Coggon, D., Barker, D. & Rose, G. *Epidemiology for the Uninitiated*. (John Wiley & Sons, 2009).

17. Snow, J., Richardson, B. W. & Frost, W. H. *Snow on cholera*. (Commonwealth Fund New York, 1936).

18. Doll, R. & Hill, A. B. Smoking and carcinoma of the lung; preliminary report. *Br. Med. J.* **2**, 739–748 (1950).

19. Peto, R. *et al.* Smoking, smoking cessation, and lung cancer in the UK since 1950: combination of national statistics with two case-control studies. *BMJ* **321**, 323–329 (2000).

20. Kannel, W. B. The Framingham Study: ITS 50-year legacy and future promise. *J. Atheroscler. Thromb.* **6**, 60–66 (2000).

21. CENTERS FOR DISEASE CONTROL. Pneumocystis pneumonia Los Angeles. *MMWR* **30**, 250–252 (1981).

22. Institute of Medicine (US) Committee on a National Strategy for AIDS. *The Centers for Disease Control's Surveillance Definition of AIDS*. (National Academies Press (US), 1986).

23. Watson, J. D. & Crick, F. H. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* **171**, 737–738 (1953).

24. Blattner, F. R. *et al.* The complete genome sequence of Escherichia coli K-12. *Science* **277**, 1453–1462 (1997).

25. Heather, J. M. & Chain, B. The sequence of sequencers: The history of sequencing DNA. *Genomics* **107**, 1–8 (2016).

26. Quick, J. *et al.* Real-time, portable genome sequencing for Ebola surveillance. *Nature* **530**, 228–232 (2016).

27. Seemann, T. Assembling NGS data.

28. Sheppard, A. E. *et al.* Nested Russian doll-like genetic mobility drives rapid dissemination of the carbapenem resistance gene blaKPC. *Antimicrob. Agents Chemother.* **60**, 3767–3778 (2016).

29. Principles of Epidemiology: Home. https://www.cdc.gov/csels/dsepd/ss1978/index.html (2021).

30. Viral Hemorrhagic Fever (VHF) 2011 Case Definition. https://ndc.services.cdc.gov/case-definitions/viral-hemorrhagic-fever-2011/ (2021).

31. van Ingen, J. *et al.* Global outbreak of severe Mycobacterium chimaera disease after cardiac surgery: a molecular epidemiological study. *Lancet Infect. Dis.* (2017) doi:10.1016/S1473-3099(17)30324-9.

32. Clermont, O., Christenson, J. K., Denamur, E. & Gordon, D. M. The Clermont Escherichia coli phylo-typing method revisited: improvement of specificity and detection of new phylo-groups. *Environ. Microbiol. Rep.* **5**, 58–65 (2013).

33. Lees, J. A. *et al.* Fast and flexible bacterial genomic epidemiology with PopPUNK. *Genome Res.* **29**, 304–316 (2019).

34. Tonkin-Hill, G., Lees, J. A., Bentley, S. D., Frost, S. D. W. & Corander, J. Fast hierarchical Bayesian analysis of population structure. *Nucleic Acids Res.* **47**, 5539–5549 (2019).

35. Didelot, X., Kendall, M., Xu, Y., White, P. J. & McCarthy, N. Genomic epidemiology analysis of infectious disease outbreaks using TransPhylo. *Curr Protoc* **1**, e60 (2021).

36. Etymologia: Escherichia coli. *Emerg. Infect. Dis.* **21**, 1310 (2015).

37. Denamur, E., Clermont, O., Bonacorsi, S. & Gordon, D. The population genetics of pathogenic Escherichia coli. *Nat. Rev. Microbiol.* **19**, 37–54 (2021).

38. Gordon, D. M. & Cowling, A. The distribution and genetic structure of Escherichia coli in Australian vertebrates: host and geographic effects. *Microbiology* **149**, 3575–3586 (2003).

39. Clermont, O. *et al.* Characterization and rapid identification of phylogroup G in Escherichia coli, a lineage with high virulence and antibiotic resistance potential. *Environ. Microbiol.* **21**, 3107–3117 (2019).

40. Hastak, P. *et al.* Genomic profiling of Escherichia coli isolates from bacteraemia patients: a 3-year cohort study of isolates collected at a Sydney teaching hospital. *Microb Genom* **6**, (2020).

41. Kallonen, T. *et al.* Systematic longitudinal survey of invasive Escherichia coli in England demonstrates a stable population structure only transiently disturbed by the emergence of ST131. *Genome Res.* (2017) doi:10.1101/gr.216606.116.

42. Goswami, C. *et al.* Genetic analysis of invasive Escherichia coli in Scotland reveals determinants of healthcare-associated versus community-acquired infections. *Microb Genom* **4**, (2018).

43. Gladstone, R. A. *et al.* Emergence and dissemination of antimicrobial resistance in Escherichia coli causing bloodstream infections in Norway in 2002–17: a nationwide, longitudinal, microbial population genomic study. *The Lancet Microbe* (2021) doi:10.1016/S2666-5247(21)00031-8.

44. Price, L. B. *et al.* The epidemic of extended-spectrum-β-lactamase-producing Escherichia coli ST131 is driven by a single highly pathogenic subclone, H30-Rx. *MBio* **4**, e00377-13 (2013).

45. Lipworth, S., Peto, T. E., Crook, D. W., Walker, A. S. & Stoesser, N. Antimicrobial resistance genes and clonal success in Escherichia coli isolates causing bloodstream infection. *Lancet Microbe* **2**, e492 (2021).

46. Nobrega, D. *et al.* Spatial distribution of Escherichia coli ST131 C subclades in a centralized Canadian urban region. *J. Antimicrob. Chemother.* **76**, 1135–1139 (2021).

47. Vimont, S. *et al.* The CTX-M-15-producing Escherichia coli clone O25b: H4-ST131 has high intestine colonization and urinary tract infection abilities. *PLoS One* **7**, e46547 (2012).

48. Sarkar, S. *et al.* Intestinal Colonization Traits of Pandemic Multidrug-Resistant Escherichia coli ST131. *J. Infect. Dis.* **218**, 979–990 (2018).

49. Alhashash, F. *et al.* Increase in bacteraemia cases in the East Midlands region of the UK due to MDR Escherichia coli ST73: high levels of genomic and plasmid diversity in causative isolates. *J. Antimicrob. Chemother.* **71**, 339–343 (2016).

50. Manges, A. R. *et al.* Widespread distribution of urinary tract infections caused by a multidrug-resistant Escherichia coli clonal group. *N. Engl. J. Med.* **345**, 1007–1013 (2001).

51. Hertz, F. B. *et al.* “Population structure of drug-susceptible,-resistant and ESBL-producing Escherichia coli from community-acquired urinary tract.” *BMC Microbiol.* **16**, 63 (2016).

52. van Hout, D. *et al.* Extended-spectrum beta-lactamase (ESBL)-producing and non-ESBL-producing Escherichia coli isolates causing bacteremia in the Netherlands (2014 - 2016) differ in clonal distribution, antimicrobial resistance gene and virulence gene content. *PLoS One* **15**, e0227604 (2020).

53. Roer, L. *et al.* WGS-based surveillance of third-generation cephalosporin-resistant Escherichia coli from bloodstream infections in Denmark. *J. Antimicrob. Chemother.* **72**, 1922–1929 (2017).

54. Friedländer, C. Ueber die Schizomyceten bei der acuten fibrösen Pneumonie. *Archiv für pathologische Anatomie und Physiologie und für klinische Medicin* **87**, 319–324 (1882).

55. Bagley, S. T. Habitat association of Klebsiella species. *Infect. Control* **6**, 52–58 (1985).

56. Raffelsberger, N. *et al.* Gastrointestinal carriage of Klebsiella pneumoniae in a general adult population: a cross-sectional study of risk factors and bacterial genomic diversity. *Gut Microbes* **13**, 1939599 (2021).

57. Henson, S. P. *et al.* Molecular epidemiology of Klebsiella pneumoniae invasive infections over a decade at Kilifi County Hospital in Kenya. *Int. J. Med. Microbiol.* **307**, 422–429 (2017).

58. Gorrie, C. L. *et al.* Antimicrobial-Resistant Klebsiella pneumoniae Carriage and Infection in Specialized Geriatric Care Wards Linked to Acquisition in the Referring Hospital. *Clin. Infect. Dis.* **67**, 161–170 (2018).

59. Webster, D. P. *et al.* Impact of a clonal outbreak of extended-spectrum β-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. *J. Antimicrob. Chemother.* **66**, 2126–2135 (2011).

60. Lam, M. M. C. *et al.* A genomic surveillance framework and genotyping tool for Klebsiella pneumoniae and its related species complex. *Nat. Commun.* **12**, 4188 (2021).

61. Lam, M. M. C. *et al.* Convergence of virulence and MDR in a single plasmid vector in MDR Klebsiella pneumoniae ST15. *J. Antimicrob. Chemother.* **74**, 1218–1222 (2019).

62. Vihta, K.-D. *et al.* Trends over time in Escherichia coli bloodstream infections, urinary tract infections, and antibiotic susceptibilities in Oxfordshire, UK, 1998–2016: a study of electronic health records. *Lancet Infect. Dis.* **18**, 1138–1149 (2018).

63. Buetti, N., Atkinson, A., Marschall, J., Kronenberg, A. & the Swiss Centre for Antibiotic Resistance (ANRESIS). Incidence of bloodstream infections: a nationwide surveillance of acute care hospitals in Switzerland 2008–2014. *BMJ Open* **7**, e013665 (2017).

64. De Angelis, G. *et al.* Incidence and antimicrobial resistance trends in bloodstream infections caused by ESKAPE and Escherichia coli at a large teaching hospital in Rome, a 9-year analysis (2007-2015). *Eur. J. Clin. Microbiol. Infect. Dis.* **37**, 1627–1636 (2018).

65. Lam, M. M. C. *et al.* Tracking key virulence loci encoding aerobactin and salmochelin siderophore synthesis in Klebsiella pneumoniae. *Genome Med.* **10**, 77 (2018).

66. Wyres, K. L. *et al.* Genomic surveillance for hypervirulence and multi-drug resistance in invasive Klebsiella pneumoniae from South and Southeast Asia. *Genome Med.* **12**, 11 (2020).

67. Ellington, M. J. *et al.* Contrasting patterns of longitudinal population dynamics and antimicrobial resistance mechanisms in two priority bacterial pathogens over 7 years in a single center. *Genome Biol.* **20**, 184 (2019).

68. Lipworth, S. I. W. *et al.* Ten year longitudinal molecular epidemiology study of Escherichia coli and Klebsiella species bloodstream infections in Oxfordshire, UK. *Genome Med.* (2021).

69. *English Surveillance Programme for Antimicrobial Utilisation and Resistance (ESPAUR)*. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/843129/English_Surveillance_Programme_for_Antimicrobial_Utilisation_and_Resistance_2019.pdf (2019).

70. Public Health England. Escherichia coli (E. coli): guidance, data and analysis. *GOV.UK* https://www.gov.uk/government/collections/escherichia-coli-e-coli-guidance-data-and-analysis (2010).

71. David, S. *et al.* Epidemic of carbapenem-resistant Klebsiella pneumoniae in Europe is driven by nosocomial spread. *Nat Microbiol* **4**, 1919–1929 (2019).

72. Wyres, K. L. *et al.* Distinct evolutionary dynamics of horizontal gene transfer in drug resistant and virulent clones of Klebsiella pneumoniae. *PLoS Genet.* **15**, e1008114 (2019).

73. Blandy, O. *et al.* Factors that impact on the burden of Escherichia coli bacteraemia: multivariable regression analysis of 2011-2015 data from West London. *J. Hosp. Infect.* **101**, 120–128 (2019).

74. Kizny Gordon, A. *et al.* Genomic dynamics of species and mobile genetic elements in a prolonged blaIMP-4-associated carbapenemase outbreak in an Australian hospital. *J. Antimicrob. Chemother.* **75**, 873–882 (2020).

75. Day, M. J. *et al.* Population structure of Escherichia coli causing bacteraemia in the UK and Ireland between 2001 and 2010. *J. Antimicrob. Chemother.* **71**, 2139–2142 (2016).

76. Holt, K. E. *et al.* Genomic analysis of diversity, population structure, virulence, and antimicrobial resistance in Klebsiella pneumoniae, an urgent threat to public health. *Proc. Natl. Acad. Sci. U. S. A.* **112**, E3574-81 (2015).

77. Long, S. W. *et al.* Population Genomic Analysis of 1,777 Extended-Spectrum Beta-Lactamase-Producing Klebsiella pneumoniae Isolates, Houston, Texas: Unexpected Abundance of Clonal Group 307. *MBio* **8**, (2017).

78. SURVEILLANCE REPORT. Surveillance of antimicrobial resistance in Europe 2018.

79. De Maio, N. *et al.* Comparison of long-read sequencing technologies in the hybrid assembly of complex bacterial genomes. *Microb Genom* **5**, (2019).

80. *Guidance on the definition of healthcare associated Gram-negative bloodstream infections*. https://www.england.nhs.uk/wp-content/uploads/2020/08/HCA_BSI_definitions_guidance.pdf.

81. Seemann, T. Shovill. https://github.com/tseemann/shovill.

82. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).

83. Bortolaia, V. *et al.* ResFinder 4.0 for predictions of phenotypes from genotypes. *J. Antimicrob. Chemother.* **75**, 3491–3500 (2020).

84. Liu, B., Zheng, D., Jin, Q., Chen, L. & Yang, J. VFDB 2019: a comparative pathogenomic platform with an interactive web interface. *Nucleic Acids Res.* **47**, D687–D692 (2019).

85. Siguier, P., Perochon, J., Lestrade, L., Mahillon, J. & Chandler, M. ISfinder: the reference centre for bacterial insertion sequences. *Nucleic Acids Res.* **34**, D32-6 (2006).

86. Carattoli, A. *et al.* In silico detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing. *Antimicrob. Agents Chemother.* **58**, 3895–3903 (2014).

87. Seemann, T. *abricate*. (Github).

88. Lam, M. M. C. *et al.* Genetic diversity, mobilisation and spread of the yersiniabactin-encoding mobile element ICEKp in Klebsiella pneumoniae populations. *Microb Genom* **4**, (2018).

89. Wyres, K. L. *et al.* Identification of Klebsiella capsule synthesis loci from whole genome data. *Microb Genom* **2**, e000102 (2016).

90. Wick, R. R., Heinz, E., Holt, K. E. & Wyres, K. L. Kaptive Web: User-Friendly Capsule and Lipopolysaccharide Serotype Prediction for Klebsiella Genomes. *J. Clin. Microbiol.* **56**, (2018).

91. Cury, J., Jové, T., Touchon, M., Néron, B. & Rocha, E. P. Identification and analysis of integrons and cassette arrays in bacterial genomes. *Nucleic Acids Res.* **44**, 4539–4550 (2016).

92. Ondov, B. D. *et al.* Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.* **17**, 132 (2016).

93. Seemann, T. *mlst*. (Github).

94. Holt, K. Kleborate. https://github.com/katholt/Kleborate.

95. Diancourt, L., Passet, V., Verhoef, J., Grimont, P. A. D. & Brisse, S. Multilocus sequence typing of Klebsiella pneumoniae nosocomial isolates. *J. Clin. Microbiol.* **43**, 4178–4182 (2005).

96. Wirth, T. *et al.* Sex and virulence in Escherichia coli: an evolutionary perspective. *Mol. Microbiol.* **60**, 1136–1151 (2006).

97. Beghain, J., Bridier-Nahmias, A., Le Nagard, H., Denamur, E. & Clermont, O. ClermonTyping: an easy-to-use and accurate in silico method for Escherichia genus strain phylotyping. *Microb Genom* **4**, (2018).

98. Tonkin-Hill, G. *et al.* Producing Polished Prokaryotic Pangenomes with the Panaroo Pipeline. *bioRxiv* 2020.01.28.922989 (2020) doi:10.1101/2020.01.28.922989.

99. Seemann, T. *snippy*. (Github, 2015).

100. Croucher, N. J. *et al.* Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res.* **43**, e15 (2015).

101. Didelot, X., Croucher, N. J., Bentley, S. D., Harris, S. R. & Wilson, D. J. Bayesian inference of ancestral dates on bacterial phylogenetic trees. *Nucleic Acids Res.* **46**, e134 (2018).

102. Plummer, M., Best, N., Cowles, K. & Vines, K. CODA: Convergence Diagnosis and Output Analysis for MCMC. *R News* vol. 6 7–11 (2006).

103. Kembel, S. W. *et al.* Picante: R tools for integrating phylogenies and ecology. *Bioinformatics* **26**, 1463–1464 (2010).

104. Arredondo-Alonso, S. *et al.* mlplasmids: a user-friendly tool to predict plasmid- and chromosome-derived sequences for single species. *Microb Genom* **4**, (2018).

105. Baker, D. N. & Langmead, B. Dashing: fast and accurate genomic distances with HyperLogLog. *Genome Biol.* **20**, 265 (2019).

106. Kalinka, A. T. & Tomancak, P. linkcomm: an R package for the generation, visualization, and analysis of link communities in networks of arbitrary size and type. *Bioinformatics* **27**, 2011–2012 (2011).

107. Oksanen, J. *et al.* vegan: Community Ecology Package. (2019).

108. Lees, J. A., Galardini, M., Bentley, S. D., Weiser, J. N. & Corander, J. pyseer: a comprehensive tool for microbial pangenome-wide association studies. *Bioinformatics* **34**, 4310–4312 (2018).

109. Jombart, T. & Ahmed, I. adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics* **27**, 3070–3071 (2011).

110. Csardi, G. & Nepusz, T. The igraph software package for complex network research. *InterJournal* vol. Complex Systems 1695 (2006).

111. StataCorp, L. P. *Stata statistical software: release 16. College Station, TX: StataCorp LP; 2019*.

112. Lunn, M. & McNeil, D. Applying Cox regression to competing risks. *Biometrics* **51**, 524–532 (1995).

113. R Core Team. R: A Language and Environment for Statistical Computing. (2019).

114. Gasparini, A. comorbidity: An R package for computing comorbidity scores. *Journal of Open Source Software* **3**, 648 (2018).

115. Eyre, D. W. *et al.* Two Distinct Patterns of Clostridium difficile Diversity Across Europe Indicating Contrasting Routes of Spread. *Clin. Infect. Dis.* **67**, 1035–1044 (2018).

116. Yan, W. J. *et al.* Molecular characterization of carbapenem-resistant Enterobacteriaceae and emergence of tigecycline non-susceptible strains in the Henan province in China: a multicentrer study. *J. Med. Microbiol.* **70**, (2021).

117. Wick, R. R., Judd, L. M., Gorrie, C. L. & Holt, K. E. Completing bacterial genome assemblies with multiplex MinION sequencing. *Microb Genom* **3**, e000132 (2017).

118. Wanat, M. *et al.* Exploring the implementation of interventions to reduce catheter-associated urinary tract infections (ENACT): main report. (2021).

119. Johnson, J. R., Delavari, P., Kuskowski, M. & Stell, A. L. Phylogenetic distribution of extraintestinal virulence-associated traits in Escherichia coli. *J. Infect. Dis.* **183**, 78–88 (2001).

120. Picard, B. *et al.* The link between phylogeny and virulence in Escherichia coli extraintestinal infection. *Infect. Immun.* **67**, 546–553 (1999).

121. Galardini, M. *et al.* Major role of iron uptake systems in the intrinsic extra-intestinal virulence of the genus Escherichia revealed by a genome-wide association study. *PLoS Genet.* **16**, e1009065 (2020).

122. Ambulatory Care Emergency Network. *Ambulatory Care Emergency Network: Oxford Case Study*. https://www.ambulatoryemergencycare.org.uk/uploads/files/1/CaseStudies/AEC%20Oxford%20Case%20Study.pdf.

123. Bennett, R. J. *et al.* Temporal GWAS identifies a widely distributed putative adhesin contributing to pathogen success in Shigella spp. *bioRxiv* 2022.08.23.504947 (2022) doi:10.1101/2022.08.23.504947.

124. MacKinnon, M. C. *et al.* Increasing incidence and antimicrobial resistance in Escherichia coli bloodstream infections: a multinational population-based cohort study. *Antimicrob. Resist. Infect. Control* **10**, 131 (2021).

125. Arredondo-Alonso, S. *et al.* Plasmids Shaped the Recent Emergence of the Major Nosocomial Pathogen Enterococcus faecium. *MBio* **11**, (2020).

126. Brodrick, H. J. *et al.* Longitudinal genomic surveillance of multidrug-resistant Escherichia coli carriage in a long-term care facility in the United Kingdom. *Genome Med.* **9**, 70 (2017).

127. Shaw, L. P. *et al.* Niche and local geography shape the pangenome of wastewater- and livestock-associated Enterobacteriaceae. *Sci. Adv.* **7**, eabe3868 (2021).

128. A Study of Three Different Doses of VAC52416 (ExPEC10V) in Adults Aged 60 to 85 Years in Stable Health. *clinicaltrials.gov* https://clinicaltrials.gov/ct2/show/NCT03819049.

129. Lipworth, S. *et al.* Ten years of population-level genomic Escherichia coli and Klebsiella pneumoniae serotype surveillance informs vaccine development for invasive infections. *Clin. Infect. Dis.* (2021) doi:10.1093/cid/ciab006.

130. Diekema, D. J. *et al.* The Microbiology of Bloodstream Infection: 20-Year Trends from the SENTRY Antimicrobial Surveillance Program. *Antimicrob. Agents Chemother.* **63**, (2019).

131. Douglas, N. M., Hennessy, J. N., Currie, B. J. & Baird, R. W. Trends in Bacteremia Over 2 Decades in the Top End of the Northern Territory of Australia. *Open Forum Infect Dis* **7**, ofaa472 (2020).

132. DHSC. *The Health and Social Care Act 2008 Code of Practice of the prevention and control of infections and related guidance*. (2015).

133. Poolman, J. T. & Wacker, M. Extraintestinal Pathogenic Escherichia coli, a Common Human Pathogen: Challenges for Vaccine Development and Progress in the Field. *J. Infect. Dis.* **213**, 6–13 (2016).

134. Frenck, R. W., Jr *et al.* Safety and immunogenicity of a vaccine for extra-intestinal pathogenic Escherichia coli (ESTELLA): a phase 2 randomised controlled trial. *Lancet Infect. Dis.* **19**, 631–640 (2019).

135. Choi, M., Tennant, S. M., Simon, R. & Cross, A. S. Progress towards the development of Klebsiella vaccines. *Expert Rev. Vaccines* **18**, 681–691 (2019).

136. Ingle, D. J. *et al.* In silico serotyping of E. coli from short read data identifies limited novel O-loci but extensive diversity of O:H serotype combinations within and between pathogenic lineages. *Microb Genom* **2**, e000064 (2016).

137. Mamani, R. *et al.* Sequence Types, Clonotypes, Serotypes, and Virotypes of Extended-Spectrum β-Lactamase-Producing Escherichia coli Causing Bacteraemia in a Spanish Hospital Over a 12-Year Period (2000 to 2011). *Front. Microbiol.* **10**, 1530 (2019).

138. Follador, R. *et al.* The diversity of Klebsiella pneumoniae surface polysaccharides. *Microb Genom* **2**, e000073 (2016).

139. Choi, M. *et al.* The Diversity of Lipopolysaccharide (O) and Capsular Polysaccharide (K) Antigens of Invasive Klebsiella pneumoniae in a Multi-Country Collection. *Front. Microbiol.* **11**, 1249 (2020).

140. Grundmann, H. *et al.* Occurrence of carbapenemase-producing Klebsiella pneumoniae and Escherichia coli in the European survey of carbapenemase-producing Enterobacteriaceae (EuSCAPE): a prospective, multinational study. *Lancet Infect. Dis.* **17**, 153–163 (2017).

141. Cardoso, T. *et al.* Classification of healthcare-associated infection: a systematic review 10 years after the first proposal. *BMC Med.* **12**, 40 (2014).

142. Inouye, M. *et al.* SRST2: Rapid genomic surveillance for public health and hospital microbiology labs. *Genome Med.* **6**, 90 (2014).

143. Joensen, K. G., Tetzschner, A. M. M., Iguchi, A., Aarestrup, F. M. & Scheutz, F. Rapid and Easy In Silico Serotyping of Escherichia coli Isolates by Use of Whole-Genome Sequencing Data. *J. Clin. Microbiol.* **53**, 2410–2426 (2015).

144. ECTyper. https://github.com/phac-nml/ecoli_serotyping/tree/master/ectyper.

145. Choby, J. E., Howard-Anderson, J. & Weiss, D. S. Hypervirulent Klebsiella pneumoniae - clinical and molecular perspectives. *J. Intern. Med.* **287**, 283–300 (2020).

146. Boston Consulting Group and The Wellcome Trust. *Vaccines to tackle drug resistant infections An evaluation of R&D opportunities*. https://vaccinesforamr.org/wp-content/uploads/2018/09/Vaccines_for_AMR.pdf.

147. Johnston, B. D. *et al.* Global molecular epidemiology of carbapenem-resistant Escherichia coli (2002-2017). *Eur. J. Clin. Microbiol. Infect. Dis.* (2021) doi:10.1007/s10096-021-04310-6.

148. van den Dobbelsteen, G. P. J. M. *et al.* Immunogenicity and safety of a tetravalent E. coli O-antigen bioconjugate vaccine in animal models. *Vaccine* **34**, 4152–4160 (2016).

149. Inoue, M. *et al.* Safety, tolerability and immunogenicity of the ExPEC4V (JNJ-63871860) vaccine for prevention of invasive extraintestinal pathogenic Escherichia coli disease: A phase 1, randomized, double-blind, placebo-controlled study in healthy Japanese participants. *Hum. Vaccin. Immunother.* **14**, 2150–2157 (2018).

150. Huttner, Hatz & Dobbelsteen. Safety, immunogenicity, and preliminary clinical efficacy of a vaccine against extraintestinal pathogenic Escherichia coli in women with a history of recurrent …. *Lancet Infect. Dis.* (2017).

151. Cryz, S. J., Jr *et al.* Safety and immunogenicity of Escherichia coli O18 O-specific polysaccharide (O-PS)-toxin A and O-PS-cholera toxin conjugate vaccines in humans. *J. Infect. Dis.* **163**, 1040–1045 (1991).

152. Kaijser, B. & Ahlstedt, S. Protective capacity of antibodies against Escherichia coli and K antigens. *Infect. Immun.* **17**, 286–289 (1977).

153. Henriksen, A. Z. & Maeland, J. A. Serum antibodies to outer membrane proteins of Escherichia coli in healthy persons and patients with bacteremia. *J. Clin. Microbiol.* **25**, 2181–2188 (1987).

154. Miajlovic, H. & Smith, S. G. Bacterial self-defence: how Escherichia coli evades serum killing. *FEMS Microbiol. Lett.* **354**, 1–9 (2014).

155. de Lastours, V. *et al.* Mortality in Escherichia coli bloodstream infections: antibiotic resistance still does not make it. *J. Antimicrob. Chemother.* **75**, 2334–2343 (2020).

156. MacKinnon, M. C. *et al.* Mortality in Escherichia coli bloodstream infections: a multinational population-based cohort study. *BMC Infect. Dis.* **21**, 606 (2021).

157. Zhao, S. *et al.* Risk factors for antibiotic resistance and mortality in patients with bloodstream infection of Escherichia coli. *Eur. J. Clin. Microbiol. Infect. Dis.* **41**, 713–721 (2022).

158. Denamur, E. *et al.* Genome wide association study of Escherichia coli bloodstream infection isolates identifies genetic determinants for the portal of entry but not fatal outcome. *PLoS Genet.* **18**, e1010112 (2022).

159. AMRFinderPlus - Pathogen Detection - NCBI. https://www.ncbi.nlm.nih.gov/pathogens/antimicrobial-resistance/AMRFinder/.

160. Wickham, H. ggplot2: Elegant Graphics for Data Analysis. (2016).

161. Biggel, M. *et al.* Horizontally acquired papGII-containing pathogenicity islands underlie the emergence of invasive uropathogenic Escherichia coli lineages. *Nat. Commun.* **11**, 5968 (2020).

162. Sanderson, T., Bileschi, M. L., Belanger, D. & Colwell, L. J. ProteInfer: deep networks for protein functional inference. *bioRxiv* (2021) doi:10.1101/2021.09.20.461077.

163. Ferjani, S., Saidani, M., Ennigrou, S., Hsairi, M. & Ben Redjeb, S. Virulence determinants, phylogenetic groups and fluoroquinolone resistance in Escherichia coli isolated from cystitis and pyelonephritis. *Pathol. Biol.* **60**, 270–274 (2012).

164. Johnson, J. R., Owens, K., Gajewski, A. & Kuskowski, M. A. Bacterial characteristics in relation to clinical source of Escherichia coli isolates from women with acute cystitis or pyelonephritis and uninfected women. *J. Clin. Microbiol.* **43**, 6064–6072 (2005).

165. Kutsukake, K., Nakashima, H., Tominaga, A. & Abo, T. Two DNA invertases contribute to flagellar phase variation in Salmonella enterica serovar Typhimurium strain LT2. *J. Bacteriol.* **188**, 950–957 (2006).

166. Yoon, C. H. *et al.* Mortality risks associated with empirical antibiotic activity in E. coli bacteraemia: an analysis of electronic health records. *bioRxiv* (2022) doi:10.1101/2022.01.22.22269642.

167. Lipworth, S. *et al.* The mobilome associated with Gram-negative bloodstream infections: A large-scale observational hybrid sequencing based study. (2022) doi:10.1101/2022.04.03.22273290.

168. Lipworth, S. *et al.* Optimized use of Oxford Nanopore flowcells for hybrid assemblies. *Microb Genom* **6**, (2020).

169. Matlock, W. *et al.* Flanker: a tool for comparative genomics of gene flanking regions. *Microb Genom* **7**, (2021).

170. Goswami, C. *et al.* Origin, maintenance and spread of antibiotic resistance genes within plasmids and chromosomes of bloodstream isolates of Escherichia coli. *Microb. Genom.* **6**, (2020).

171. Acman, M., van Dorp, L., Santini, J. M. & Balloux, F. Large-scale network analysis captures biological features of bacterial plasmids. *Nat. Commun.* **11**, 2452 (2020).

172. Redondo-Salvo, S. *et al.* Pathways for horizontal gene transfer in bacteria revealed by a global map of their plasmids. *Nat. Commun.* **11**, 3602 (2020).

173. Arredondo-Alonso, S. *et al.* A high-throughput multiplexing and selection strategy to complete bacterial genomes. *Gigascience* **10**, (2021).

174. Wick, R. R., Judd, L. M. & Holt, K. E. Deepbinner: Demultiplexing barcoded Oxford Nanopore reads with deep convolutional neural networks. *PLoS Comput. Biol.* **14**, e1006583 (2018).

175. Wick, R. R., Judd, L. M., Gorrie, C. L. & Holt, K. E. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput. Biol.* **13**, e1005595 (2017).

176. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.* **37**, 540–546 (2019).

177. Holley, G. *et al.* Ratatosk: hybrid error correction of long reads enables accurate variant calling and assembly. *Genome Biol.* **22**, 28 (2021).

178. Walker, B. J. *et al.* Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9**, e112963 (2014).

179. Snipen, L. & Liland, K. H. micropan: an R-package for microbial pan-genomics. *BMC Bioinformatics* **16**, 79 (2015).

180. Xie, Y. *et al.* TADB 2.0: an updated database of bacterial type II toxin-antitoxin loci. *Nucleic Acids Res.* **46**, D749–D753 (2018).

181. Konopka, T. umap: Uniform Manifold Approximation and Projection. (2020).

182. Castro, M. A. A., Wang, X., Fletcher, M. N. C., Meyer, K. B. & Markowetz, F. RedeR: R/Bioconductor package for representing modular structures, nested networks and multiple levels of hierarchical associations. *Genome Biology* vol. 13 R29 (2012).

183. Orlek, A. *et al.* Ordering the mob: Insights into replicon and MOB typing schemes from analysis of a curated dataset of publicly available plasmids. *Plasmid* **91**, 42–52 (2017).

184. Robertson, J., Bessonov, K., Schonfeld, J. & Nash, J. H. E. Universal whole-sequence-based plasmid typing and its utility to prediction of host range and epidemiological surveillance. *Microb Genom* **6**, (2020).

185. Redondo-Salvo, S. *et al.* COPLA, a taxonomic classifier of plasmids. *BMC Bioinformatics* **22**, 390 (2021).

186. Matlock, W. *et al.* Genomic network analysis of environmental and livestock F-type plasmid populations. *ISME J.* **15**, 2322–2335 (2021).

187. Wu, T. NMI: Normalized Mutual Information of Community Structure in Network. (2016).

188. Pöntinen, A. K. *et al.* Apparent nosocomial adaptation of Enterococcus faecalis predates the modern hospital era. *Nat. Commun.* **12**, 1523 (2021).

189. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

190. Wilkins, D. gggenes: Draw Gene Arrow Maps in “ggplot2.” (2019).

191. Gilchrist, C. L. M. & Chooi, Y.-H. Clinker & clustermap.js: Automatic generation of gene cluster comparison figures. *Bioinformatics* (2021) doi:10.1093/bioinformatics/btab007.

192. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).

193. Barry, K. E. *et al.* Don't overlook the little guy: An evaluation of the frequency of small plasmids co-conjugating with larger carbapenemase gene containing plasmids. *Plasmid* **103**, 1–8 (2019).

194. Brueggemann, A. B., Pai, R., Crook, D. W. & Beall, B. Vaccine escape recombinants emerge after pneumococcal vaccination in the United States. *PLoS Pathog.* **3**, e168 (2007).

195. Young, B. C. *et al.* Severe infections emerge from commensal bacteria by adaptive evolution. *Elife* **6**, (2017).

196. Ludden, C. *et al.* One Health Genomic Surveillance of Escherichia coli Demonstrates Distinct Lineages and Mobile Genetic Elements in Isolates from Humans versus Livestock. *MBio* **10**, (2019).

197. Findlay, J. *et al.* Molecular Epidemiology of Escherichia coli Producing CTX-M and pAmpC β-Lactamases from Dairy Farms Identifies a Dominant Plasmid Encoding CTX-M-32 but No Evidence for Transmission to Humans in the Same Geographical Region. *Appl. Environ. Microbiol.* **87**, (2020).

198. Matlock, W. *et al.* Enterobacterales-associated plasmid sharing amongst human bloodstream infections, livestock, wastewater, and waterway niches: a genomic surveillance study in Oxfordshire, UK. *bioRxiv* 2022.05.06.490774 (2022) doi:10.1101/2022.05.06.490774.