

An investigation into the genetic background of *Escherichia coli* lineages containing multi-drug resistant clones

By

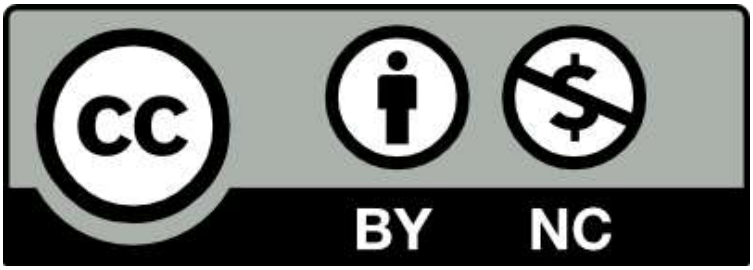
Elizabeth Cummins

A thesis submitted to
the University of Birmingham
for the degree of
DOCTOR OF PHILOSOPHY

Institute of Microbiology and Infection
School of Biosciences
College of Medical and Dental Sciences
University of Birmingham
September 2023

UNIVERSITY OF BIRMINGHAM

University of Birmingham Research Archive
e-theses repository



This unpublished thesis/dissertation is under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) licence.

You are free to:

Share — copy and redistribute the material in any medium or format

Adapt — remix, transform, and build upon the material

The licensor cannot revoke these freedoms as long as you follow the license terms.

Under the following terms:



Attribution — You must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.



NonCommercial — You may not use the material for commercial purposes.

No additional restrictions — You may not apply legal terms or technological measures that legally restrict others from doing anything the license permits.

Notices:

You do not have to comply with the license for elements of the material in the public domain or where your use is permitted by an applicable exception or limitation.

No warranties are given. The license may not give you all of the permissions necessary for your intended use. For example, other rights such as publicity, privacy, or moral rights may limit how you use the material.

Unless otherwise stated, any material in this thesis/dissertation that is cited to a third-party source is not included in the terms of this licence. Please refer to the original source(s) for licencing conditions of any quotes, images or other material cited to a third party.

© Copyright by ELIZABETH CUMMINS, 2023
All Rights Reserved

ABSTRACT

Escherichia coli is a commensal resident of the human gastrointestinal tract that can also cause disease at multiple extra-intestinal body sites. *E. coli* is globally recognised as a pathogen of concern due to the high prevalence of antimicrobial resistance (AMR) present within the species. However, the prevalence of AMR is not uniform across the enormously diverse *E. coli* population. Some *E. coli* lineages remain largely susceptible to antibiotics, whilst others are resistant to many antibiotic classes and are therefore multi-drug resistant (MDR). A wide spectrum of resistance phenotypes have been observed in strains from throughout the *E. coli* phylogeny, meaning that the ability to acquire and accumulate resistance determinants is not an exclusively vertically-inherited trait. A simple explanation for the phylogenetic spread of MDR phenotypes could be that the lineages that encounter antibiotics more frequently are more resistant. Yet this is not the case, as multiple longitudinal studies have shown that large, stable proportions of *E. coli* bloodstream and urinary tract infections are caused by lineages that do not exhibit high levels of resistance. Compounding the problem, MDR *E. coli* are adept at colonising the gastrointestinal tracts of healthy humans who are not taking antibiotics. Studies of international travellers have documented colonisation with MDR *E. coli* occurring within days of arriving in AMR endemic areas.

Specific MDR clones being responsible for the majority of resistant infections, in addition to their rapid colonisation capabilities, pose interesting evolutionary questions. What prerequisites are required for a lineage to become MDR? How are MDR *E. coli* able to displace commensals so rapidly? Are there background genomic features that can potentiate the evolution of resistance? Understanding why AMR is concentrated in specific lineages and how these clones are transmitted will aid efforts to prevent the spread of MDR *E. coli*.

In this study, large-scale and high-resolution genomic approaches have been employed to investigate the success of significant MDR *E. coli* lineages. We hypothesised that antagonistic type VI secretion systems (T6SSs) are not responsible for MDR *E. coli* displacing resident commensals and, as a result, efficiently colonising the human gut. Testing this hypothesis revealed the loss of T6SS components occurring within clades containing MDR lineages. We then demonstrate that lineages of *E. coli* do not evolve in the same way, using a specifically developed comparative pangenomics technique. *E. coli* lineages have the potential to act as genetic sources and sinks for one another and lineage core genomes are enriched in distinct biological processes. This extends to sequence variation in core metabolic genes being present between lineages. Finally, we test the capability of machine learning algorithms to use sequence variation data to predict an MDR genotype. This thesis contributes a greater understanding of how MDR *E. coli* have evolved and highlights the importance of gene-level variation in this process.

DEDICATION

Dedicated to my family.

ACKNOWLEDGEMENTS

I would like to thank my supervisors for their guidance and support throughout my PhD project. Firstly, Alan McNally for agreeing to supervise and introduce me to the field of microbial genomics. Your infectious enthusiasm for science has spurred my academic endeavours and I appreciate the freedom and independence you have given me during my PhD to become the scientist I am today. I have learnt so much from your mentoring, and I am grateful for the knowledge, confidence, and opportunities you have provided me with throughout my time in Birmingham, alongside innumerable Staff House pints. Secondly, Nicole Wheeler for expanding my scientific skill set, serving as an inspiring role model, and providing support in unfamiliar territory. Your expertise and rigour has bettered me as a scientist. I also acknowledge the Wellcome Trust for funding this work.

I am extremely fortunate to have worked in a great lab group and among a fantastic office of scientists, but firstly friends. To every member of the McNally lab past and present - Bex, Rob, Matt, Ann, Izzie, Chris, Steven, and Emma - thank you for your support and making being a 'McNally' a fun job. Burrito Wednesdays, 'labminton', and weekly staff house trips will always be treasured memories from my PhD, so thank you also to Freya, Ryan, Greg, Stan, Ross, Ruth, and everyone else I have met during my time at the IMI. A special thank you to Laura for your ongoing support, I appreciate you.

Finally, I would like to thank my family for their love and support, both throughout my PhD and in everything I do. Your belief in me throughout my life has shaped the person that I am today and taught me that I can achieve anything I put my mind to, so for that I am forever grateful. Thank you for being proud of me.

Contents

Page

1 Introduction **1**

1.1 *Escherichia coli* . . . 2

1.1.1 Population structure . . . 2

1.1.2 Pathotypes of *E. coli* . . . 4

1.2 Antibiotics . . . 5

1.3 Antibiotic resistance . . . 6

1.3.1 Multi-drug resistance in *E. coli* . . . 8

1.4 Colonisation capabilities of MDR *E. coli* . . . 16

1.5 Microbial genomics . . . 18

1.5.1 Pangenomes . . . 19

1.6 Microbial genome evolution . . . 20

1.6.1 Functional gene divergence . . . 21

1.7 Microbial genomics in the era of machine learning . . . 22

1.8 Thesis aims . . . 24

1.8.1 Wider importance . . . 25

2 Parallel loss of type VI secretion systems in two multi-drug resistant *Escherichia coli* lineages **26**

2.1 Introduction . . . 27

2.2 Methods . . . 30

2.2.1 *E. coli* dataset curation . . . 30

2.2.2 T6SS database creation and interrogation . . . 32

2.2.3 Genomic screening for insertion and deletion event sequences, bacteriocins, and phage . . . 32
2.2.4 Strains and culture conditions . . . 33
2.2.5 Competition assay . . . 34
2.3 Results . . . 35
2.3.1 Distribution of T6SS genes in *E. coli* . . . 35
2.3.2 Genomic analysis of ST410 T6SS . . . 39
2.3.3 Genomic analysis of ST131 T6SS . . . 44
2.3.4 Competitive growth of strains with and without a type VI secretion system . . . 53
2.4 Discussion . . . 55

3 Comparative pangenomics identifies distinct evolutionary trajectories within sequence types 59
3.1 Introduction . . . 60
3.2 Methods . . . 62
3.2.1 Sampling diversity of dataset . . . 62
3.2.2 Genome annotation, AMR Identification . . . 63
3.2.3 Pangenome assembly . . . 63
3.2.4 Functional pangenome annotation . . . 64
3.2.5 Distribution of core genomes . . . 65
3.2.6 Core metabolic reconstructions . . . 65
3.2.7 Identifying unique core metabolic genes . . . 65
3.3 Results . . . 66
3.3.1 Dataset properties . . . 66
3.3.2 Pangenome properties . . . 70
3.3.3 Functional pangenomics . . . 73
3.3.4 Allelic variations of metabolic genes . . . 89
3.4 Discussion . . . 93

4 MDR *E. coli* do not share a common evolutionary trajectory of gene degradation 97
4.1 Introduction 97
4.2 Methods 100
4.2.1 Merging pangenomes 100
4.2.2 Bitscore analysis 100
4.2.3 Bitscore quality control 101
4.2.4 Phylogenetically-informed subsampling 101
4.2.5 Delta-bitscore analysis 102
4.2.6 Machine learning 103
4.3 Results 104
4.3.1 A 54,597 gene cluster *E. coli* pangenome 104
4.3.2 Core genome bitscore analysis 105
4.3.3 Predicting multi-drug resistance from functionally significant mutations 113
4.4 Discussion 121

5 Conclusions and Future Work 126

References 130

Supplementary Data 161

List of Figures

1.1 *E. coli* phylogroups 3
1.2 Clade structures of MDR lineages 12
1.3 Evolutionary timeline of MDR lineages 16
1.4 Pangenome schematic 19

2.1 T6SS sub-types in *E. coli* 29
2.2 Distribution of T6SS genes across 21 *E. coli* STs 36
2.3 Presence of structural T6SS genes in ST410 and ST131 38
2.4 Schematic of alignment used to determine the ST410 T6SS locus 41
2.5 The ST410-A/H53 chromosomal T6SS locus 41
2.6 Deletion event in ST410 T6SS locus 43
2.7 The ST131 chromosomal T6SS locus 45
2.8 Insertion event in ST131 T6SS locus 48
2.9 Presence of structural T6SS genes in B2 lineages ST73 and ST95 . 50
2.10 Presence of structural T6SS genes in ST167 and ST648 52
2.11 Pairwise competition assays 54

3.1 Sample source information for genome dataset 68
3.2 Resistance gene carriage by lineage 69
3.3 *E. coli* core genome functions 74
3.4 Significant correlations between the proportion of accessory genome assigned to COG categories and carriage of AMR genes. 78
3.5 Significant correlation between MDR and the proportion of accessory genome assigned to COG categories. 81

3.6 Pangenome segmentation 82
3.7 Functional classification of ST-specific core genomes 83
3.8 Presence of core energy production and conversion genes in ST410 85
3.9 Presence of core energy production and conversion genes in ST648 86
3.10 Presence of core carbohydrate metabolism genes in ST10 86
3.11 Presence of core transcription genes in ST10 87
3.12 Presence of core cell wall/membrane/envelope biogenesis genes in ST95 87
3.13 Presence of core carbohydrate metabolism genes in ST131 88
3.14 Phosphotransferase system component prevalence 91
3.15 Presence of *crt*, *fhuA*, *mntH* and *pduCDEF* alleles 92

4.1 Bitscore analysis workflow 105
4.2 Core genome delta-bitscore density distributions 107
4.3 Q-Q plot of delta-bitscores 108
4.4 Proportion of COG categories with significant delta-bitscore differences between MDR and non-MDR populations. 109
4.5 Upregulated genes bitscore plots 112
4.6 Balanced accuracy of random forest model training 114
4.7 Random forest feature importance 115
4.8 Balanced accuracies of XGBoost model training 116
4.9 XGBoost feature importance 117
4.10 Machine learning predictions 119
4.11 ST131 MDR predictions 121

List of Tables

1.1 Antibiotics 6

2.1 Strains used in this study. 34

3.1 Sequence type pangenome structures and and sampling metrics . 71
3.2 Numbers of genes uniquely core to a single phylogroup and sequence type. 73
3.3 COG category correlation with carriage of AMR genes for lineage core genomes. 76
3.4 COG category correlation with carriage of AMR genes for lineage accessory genomes. 77
3.5 COG category correlation with multi-drug resistance for lineage core genomes. 79
3.6 COG category correlation with multi-drug resistance for lineage accessory genomes. 80
3.7 Numbers of genes and distinct alleles per enriched functional COG category. 90

4.1 COG category correlation with number of genes with significant delta-bitscore differences between MDR and non-MDR populations 110
4.2 Random forest top ten important features 115
4.3 XGBoost top ten important features 117
4.4 Machine learning accuracy by ST 120

1 Phylogroup unique core genes 161

2 ST131-specific core carbohydrate metabolism genes 176
3 ST10-specific core carbohydrate metabolism genes 179

Acronyms

PCR - polymerase chain reaction

MLST - multi locus sequence typing

ST - sequence type

InPEC - intestinal pathogenic *E. coli*

ExPEC - extraintestinal pathogenic *E. coli*

EPEC - enteropathogenic *E. coli*

EHEC - enterohemorrhagic *E. coli*

ETEC - enterotoxigenic *E. coli*

EAEC - enteroaggregative *E. coli*

EIEC - enteroinvasive *E. coli*

UPEC - uropathogenic *E. coli*

NMEC - neonatal meningitis *E. coli*

APEC - avian pathogenic *E. coli*

AMR - antimicrobial resistance

WHO - World Health Organisation

MDR - multi-drug resistant

ESBL - extended-spectrum beta-lactamase

CREC - carbapenem resistant *E. coli*

MRCA - most recent common ancestor

FQR - fluoroquinolone resistant

VF - virulence factor

TCA - tricarboxylic acid

SNP - single nucleotide polymorphism

COMBAT - Carriage Of Muli-resistant Bacteria After Travel

NGS - next-generation sequencing

TGS - third-generation sequencing

HGT - horizontal gene transfer

HMM - hidden Markov model

T6SS - type VI secretion system

LB - lysogeny broth

COG - clusters of orthologous groups

GOI - gene of interest

LPS - lipopolysaccharide

KS - Kolmogorov-Smirnov

RF - random forest

XGB - XGBoost

Chapter One

Introduction

Sections of this chapter are taken from the following publications:

Cummins EA*, Snaith AE*, McNally A, Hall RJ. *The role of potentiating mutations in the evolution of pandemic Escherichia coli clones*. European Journal of Clinical Microbiology & Infectious Diseases. 2021.

Cummins EA*, Hall RJ*, McInerney JO, McNally A. *Prokaryote pangenomes are dynamic entities*. Current Opinion in Microbiology. 2022;6.

1.1 *Escherichia coli*

Escherichia coli is a Gram-negative rod-shaped bacterium that was discovered in 1886 by the paediatrician Theodor Escherich [180], whom it is now named after. Today, *E. coli* is widely considered to be one of the most extensively studied organisms and is frequently used as a model organism in scientific research. Beyond the laboratory, *E. coli* exhibits a vast array of variable lifestyles, including environmental, commensal, and pathogenic organisms. The species also accommodates a huge breadth of intra-specific variation, and a lifestyle spectrum from commensal to globally disseminated pathogenic clones [32]. The diversity contained within this species, due to the habitation of multiple distinct and disparate niches, is reflected by its complex population structure.

1.1.1 Population structure

Phylogroups

E. coli can be divided into eight main phylogroups (A, B1, B2, C, D, E, F, and G) alongside additional cryptic clades I-V. The main phylogroups represent major ancestral divergences in the *E. coli* population and are shown in Figure 1.1. A, B1, B2, and D were the first four groups to be established [73], with further groups proposed as sequencing technologies developed [36, 37, 191]. A method of phylotyping using a triplex polymerase chain reaction (PCR) was proposed in 2000 by Clermont and colleagues [34]. This has since been superseded by ClermonTyping, a bioinformatic tool that provides an '*in silico* PCR assay' to phylotype *E. coli* assemblies [7]. Phylogroups can be a good indicator of various isolate properties, including genome size [9], host organism [191], and pathotype [48]. This information can provide insight into whether a trait has arisen independently on multiple occasions across the phylogeny or is strictly vertically in-

herited.

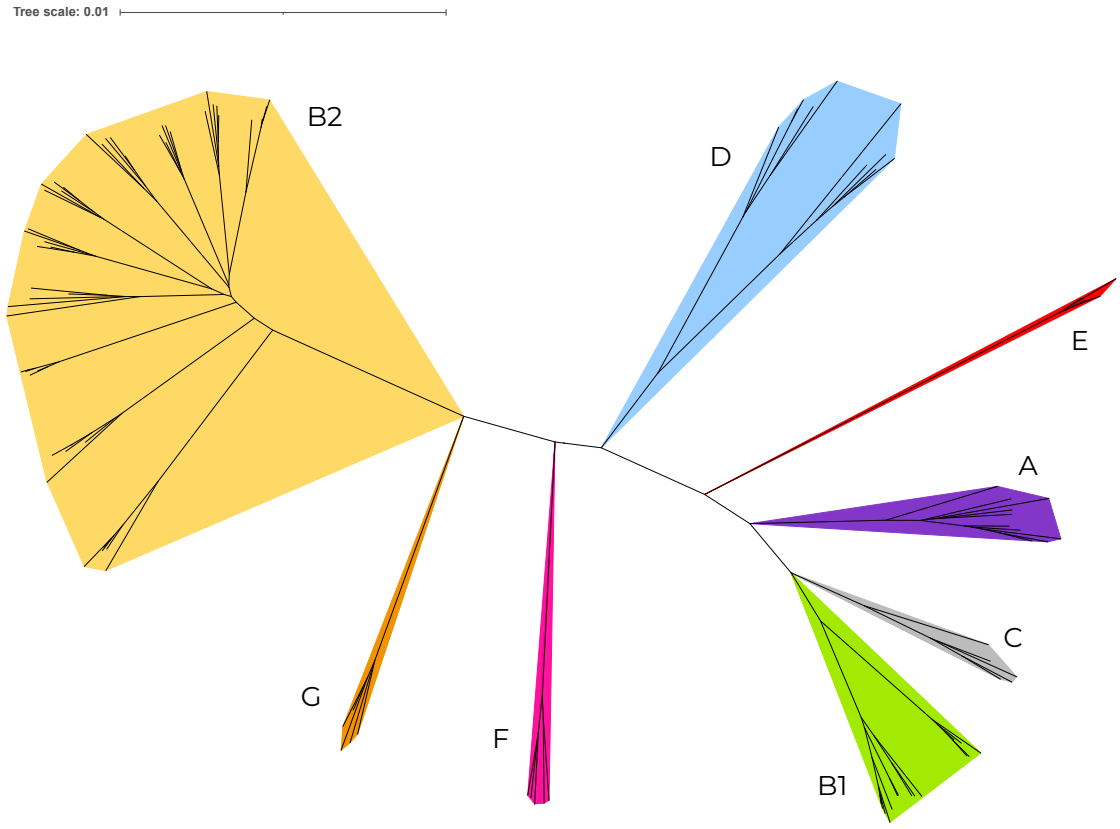


Figure 1.1: Phylogenetic tree of *E. coli*. Representative sampling of the data set used in this work to demonstrate the species population structure. The major phylogroups are coloured and labelled.

Sequence types

Multi Locus Sequence Typing (MLST) is a technique used to delineate clonal groups within a phylogroup. MLST involves the comparison of nucleotide sequences from multiple well conserved chromosomal house-keeping genes. There are three available MLST schemes; Warwick [210], Pasteur [80], and Michigan [163]. The combination of alleles of scheme-specific house-keeping genes constitute a profile that is referred to as an isolate's Sequence Type (ST). The

number and type of loci used to type an isolate varies by MLST scheme, but all loci were chosen as they are conserved across the *E. coli* population. The Warwick Achtman seven gene scheme is the most commonly used and has been deemed the most phylogenetically accurate [35]. Unlike phylogroups, a ST does not usually encompass a spectrum of pathotypes.

1.1.2 Pathotypes of *E. coli*

Pathogenic *E. coli* can cause a range of diseases based on their genetic factors and physical location in the body. Broadly, they can be divided into intestinal pathogenic *E. coli* (InPEC) or extraintestinal pathogenic *E. coli* (ExPEC), both of which are found throughout the *E. coli* phylogeny. There are a variety of pathotypes within InPEC, including enteropathogenic *E. coli* (EPEC), enterohemorrhagic *E. coli* (EHEC), enterotoxigenic *E. coli* (ETEC), enteroaggregative *E. coli* (EAEC), and enteroinvasive *E. coli* (EIEC). EPEC were the first pathogenic *E. coli*, identified in 1945 [22], and are characterised by the attaching and effacing lesions they cause on intestinal epithelial cells [48]. EHEC and ETEC are characterised by their specific Shiga toxin *stx* and enterotoxin genes. EAEC and EIEC can be defined by their specific pathogenicity behaviours, the former adheres to intestinal epithelial cells whilst the latter invades the epithelial cells. All InPEC pathotypes are found across multiple phylogroups [48].

ExPEC is the dominant pathotype of multi-drug resistant (MDR) *E. coli* and will be the main focus of this thesis. ExPEC are commonly categorised based on their target organ or host. For example, uropathogenic *E. coli* (UPEC) target the urinary tract and neonatal meningitis *E. coli* (NMEC) is found in the cerebrospinal fluid of newborns. Some ExPEC pathotypes are associated with specific phylogroups. For example, avian pathogenic *E. coli* (APEC) cause disease in birds and are often specifically associated with phylogroup G [36]. Generally, B2

is the dominant phylogroup for ExPEC but they are also found in A, D, and F [48]. ExPEC can reside in the host gut without causing disease and are often found in healthy individuals, therefore earning the label of 'facultative pathogen' [144]. Some have hypothesised that ExPEC pathogenicity is an accidental outcome of enhanced intestinal colonisation capabilities [100, 102]. Colonisation factors can include adhesins, iron uptake systems, fimbriae, and secretion systems [65, 114, 131, 141].

1.2 Antibiotics

Antibiotics are compounds that target bacterial cells to either inhibit their growth (bacteriostatic) or induce cell death (bactericidal). Antibiotics can be classified based on the cellular processes and bacterial structures that they target. The antibiotic classes relevant to this thesis are aminoglycosides, beta-lactams, fluoroquinolones, polymyxins (which are considered a 'last-resort' treatment option for Gram-negative multi-drug resistant (MDR) infections), and rifamycins. The cellular targets and modes of action of these antibiotic classes, alongside examples are shown in Table 1.1. Unfortunately, antibiotics are not a final solution to bacterial infections as bacteria have evolved mechanisms to evade these drugs.

Table 1.1: Characteristics of antibiotics relevant to this thesis.

Antibiotic class	**Example**	**Target**	**Mode of Action**	**Ref.**
Aminoglycosides	Kanamycin	30S ribosomal subunit	Inhibit protein synthesis	[198]
Beta-lactams	Cefotaxime	Transpeptidase enzymes involved in cross-linking peptidoglycan chains	Inhibit cell wall synthesis	[106]
Fluoroquinolones	Ciprofloxacin	DNA gyrase or topoisomerase IV	Inhibit DNA replication and transcription by preventing DNA unwinding	[75]
Polymyxins	Colistin	Outer membrane	Increase permeability of outer membrane	[156]
Rifamycins	Rifampicin	Beta subunit of RNA polymerase	Inhibit RNA synthesis	[139]

1.3 Antibiotic resistance

The phenomenon of antimicrobial resistance (AMR) is a significant problem in modern medicine. Resistance to antibiotics hinders treatment of bacterial infections. Human pathogens have been evolving resistance to clinical antibiotic treatments for over 60 years [186], but AMR genes have been shown to predate antibiotics by millions of years in studies of ancient permafrost [165]. There are numerous resistance mechanisms that bacteria can use to evade antibiotics.

Many antibiotics target essential cellular processes, such as transcription (rifamycins), DNA replication and repair (fluoroquinolones), and cell wall synthesis (beta-lactams). The antibiotics act by occupying binding sites used in these

cellular processes which inhibits overall activity. Bacteria can evade this strategy by acquiring mutations in antibiotic target sites which reduce antibiotic binding, allowing the essential cellular processes to proceed uninhibited. For example, specific mutations in *gyrA* and *gyrB* encoding DNA gyrase and *parC* and *parE* encoding topoisomerase IV provide fluoroquinolone resistance in *E. coli* [152]. Bacteria can also actively modify an antibiotic to reduce its binding capability. This resistance mechanism is carried out by production of a specific enzyme that will alter the structure of the antibiotic, making it less efficient at binding to the targeted active site. For example, *E. coli* can become resistant to beta-lactam antibiotics by producing beta-lactamase enzymes that hydrolyse the functional group (beta-lactam ring) of this antibiotic class [27]. Using alternatives to antibiotic targets is another route to resistance in bacteria. Resistance can be conferred *via* acquisition of alternative genes. For example, *E. coli* can become resistant to sulphonamides through acquisition of exogenous *sul* genes that encode alternative, sulphonamide-resistant dihydropteroate synthase enzymes [184].

Increased efflux or decreased membrane permeability confer resistance by lowering the concentration of antibiotic within a cell [15]. Mutations in the porin-encoding *ompF* in *E. coli* reduce the absorbance of beta-lactam antibiotics. Efflux pumps can provide resistance to a large range of antibiotic classes due to their non-specificity [15]. In *E. coli*, multi-drug efflux pumps provide resistance to several classes of antibiotics, including macrolides, tetracyclines, and aminoglycosides [15].

Antibiotic resistance is increasing within *E. coli* and this species currently ranks third in the World Health Organisation (WHO) list of priority pathogens [208]. In some cases pathogens have evolved to become resistant to multiple different classes of antibiotics, gaining the title ‘multi-drug resistant’ (MDR). Multi-drug resistance furthers the difficulty in managing the spread and dissemination of human pathogens.

1.3.1 Multi-drug resistance in *E. coli*

The emergence of MDR *E. coli* clones has been rapid, with a notable rise in the last three decades that is attributed widely to increased global travel [4, 11, 19] and the current usage of antimicrobials in humans and animals. MDR *E. coli* are typically ExPEC, and the pandemic *E. coli* ST131 lineage is the dominant cause of MDR ExPEC infections across the globe [151, 185]. Infections of this type include urinary tract infections and bloodstream infections. Clinical causes of concern are extended-spectrum beta-lactamase (ESBL)-producing *E. coli*, resistant to beta-lactam antibiotics including penicillins, cephalosporins and monobactams [110], and carbapenem resistant *E. coli* (CREC). The increasing presence of CREC further limits antibiotic treatment options, with colistin being one of the few remaining drugs of last resort. Intriguingly, multi-drug resistance is only observed in some *E. coli* lineages and can be restricted to specific clones within them [185]. Understanding this restricted distribution of multi-drug resistance in *E. coli* is essential if we are to predict and prevent the emergence of new clones.

There are particular STs of interest in the context of multi-drug resistance that will be discussed in detail throughout this work. These STs either contain a MDR clone (ST131 and ST410), or the lineage itself is generally MDR (ST648 and ST167).

ST131

ST131 belongs to phylogroup B2 and is a widely studied lineage of *E. coli*, capable of frequent inter-species transmission and containing within it several important pandemic MDR clones [130]. The most recent common ancestor (MRCA) of ST131 is estimated to have emerged in the mid 1800s [185] (Figure 1.3). The ST131 phylogeny has a three-clade structure that corresponds with the pos-

session of different *fimH* alleles; H41, H22, and H30 (alternatively referred to as clades A, B, and C in the Petty classification) [151, 185] (Figure 1.2A). A prominent sub-lineage, ST131-C/H30 incorporates a H30S (susceptible) basal component (known alternatively as C0), which gave rise to the ST131-C/H30R (resistant) clade [8, 56, 82, 128].

ST131-C/H30R, characterised by the FimH30 allele, is thought to have diverged from the drug-susceptible H30S basal component in 1982, before splitting into parallel sister sub-clades H30R1 and H30Rx [8, 86]. Both ST131-C/H30R1 and ST131-C/H30Rx clones are fluoroquinolone resistant and carry $bla_{\text{CTX-M}}$ genes that confer resistance to ESBLs [158, 185]. ST131-C/H30R, possessing $bla_{\text{CTX-M-14}}$, was prevalent in the early 2000s but numerous studies indicate that it has since been superseded by ST131-C/H30R1 ($bla_{\text{CTX-M-27}}$) and ST131-C/H30Rx ($bla_{\text{CTX-M-15}}$) [13, 90, 127, 128, 134]. The extent of the prevalence of these different ESBLs varies by geography and study location (e.g. community, hospital) [13, 90, 127, 128, 134]. The ST131-C/H30Rx subclade carries $bla_{\text{CTX-M-15}}$ [158, 185]. Global nosocomial infections of ST131-C/H30Rx were reported by 2008 and this clone has since become dominant [128, 142, 143]. ST131 is perhaps an unusual lineage because it is a successful pathogen whilst encoding multiple plasmid-encoded resistance genes, thereby contradicting the expectation that fitness costs associated with plasmid carriage would decrease pathogenicity [142]. Plasmids carrying ESBL genes have been shown to incur no general fitness cost in ST131, and in some cases even increase virulence properties [161].

It has been suggested that mutations in gene regulatory regions may compensate for plasmid fitness costs and contribute to the dominance of ST131 as an MDR pathogen in the clinical setting [55, 66, 130, 131]. Analysis of gene regulatory regions across ST131-A/H41, ST131-B/H22 and ST131-C/H30 has shown that the alleles found in ST131-A/H41 and ST131-B/H22 are more sim-

ilar to one another than those of ST131-C/H30, suggesting that the long-term presence of multi-drug resistance plasmids in the latter could be the influential factor [130]. Numerous studies have shown that the genetic background of the recipient heavily influences the fitness impact of a plasmid, although more work is needed to pinpoint the specific epistatic features that enable these clones to carry multi-drug resistance plasmids so efficiently [50, 66, 97].

ST131-C/H30Rx is thought to be particularly proficient in acquiring and maintaining large multi-drug resistance plasmids [131]. Acquisition of the $bla_{\text{CTX-M-15}}$ gene via FII-type plasmids was a key evolutionary event in the formation of ST131-C/H30Rx, a subclade of ST131-C/H30R [167, 185]. The dominant variant harboured by ST131-C/H30Rx is $bla_{\text{CTX-M-15}}$, but other variants including $bla_{\text{CTX-M-14}}$ and $bla_{\text{CTX-M-27}}$ can be found within ST131-C/H30R1 at lower frequencies [185]. Beyond resistance, ESBL plasmids can also influence beneficial colonisation traits including biofilm formation, further solidifying the repertoire of enhanced survival capabilities [130, 175].

Pathogenic lineages of *E. coli* are heavily shaped by foothold genetic changes that can be considered primary events that give rise to phenotypes favouring colonisation [51]. A key factor that shapes the ability of *E. coli* lineages to colonise a novel niche, particularly their transition from commensal to pathogenic, is their complement of virulence factors (VFs). VFs can be introduced via mobile genetic elements or located on chromosomal pathogenicity islands [44], with the former being particularly problematic. A mobile VF that improves the fitness of a pandemic clone may result in rapid spread through a community or healthcare setting.

ExPEC-specific traits, such as enhanced iron-uptake and adhesion to extraintestinal tissue, are encoded by VFs that provide ST131 with a competitive advantage in host colonisation [131, 153, 175]. ST131 clones can successfully overcome host restriction of iron availability, a common defence strategy to com-

bat bacterial pathogens [74]. Siderophores are secreted by a clone to bypass this nutritional immunity as siderophores bind to Fe^{3+} with a greater affinity than the host proteins, transferrin and lactoferrin, to relinquish sequestered iron [74]. *E. coli* ST131 clones are known to encode siderophore-related genes including the aerobactin siderophore receptor *iutA* [153]. VFs may therefore increase the likelihood that a clone will colonise successfully through their contribution to cellular invasion and adhesion [126, 177].

Alongside colonisation, proficient metabolic capabilities are required for wider dissemination of a clone. Genes involved in anaerobic metabolism were found to be more diverse in the drug-resistant ST131-C/H30 in comparison to the drug-susceptible ancestral clades [131]. For example, seven variants of the *eutA* gene, encoding an ethanolamine ammonia-lyase reactivase that acts in the ethanolamine degradation pathway, were found in ST131-C/H30 genomes. A distinct allele of the *cobW* gene, involved in the biosynthesis of the cofactor cobalamin, was also found in the drug-resistant clade. Together, these imply selection for anaerobic metabolism [131]. A recent study has also found that, in the presence of antibiotics, strains of *E. coli* from clinically relevant lineages will develop mutations in core metabolic genes [112]. This includes mutations in the *icd* gene that encodes isocitrate dehydrogenase, a key enzyme within the tricarboxylic acid (TCA) cycle. That variation can occur in core central metabolism genes when under selection pressure from antibiotics underscores the likeliness of a relationship between metabolism and resistance that should be explored further.

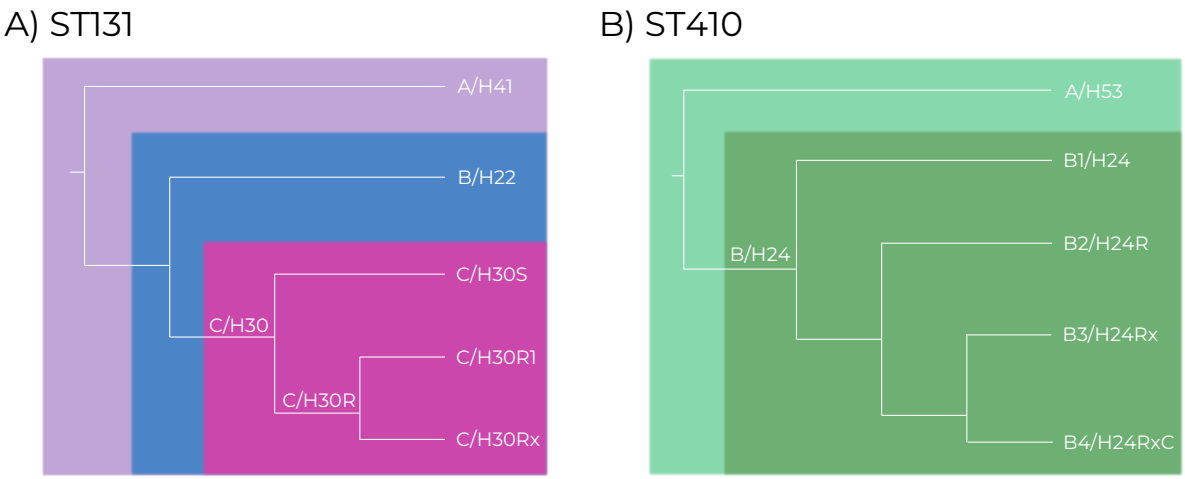


Figure 1.2: Representative phylogenies demonstrating the clade structure of MDR lineages. A) The three clade structure of ST131: A/H41 (purple), B/H22 (blue), C/H30 (pink). B) The two clade structure of ST410: A/H53 (light green) and B/H24 (dark green). Not drawn to scale.

ST410

ST410 belongs to phylogroup A and has a two-clade phylogenetic structure (A and B, the latter of which contains an MDR clone, Figure 1.2B). Like ST131, it is estimated to have emerged in the 1800s, with key clones, specifically B2/H24R and B3/H24Rx, arising in the 1980s [167] (Figure 1.3). This was followed by the emergence of the carbapenem-resistant clone ST410-B4/H24RxC, carrying the $bla_{\text{OXA-181}}$ and $bla_{\text{NDM-5}}$ carbapenemase genes in addition to the $bla_{\text{CTX-M-15}}$ encoded by ST410-B3/H24Rx [60, 167]. Crucially, multiple polymorphisms have been identified that have influenced the evolution of ST410-B4/H24RxC and that are associated with acquisition of these carbapenemase genes. These potentiating mutations include porin genes *ompC* and *ompF*, and the beta-lactam target *ftsI* [148]. The acquisition of a $bla_{\text{CTX-M-15}}$ gene, via an FII-type plasmid, is another feature of ST410-B4/H24RxC in common with ST131 clones [167, 185].

Clones within the ST410 lineage display other parallelisms with ST131 pandemic clones in several regards. ST410-B4/H24RxC is for example able to sequester iron from lactoferrin [60], a property which promotes gut colonisation. The genetic causes of this key phenotype are single nucleotide polymorphisms (SNPs) and clone-specific core gene alleles, most notably a unique allele of *fhuA*, encoding the ferrichrome porin responsible for ferric hydroxamate uptake [120]. In this instance, the unique *fhuA* allele was introduced via a homologous recombination event early in the formation of the clone. ST410-B4/H24RxC is also associated with distinct clone-specific alleles of *fimH*, Fim24 [167], and fimbriae-like protein *yadC* [60]. The speculation that the FimH30 allele of ST131-C/H30Rx improves fitness of the clone may reasonably be extended to ST410-B4/H24RxC and FimH24. Another mirrored trait between ST410 and ST131 pandemic clones is the observation that ST410-B4/H24RxC does not show an augmented virulence phenotype in comparison to other isolates of the ST410 lineage [60].

Furthermore, there is evidence for clone-specific SNPs in anaerobic metabolism loci in ST410-B4/H24RxC [60] that mirror prior findings within ST131 [130]. Specifically, a study of clinical CREC strains identified 382 SNPs in ST410-B4/H24RxC, of which 60 were non-synonymous mutations in 48 genes [60]. The authors note that the majority of these genes function in metabolism, including *araAB* (L-arabinose degradation), *thiBPQ* (thiamine transport), and three genes involved in anaerobic metabolism, *lueB*, *lpdA*, and the putative shikimate dehydrogenase gene. In another parallelism with ST131, fluoroquinolone resistance (FQR), arising from mutations in *gyrA* and *parC*, is known to be present in ST410 clones [167].

The ST410-B4/H24RxC MDR clone is further distinguished from the rest of the ST410 lineage by specific SNPs in genetic loci involved in gene regulation [60]. The intergenic regions in which these SNPs have been found include those

upstream of *fhuA*, and downstream of *sgrR* and *sgrT*, the regulator and inhibitor, respectively, of a glucose transporter [60]. Intergenic SNPs, including those related to metabolic processes, may therefore be considered as a species-wide evolutionary marker of an MDR clone, given the parallelisms between ST410 and ST131.

ST167

ST167 is, like ST410, a lineage within phylogroup A. ST167 emerged from the ST10 clonal complex and has only been reported clinically within the last 10 years [215]. It is now a significant CREC lineage that is known to carry the carbapenemase gene $bla_{\text{NDM-5}}$ [215, 220]. The evolution of ST167 parallels ST131-C/H30 in the diversity exhibited in its anaerobic metabolism genes [43]. It is known to encode, for example, unique alleles of *tdcD*, a propionate kinase that functions in the anaerobic degradation of L-threonine [220]. A clone-specific allele of *eutA* is present in ST167, a reactivase that also has clone-specific alleles present within ST131 [131, 220]. While ST410-B4/H24RxC is known to have a unique allele of *fhuA*, ST167 possesses a variant of *fhuB*, encoding an iron(III) hydroxamate ATP-binding cassette (ABC) transporter membrane subunit [220]. These variations, and that observed in the anaerobic metabolism related genes of other MDR clones, may contribute to an increased fitness in the environment of the mammalian gut, and to the overall success of a pandemic clone.

In a similar manner to clones of ST131 and ST410, SNPs also likely played a role in the expansion of the ST167 lineage. ST167 exhibits clone-specific compensatory mutations in intergenic regions, namely IGR0012 and IGR0014, mirroring observations in ST131 and ST410. These mutations in intergenic regions are proposed to be an evolutionary event underpinning the successful integration of large AMR plasmids and hence the emergence of multi-drug resistance

[220].

This CREC lineage is also known to have acquired mutations that link to resistance, notably *ftsI* which encodes the penicillin-binding protein 3. This has high similarity to an allelic variation of the same gene that is present in carbapenem resistant ST410 strains [148, 220]. It could therefore be concluded that ST167 displays many of the same hallmarks of multi-drug resistance evolution that are epitomised in ST131.

ST648

ST648 is a globally disseminated ExPEC that is known to encode resistance to ESBLs [177, 212, 219] that was first reported in 2009 [181] (Figure 1.3). ST648 is known to harbour more virulence-associated genes than the mostly commensal ST10 lineage [175]. This includes VFs that promote colonisation, a trait observed in ST131 pandemic clones. VFs that relate to colonisation and competitiveness and that are common to both lineages include the previously mentioned *iutA* and *ompT* [177]. The ST648 pandemic lineage also exhibits enhanced curli fiber production in comparison to ST10, aiding adherence to epithelial cells, and biofilm formation [175]. This phenotype is particularly interesting given that the ST648 isolates investigated by Schaufler and colleagues were found to lack certain genes, including the *pap* operon, that are typically associated with urinary tract infections [175]. There are however other isolates of ST648 that are predicted to encode *pap* genes [218], indicating there may be interesting differences within this lineage. ST648 has been shown experimentally to have colonisation capabilities comparable with ST131 in a chicken *in vivo* assay [175]. ST648 is categorised into phylogroup F, demonstrating that multi-drug resistance can be found across the *E. coli* phylogeny. As yet, little is known about the potential for important metabolic variations within ST648. Given the

observations surrounding SNPs in anaerobic metabolism genes in other *E. coli* lineages, this should be investigated as a priority in ST648 to establish whether metabolic variation is an important part of its evolution.

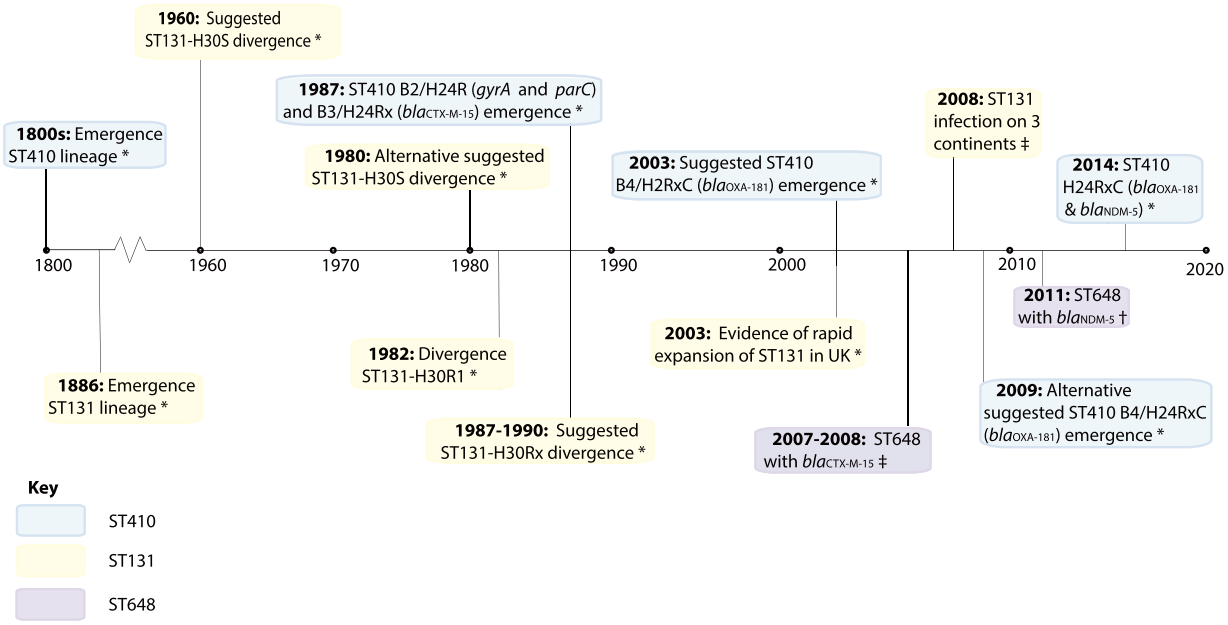


Figure 1.3: Key evolutionary dates for ST131 (yellow), ST410 (blue), and ST648 (purple). Timeline demonstrates emergence dates estimated by studies using time-scaled phylogeny techniques (*). Examples of key dates first reported (†) and first observed (‡) in published papers are also noted. Estimated dates vary between different datasets and these have been included as alternative suggestions. Key references are provided at the relevant points in text. Adapted from [43].

1.4 Colonisation capabilities of MDR *E. coli*

Assessing the colonisation capabilities of MDR *E. coli* is essential for implementing knowledge-driven prevention strategies to reduce transmission. Multiple travel studies sampling participants prior to and following travel have provided insight into the colonisation capabilities of these pathogens. An observation

study of travellers to South Asia reported extensive colonisation by ESBL-producing *E. coli* with 89% of the cohort carrying $bla_{CTX\text{-}M}$ positive strains post travel that were all $bla_{CTX\text{-}M}$ negative prior to travel [11]. The Carriage Of Multi-resistant Bacteria After Travel (COMBAT) study reported similar findings with a much larger cohort [4]. This study also investigated household transmission post-travel, reporting a 12% probability of passing an ESBL-producing *Enterobacteriaceae* to another household member, highlighting a further threat posed by these pathogens [4]. Pre- and post-travel sampling however cannot capture colonisation dynamics during travel. Kantele, Kuenzli, Dunn and colleagues carried out a study on 20 participants travelling from Switzerland to Laos for a medical course in 2015 [89]. Over the duration of this study, daily faecal samples were provided and cultured to examine the real-time dynamics of intestinal colonisation by MDR bacteria. They found that all participants became colonised by MDR bacteria within days of arriving at their destination. In initial pre-travel sampling of the cohort only 35% were positive for MDR bacteria, showing that this colonisation process occurs rapidly. The ability to rapidly displace susceptible commensals has been hypothesised to be a result of variation in metabolic capabilities [55].

It is important to note that clonal success may not depend purely on the possession of a resistant phenotype. A recent study of surveillance isolates collected in Norway found that the largely drug-susceptible ST131-A/H41 of ST131, notably lacking $bla_{CTX\text{-}M}$ and FQR genes, was well-established and dominant, relative to more resistant ST131-B/H22 isolates [68]. If resistance was the sole driver of the success of a clone, one might expect that the MDR clone, here from ST131-C/H30, would over time become the dominant isolate. A reanalysis of this dataset has found that 61 of the 75 ST131-A/H41 genomes did however encode non-$bla_{CTX\text{-}M}$ beta-lactamases, raising the question of whether ST131-A/H41 can in fact be defined as drug-susceptible [107]. Further, a longitudinal study of *E. coli* isolates in Oxfordshire, UK (2008-2018) has found that the pres-

ence of AMR genes is linked to increasing incidences of bloodstream infections [108]. The drivers of *E. coli* clonal success are therefore likely to be multifaceted, and with AMR not necessarily essential for the success of a clone [68].

1.5 Microbial genomics

A new era of microbiology began with the emergence of microbial genomics, driven by rapid advancements in sequencing technologies. The first complete microbial genome was sequenced in 1995 [62] and there are now over 660,000 genomes publicly available [14], on top of read data that is increasingly being submitted without associated assemblies. Two key milestones in the era of microbial genomics were the development of next-generation sequencing (NGS) and third-generation sequencing (TGS). NGS, exemplified by Illumina, was developed in the mid-2000s and enabled rapid, parallel sequencing of DNA fragments. The reads produced are relatively short in length ($\sim$ 100 bp) [111]. A caveat of this method is that genomic repeat sequences longer than read lengths are difficult to resolve at the assembly stage. NGS was then followed by the development of TGS, exemplified by PacBio and Oxford Nanopore technologies, which overcame previous limitations of NGS by producing longer reads ($\sim$ 100 kb) and real-time data with bacterial genomes being sequenced in a matter of hours. TGS can also enhance genome assembly and help to resolve complex genomic structures through hybrid assembly, where long and short read sequencing data are used in combination [95]. These sequencing technologies have led to an exponential increase in the volume of microbial genome data available. Consequently, this has motivated advances in bioinformatic tools and techniques to facilitate the analysis and interpretation of data.

1.5.1 Pangenomes

The field of pangenomics emerged with the acceleration of genome sequencing and the observation of intra-specific genomic variation. It is therefore a young, rapidly expanding area of research. As the union of all genes occurring within a set of genomes, the pangenome is often considered in terms of two major 'sections'. The inter-sectional area is referred to as the core genome and all genes outside of this intersection are classified as the accessory genome (Figure 1.4), although in practise a defined core genome is often not this strict. Shell and cloud are common subdivisions within the accessory genome that classify genes at more specific presence thresholds [94]. Shell genes are moderately conserved and present at intermediate frequencies, whereas cloud genes have very low frequencies or are uniquely present in a single genome [94].

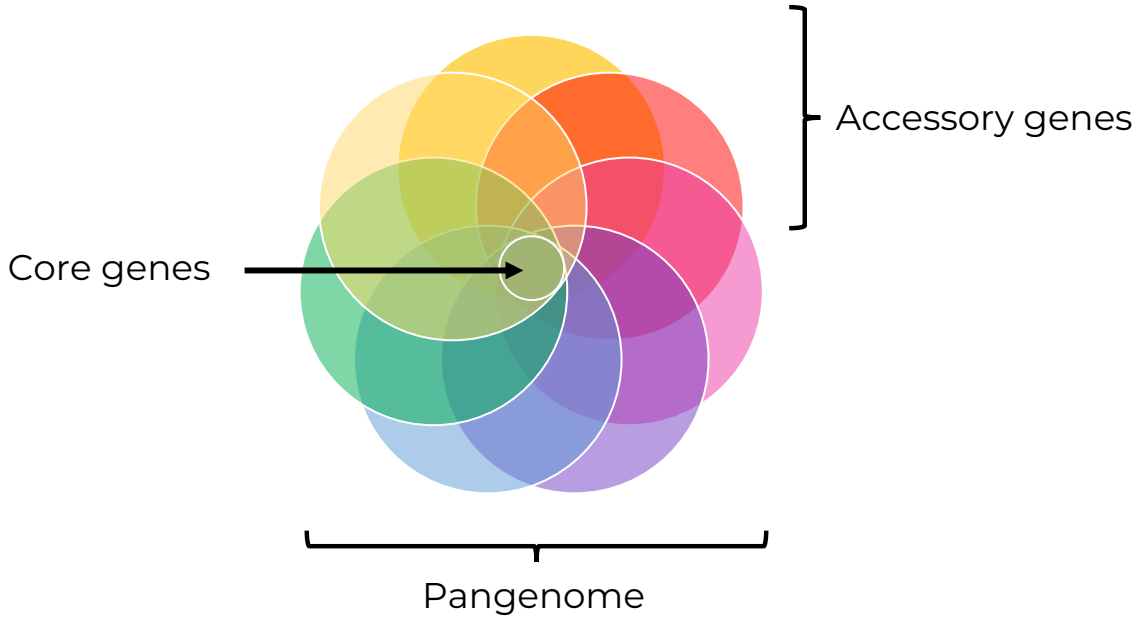


Figure 1.4: Schematic of a pangenome as a Venn diagram.

The first pangenome study in prokaryotes was carried out in 2005 [192].

Mathematical analysis concluded that the size of the *S. agalactiae* pangenome was essentially infinite; every genome sequenced was predicted to add approximately 33 new genes to the pangenome [192]. Since the publication of this work the pangenome concept has been applied to multiple phylogenetic classifications including genus and super kingdom [99, 101]. When the size of a pangenome expands with every addition of an independent strain it is described as an 'open' pangenome. *E. coli* is considered to possess an open pangenome as it displays an extensive range of genetic diversity [64].

The pangenome has been a great boon for evolutionary discoveries. The seminal work by Rasko and colleagues in 2008 identified few pathotype-specific genes, providing evidence that pathogenic lineages have evolved multiple times within *E. coli*. Comparisons including the first human commensal genome (HS) shed light on the 'mosaic structure' of HS and the lack of HS-unique genes which pointed towards commensal strains acting as genetic repositories for virulence factors [162]. More recent studies have uncovered significant associations between resistance and the accessory genome [137] and between multi-drug resistance and purine metabolism [149] in the *E. coli* pangenome. There have been multiple pangenomic studies with specific focus on ST131 [47, 130, 131]. Allelic diversity of anaerobic metabolism genes, alongside other colonisation factors, in the ST131 pangenome have been shown to correlate with multi-drug resistance [131]. Pangenome studies have also been applied to pathotypes. An APEC pangenome also revealed variation in metabolism, this time linked to pathogenicity, alongside AMR, toxicity, and heat shock response genes [121].

1.6 Microbial genome evolution

Genome evolution is driven by mutations, gene acquisition *via* horizontal gene transfer (HGT), and extensive gene loss [159, 195]. Bacterial genomes have high

mutation rates and a strong bias toward deletions [135]. Gene loss can occur via two evolutionary processes; selection or genetic drift. Adaptive loss, caused by selection, is the removal of specific genes due to their fitness cost or unnecessary functional burden in a given environment. Conversely, random mutations arise all the time, and in the absence of selection or during inefficient selection, these mutations may persist. The accumulation of random mutations may eventually result in gene inactivation. These evolutionary concepts are important when considering antibiotic resistance as recent studies on *E. coli* have established that the genetic background and potentiating mutations arising in genes not usually associated with AMR in response to antibiotic treatment influence evolution of resistance [29]. This thesis focuses on identifying genes that have accumulated mutations that alter function, and identifying associations between these events and multi-drug resistance.

1.6.1 Functional gene divergence

Profile hidden Markov models (HMMs) are statistical models informed by multiple sequence alignments to capture the probabilities of observing specific residues (amino acid or nucleotide) at each position in the alignment, therefore acting as a position-specific scoring system [58]. Delta-bitscoring is a profile HMM-based method developed by Wheeler and colleagues to measure the impact of sequence variation on protein function that is suitable for use in large comparative genomic studies [205]. Examining functional disruption of genes may provide better insight for predicting genomic shifts that relate to a change in phenotype, as subtle mutations can be detected prior to the more obvious signature of complete gene loss [98]. A successful application of the delta-bitscore method was able to identify signatures of host adaption in *Salmonella enterica* [205]. Delta-bitscore output was used as training variables for a machine learning model that correctly identified loss of function mutations in genes involved

in metabolic pathways as being a signature of invasive serovars.

In *E. coli*, mutations in metabolic genes have been linked to antibiotic resistance, although the functional impact of these mutations have yet to be explored in depth. McNally and colleagues have shown that ST131 clade C has acquired and maintained multi-drug resistance plasmids through compensatory mutations that possibly offset plasmid fitness cost [130]. A pangenome association analysis within the same study revealed 292 metabolism loci that significantly associated with ST131 compared to non-ST131 *E. coli* genomes. A follow up study in 2019 identified diversity in anaerobic metabolism genes within the MDR clade C that was not present elsewhere in the lineage [131]. Other studies have found distinct alleles of anaerobic metabolism genes in the MDR lineages ST167 [220] and ST410 [60]. This diversity was initially thought to be a signature of metabolic flexibility [43, 55], but further analysis suggests this diversity may be a reflection of loss of function mutations through recombination [40]. Metabolic streamlining is a well established evolutionary trajectory in bacterial pathogens [38, 129, 205]. To exemplify, a study in *Yersinia*, the model organism for pathogen evolution, revealed reduction in metabolic capacity as a hallmark of pathogenicity that was attributed to functional gene loss and pseudogenisation [164] .

1.7 Microbial genomics in the era of machine learning

The amount of sequencing data produced as a result of the development of NGS and TGS has spurred the exploration of new methods to analyse big data in the field of microbial genomics. Machine learning has emerged as a powerful tool, facilitating the analysis and interpretation of complex biological data. It has

been speculated that machine learning could bring a new era of data-driven, rather than knowledge-driven, science, where mining for meaningful relationships in genomic data replaces the need for hypothesis generation [93]. There are numerous machine learning algorithms that have been developed for classification problems, a common application within biological sciences. Two algorithms which are gaining popularity in this area are the random forest and gradient boosting algorithms.

Random forests were developed in 2001 [23] and are a method of ensemble learning. The term 'forest' is an intuitive name as the algorithm works by combining the votes of individual decision trees, that are weak predictors, into a composite forest of trees with high performance. Random forests are well suited to tackling biological data as they account for interactions between variables which is essential for capturing epistasis [33]. Gradient boosting algorithms are also frequently used for the interpretation of biological data [137, 149, 178]. Gradient boosting is a relatively older concept, introduced by Schapire in 1990 [173]. Also an ensemble learning technique, gradient boosting works by building decision trees based on the errors of the previous trees and can be described as 'self-correcting'.

Both models have been successfully used on *E. coli* datasets to accurately predict resistance. Moradigaravand and colleagues found that the accessory genome was the most informative feature for predicting resistance for both random forests and gradient boosting trees, although further work was not done to interpret any specific gene interactions [137]. Pearcy and colleagues went a step further, using gradient boosting to identify known resistance genes and multiple genes relating to various metabolic processes that were enriched in resistant populations [149]. These studies are of particular interest, as k-mer and SNP-based input data was used to allow identification of mutations as genetic determinants for resistance.

Other machine learning applications using mutation-focused input data have provided accurate predictions and successfully elucidated the biological mechanisms behind these predictions. Wheeler and colleagues have successfully combined random forest models with delta bitscores to identify loss of function mutations that delineate invasive and gastrointestinal *Salmonella enterica* serovars [204] and extraintestinal and gastrointestinal strains of *Campylobacter jejuni* [206]. A focus of this thesis will be to apply this same technique to identify potentiating mutations in the *E. coli* pangenome that give rise to multi-drug resistance.

1.8 Thesis aims

There are multiple successful globally disseminated MDR *E. coli* clones that are adding to the multi-faceted burden of MDR infections. Evidence from travel studies demonstrates that MDR *E. coli* are efficient colonisers. This insight provides a potential alternative phenotype to drug-resistance to target in the efforts to reduce transmission and overall success of these pathogens. Work to determine the genetic factors responsible for their enhanced colonisation capabilities is needed first to elucidate potential targets for intervention. Firstly, we hypothesise that displacement via cell-to-cell contact killing, facilitated by type VI secretion systems, is responsible for the rapid colonisation capabilities of MDR *E. coli*. Chapter 2 will test this hypothesis by:

1. Assessing the prevalence of type VI secretion system genes across the *E. coli* phylogeny. There will be further focus on the prevalence of structural T6SS genes in known MDR clones.
2. Providing the first characterisation of the T6SS loci in MDR clones.
3. Using *in vitro* competition assays to assess whether the MDR clone of ST131

out-grows *E. coli* MG1655 better than its non-MDR ancestor that possesses a differing T6SS.

We further hypothesise, based on existing literature, that genes relating to anaerobic metabolism, and metabolism more broadly, are under selective pressure in MDR *E. coli*. Chapter 3 will test this hypothesis by:

1. Developing a method of comparative pangenomics to assess whether certain STs are enriched in specific biological processes.
2. Assessing the level of metabolic variation across ST pangenomes.

and Chapter 4 will test this hypothesis by:

1. Looking for core gene loss of function mutations that differ in prevalence between MDR and non-MDR populations.
2. Assessing how well loss of function mutations can act as predictors for multi-drug resistance using machine learning models.

1.8.1 Wider importance

Understanding how MDR *E. coli* rapidly displace commensal *E. coli* and colonise hosts is critical for developing targeted approaches to obstruct transmission of MDR *E. coli*. This knowledge is necessary to tackle the growing threat these pathogens pose to global health. De-colonising or preventing colonisation by MDR *E. coli* are alternative strategies to antibiotic treatments that won't exacerbate the increasing problem of antibiotic resistance. Information gained about an already well-studied species may be applicable to other pathogens and aid tackling MDR bacteria more generally.

Chapter Two

Parallel loss of type VI secretion systems in two multi-drug resistant *Escherichia coli* lineages

The work described in this chapter is taken from the following publication:

Cummins EA, Moran RA, Snaith AE, Hall RJ, Connor CH, Dunn SJ, McNally A. *Parallel loss of type VI secretion systems in two multi-drug-resistant Escherichia coli lineages*. Microbial Genomics. 2023;9(11).

2.1 Introduction

Escherichia coli has been ranked globally as the number one causal pathogen of deaths associated with bacterial antimicrobial resistance (AMR) [140] and AMR has been declared by the World Health Organisation as a top ten global public health threat (www.who.int/news-room/fact-sheets/detail/antimicrobial-resistance). The emergence and evolution of multi-drug resistant (MDR) *E. coli* is a pressing and relevant issue for global healthcare as their extensive resistance profiles result in diminishing therapeutic options for treating infections.

A trait increasingly ascribed to MDR *E. coli* lineages is the ability to rapidly, and asymptomatically, colonise the human intestinal tract [40]. Multiple travel studies have now shown that people travelling from areas of low AMR incidence to AMR endemic regions become colonised by extended spectrum beta-lactam (ESBL)-resistant or MDR *E. coli* during travel [4, 11, 89]. Furthermore, genomic analysis has shown that the gain of MDR *E. coli* was due to the acquisition of a new MDR strain and not the preceding commensal *E. coli* becoming MDR [11]. The ability to displace and colonise may be attributed to the MDR phenotype itself, but longitudinal studies from the UK and Norway [68, 86] have shown that multi-drug resistance alone is not a sufficient driver for epidemiological success of *E. coli*. Recent metagenomic analysis has also revealed that colonisation by MDR *E. coli* does not disrupt the wider gut microbiome's composition or diversity [45].

One possible hypothesis for the ability of drug-resistant *E. coli* to displace resident commensal *E. coli* so rapidly is a result of the drug-resistant *E. coli* possessing a type VI secretion system that allows contact-dependent killing of the resident commensal *E. coli*. The type VI secretion system (T6SS) is a multi-functional apparatus that some Gram-negative bacteria possess to facilitate nutrient uptake [125], manipulation of host cells [160], and the killing of

competing bacteria [10, 170, 187, 196]. Type VI secretion systems (T6SSs) are multi-protein complexes that are situated on the inner and outer membrane of some *Proteobacteria*. The system contains a contractile sheath, that upon contraction releases an inner tube that contains effector proteins that are released into a target cell or extracellular space. This machinery has been likened to the contractile tail of bacteriophage [84]. The secretion apparatus is encoded by at least 10 genetic sub-units, the exact number of which varies depending on the particular sub-type of T6SS. T6SS distribution varies by environment [91, 197] and the presence of T6SSs varies on all taxonomic levels [1, 67, 197] with over-representation in Gammaproteobacteria [12]. T6SS genes can be gained and lost via horizontal gene transfer (HGT) [67, 170] and T6SS presence may be influenced by genomic incompatibilities, either between mobile genetic elements, donor and recipient bacteria, or via assembly of multiple T6SSs [197]. T6SSs are particularly prevalent in complex microbial communities that are host-associated [67, 91] and are commonly found in pathogens, including *E. coli* [46, 84, 117, 197]. There are multiple sub-types of T6SSs that occur and vary in prevalence between species.

T6SSs are found in most *E. coli* and can be assigned into three sub-types based on the presence and order of their genetic sub-units (Figure 2.1). All sub-types have been shown to be directly involved in pathogenesis and antibacterial activity [84]. Sub-types T6SS-1 and T6SS-2 are more commonly found in chromosomes than T6SS-3 [115]. The majority of genes encode structural components, alongside *vgrG* that encodes a spike protein and regions of variable effector proteins, some of which include recombination hotspot (*rhs*) genes (T6SS-2 only). The GC content of T6SS regions in *E. coli* varies from that of the core genome, which suggests that they may have been acquired by horizontal gene transfer (HGT) [84]. *E. coli* is a relatively understudied organism within the research field of T6SSs, with many studies only focusing on specific pathotypes [25, 46, 117].

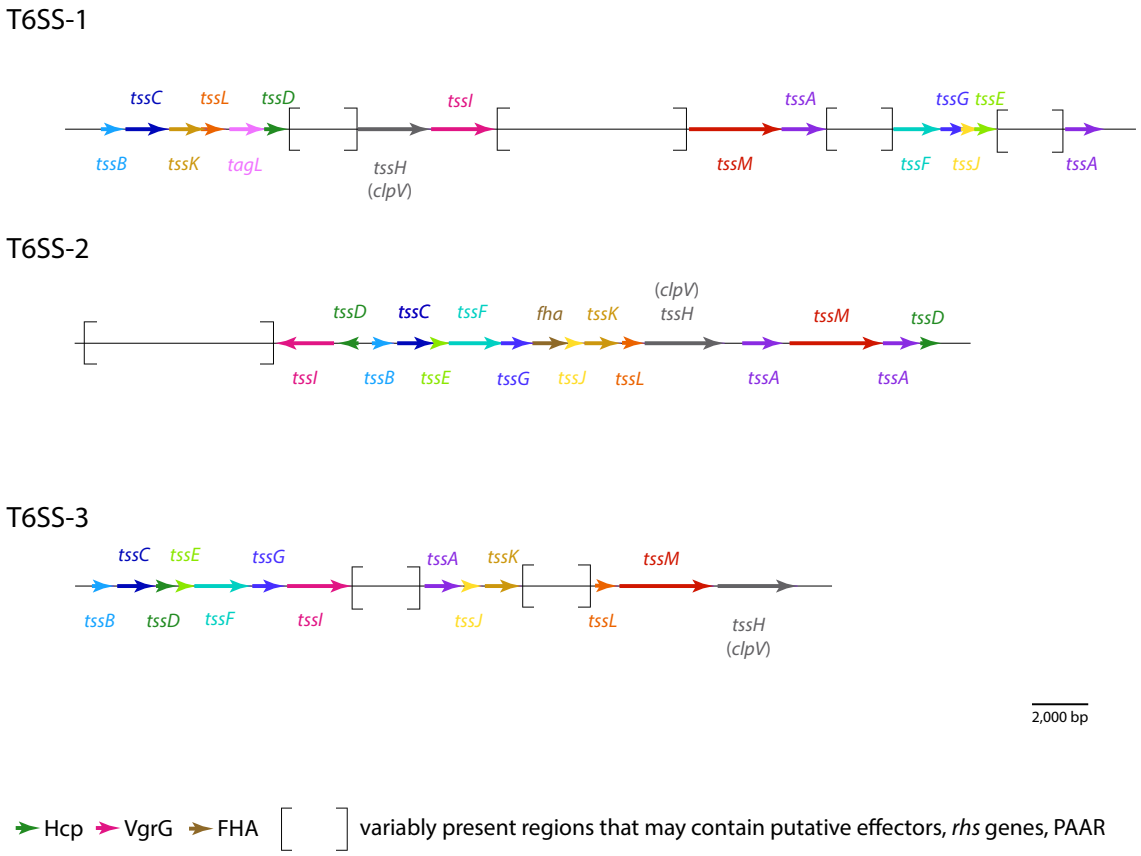


Figure 2.1: Genes encoding the three distinct phylogenetic sub-types of T6SSs found in *E. coli*. Colouring represents gene homology. Variable regions indicated by blank bracketed segments are not drawn to scale. Figure adapted from [84].

The existing connection between T6SSs and multi-drug resistance in *Acinetobacter baumanii* where MDR plasmids repress T6SSs to enable their dissemination [199, 201] makes T6SSs in *E. coli* an intriguing avenue of exploration in the context of pandemic MDR clones and how they become successful. Here, we assess the prevalence of T6SS genes across *E. coli* and identify no association with MDR, or indeed any, lineages. We characterise the T6SS loci within MDR lineages ST410 and ST131 and show that successful MDR clones ST410-B4/H24RxC and ST131-C2/H30Rx have lost a functional T6SS through deletion and insertion event respectively, thus disproving the hypothesis that MDR *E. coli* use T6SSs to rapidly displace commensal *E. coli*.

2.2 Methods

2.2.1 *E. coli* dataset curation

Existing dataset

An existing dataset of 19,571 *E. coli* genome assemblies from EnteroBase provided the majority of genome data used throughout this work. The 19,571 genomes were previously selected by Dr. Connor [40] to cover multiple highly resistant lineages (ST131, ST167, ST648) and also comparator lineages that share the same pathogenic characteristics without the MDR phenotype (ST73, ST95) [44, 124, 166]. Other pathotypes, including enterohemorrhagic *E. coli* (EHEC) and enteropathogenic *E. coli* (EPEC) (ST3, ST11, ST17, ST21, ST28) were also represented alongside the generalist and commensal lineage ST10. In total, the existing dataset covered 20 STs: 3, 10, 11, 12, 14, 17, 21, 28, 38, 69, 73, 95, 117, 127, 131, 141, 144, 167, 372 and 648. These lineages span six phylogroups A, B1, B2, D, E, and F (see Table 3.1). Phylogroup G is an intermediate group between B2 and F that was characterised in 2019 by Clermont et al. [36]. The introduction of phylogroup G succeeded the date of data collection for this work and is therefore not included in our phylogroup assignments. The diversity captured within this dataset will ensure that any observations are not unique to a specific resistant lineage and will therefore generalise to multiple resistant populations. However, the emerging MDR ExPEC lineage ST410 was not included in the original dataset, and due to the nature of its global dissemination and multidrug resistance [60, 167] we wanted to include this lineage in our analyses.

Additional 1,089 ST410 genome assemblies downloaded from EnteroBase

All available ST410 genome assemblies along with associated metadata were downloaded from EnteroBase [217] using a python script [github.com/C-Connor/EnterobaseGenomeAssemblyDownload]. EnteroBase is a freely accessible online public database of enteric bacterial genomes. The data is formed by pulling reads from publicly available sequencing repositories, such as NCBI, SRA and ENA, which are then processed using a custom pipeline to generate genome assemblies. All genome assemblies available on EnteroBase undergo quality checks to remove significantly contaminated or poor quality assemblies, by passing the following quality criteria: being between 3.7 and 6.4 Mbp, having an N50 value above 20 kb, having fewer than 800 contigs, being less than 3% scaffolding placeholders, and having above 70% of contigs assigned to a species.

In line with the methods used for the initial curation of the dataset [40], the possibility that identical isolates may be present within EnteroBase due to repeated sequencing was accounted for. The 1,089 downloaded ST410 genome assemblies were dereplicated using Mash (v1.1.1) with the default parameters [147] and a custom R script [github.com/C-Connor/MashDistDeReplication]. Mash provides a method of large scale whole genome comparison and clustering by condensing genomes into short k-mer sketches before computing a distance matrix for all isolates. The sketches contain only the most important and unique k-mers from genomes, making the data much smaller and faster to work with. The distance matrix was analysed with the custom R script to select a single representative genome from clusters of genomes with 0 mash distance. Using this established dereplication pipeline, 83 genome assemblies were removed from further analysis, adding a total of 1,006 ST410 genome assemblies to the overall dataset (Table 3.1).

2.2.2 T6SS database creation and interrogation

The SecReT6 [104, 213] experimentally validated database was retrieved and reformatted so that sequence identifiers were a suitable format for a custom ABRicate database. ABRicate (v0.9.8) was run using the default 75% minimum nucleotide sequence identity to obtain a list of T6SS genes that are present across our 21 *E. coli* sequence types. The results were summarised and partial hits (instances where a gene hit was split over multiple contigs) were accounted for and processed with a custom Python script [github.com/lillycummins/InterPangenome/blob/main/process_partial_hits.py]. A gene was determined as present in a genome if at least 85% of the gene was covered with a nucleotide sequence identity above 75%. For the focused functional gene unit analysis, any matches to genes annotated as effector or immunity protein-encoding were removed from the search results.

2.2.3 Genomic screening for insertion and deletion event sequences, bacteriocins, and phage

The following method was used to identify the prevalence of specific insertion and deletion events. Gene Construction Kit (v4.5.1) (Textco Biosoftware, Raleigh, USA) was used to annotate and visualise DNA sequences. Insertion sequences (ISs) were identified using the ISFinder database [182]. Target site duplications (TSDs) flanking ISs were identified manually. Two junction sequences specific for a given insertion were generated by taking 100 bp contiguous sequences that span the left and right ends of the insertion. Each junction sequence was comprised of 50 bp of the target IS and 50 bp of its adjacent sequence. To generate an ancestral sequence *in silico*, the interrupting IS and one copy of its associated TSD were removed manually. A 100 bp sequence that spanned the insertion point was taken to represent the naïve site. Databases of all ST131 and ST410

genomes in the collection was generated for screening with standalone BLASTn [28]. The database was queried with insertion-junction sequences and the naïve site sequence to determine whether an insertion was present, with only complete and identical matches to the 100 bp query sequences considered positive matches. Identical matches were required to pinpoint the exact IS and location of the insertion event. The same 100 bp indicator sequence method was used to identify deletion events.

Genomes were screened for bacteriocins and prophage regions using the BACTIBASE [72] online database and PHASTER [5, 216], respectively.

2.2.4 Strains and culture conditions

Three *E. coli* isolates were selected based on their phenotypic resistance profiles to allow for selective differential growth after co-culture. The strains used in this study are described in Table 2.1. *E. coli* strains were grown in Lysogeny broth (LB) at 37°C shaking or on LB agar plates. Strains were grown in the presence of 30 μg/mL kanamycin, 4 μg/mL cefotaxime or 100 μg/mL rifampicin as relevant.

The reads for MG1655-GFP (ENA accession: ERR1957959) and UTI62 (ENA accession: ERR062295) were downloaded and assembled with SPAdes (v3.12.0) with the default parameters and the careful flag. No existing sequence data was available for USVAST219. (The long read sequencing was performed by Dr. Snaith but details are included in this section for completeness). Therefore, DNA was extracted using the Monarch Genomic DNA Extraction Kit (New England Biolabs, Massachusetts, USA) before sequencing with MinION (Oxford Nanopore Technology, UK) using R9.4.1 flow cells. The data was basecalled with Guppy (v6.0.1) [github.com/nanoporetech/pyguppyclient] and adapters were trimmed with qcat (v1.1.0) [github.com/nanoporetech/qcat]. Short-read genome sequencing data was provided by MicrobesNG [microbesng.com]; DNA

was prepared using the Nextera XT kit (Illumina, San Diego, CA, USA) and sequenced on the Illumina HiSeq platform (Illumina). Long and short-read data are both available within BioProject PRJNA943186. A hybrid assembly using both the long and short-read sequencing data was created with Unicycler (v0.4.8) [209]. A complete genome of USVAST219 was needed to describe a complete T6SS locus of ST131 clade A/B as our reference region.

Table 2.1: Strains used in this study.

Strain	Lineage	Isolation	Selective resistance	Ref.
MG1655	ST10	-	Kanamycin	[21]
USVAST219	ST131-A/H41	Urine sample from Veterans Affairs Medical Centre, Michigan (2011)	Rifampicin	[39]
UTI62	ST131-C2/H30Rx	Urine agar plate from Nottingham University Hospitals (2008-09)	Cefotaxime	[41]

To generate a rifampicin-resistant mutant, 100 μL of USVAST219 overnight culture was spread on a LB plate containing 100 μg/mL rifampicin and incubated at 37 °C overnight. The following day, a single colony was chosen to inoculate 5 μL LB + rifampicin (100 μg/mL) broth. The broth was incubated at 37 °C overnight and a glycerol stock was stored at -80 °C the following day.

2.2.5 Competition assay

Overnight cultures of *E. coli* strains were diluted to an OD_{600} of 1.0, mixed at a 1:1 ratio on sterile filters and placed on LB agar plates. After four hours incubation at 37 °C, bacteria were harvested from the filter by vigorous vortexing followed by centrifugation at 12100 × g for two minutes. Harvested cells were then serially diluted and plated onto LB agar plates supplemented with the appropriate

antibiotic to enumerate the surviving *E. coli*. All plates, including samples from time point 0, were incubated at 37 °C overnight and CFUs were counted the following day. Relative fold change in growth was calculated as the four-hour growth (cfu/ml) of a strain following incubation in isolation divided by the four-hour growth of the strain following co-incubation. In a separate experiment, the supernatant was extracted from UTI62 overnight culture with a 0.2 μm sterile syringe filter. K-12 MG1655 was grown for four hours at 37 °C in the presence and absence of UTI62 supernatant to test for the presence of potentially lethal phages, bacteriocins, or excreted metabolites. Significant differences in growth were determined by one way ANOVA performed using the `f_oneway` function in `scipy.stats`.

2.3 Results

2.3.1 Distribution of T6SS genes in *E. coli*

T6SS gene presence varies between and within phylogroups and sequence types

All 20,577 *E. coli* genomes interrogated contained at least one T6SS gene from the SecReT6 experimental database [104, 213]. There was variation in the average number of T6SS genes present per ST and the average number of genes per ST within phylogroups (Figure 2.2). ST131 exhibited the largest range (2 to 29) of T6SS genes, highest number of T6SS genes in a single genome (n = 29), and highest average number of T6SS genes (joint with ST12, ST14, ST17, and ST127). The distribution of the number of T6SS genes present per genome did not align with phylogroups. STs known to contain MDR lineages, ST131, ST167, ST410, and ST648, did not possess a similar range or average of T6SS genes

(Figure 2.2). Grouping genes into functional units is therefore necessary to gain insight into how many potentially functional T6SSs are present within STs and whether this may correlate with multi-drug resistance.

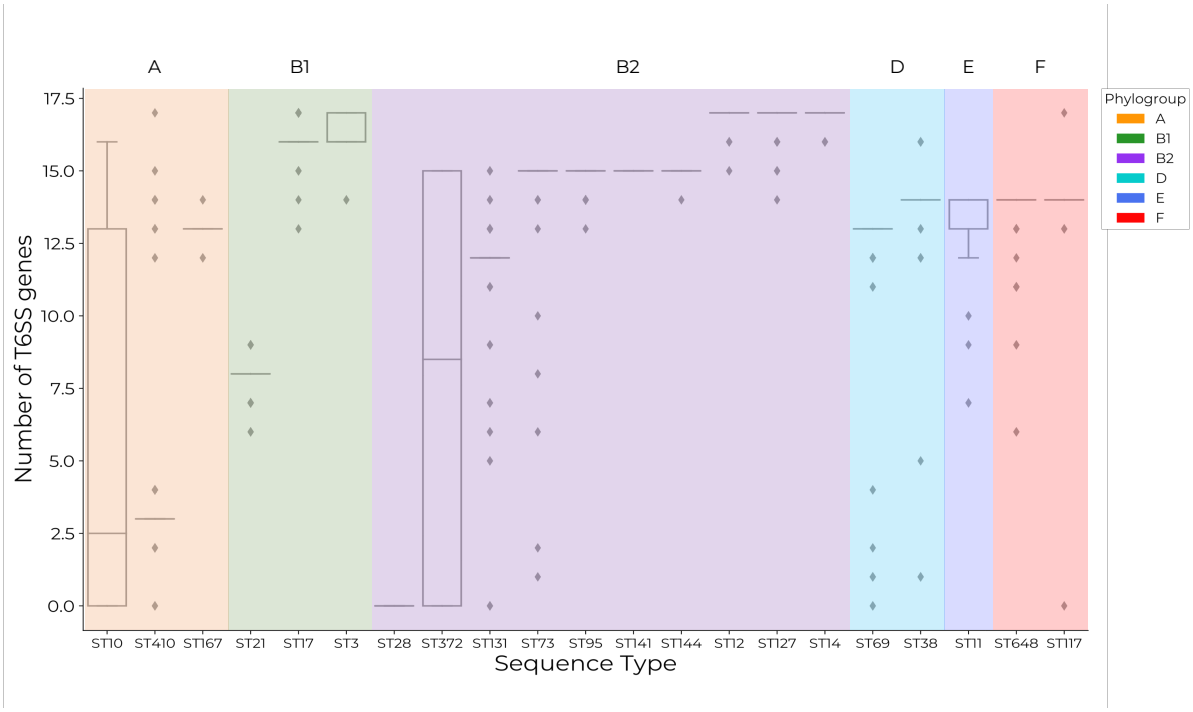


Figure 2.2: Distributions of the number of type VI secretion system (T6SS) genes from the SecReT6 experimentally validated database [104, 213] across 21 sequence types of *E. coli* shows gene carriage varies by sequence type and phylogroup. Sequence types are colour coded according to their respective phylogroups: orange = A; green = B1; purple = B2; light blue = D; blue = E; red = F. A gene is determined as present if it exceeds a DNA sequence identity of 75% and a coverage of 85%.

Presence of structural T6SS genes in lineages containing MDR clones

We examined T6SS gene units in STs containing MDR clones in greater detail. The MDR clones ST131-C2/H30Rx and ST410-B4/H24RxC were selected for further investigation due to their global distribution and clinical burden [68, 86, 151, 167]. Figure 2.3 displays the presence of structural T6SS genes within ST410 (Figure 2.3a) and ST131 (Figure 2.3b) with the corresponding clades high-

lighted on the phylogenies. Structural genes were the focus of this search because they are well characterised and vary less in comparison to effector proteins [84]. In both STs, we found a correlation between clades possessing MDR clones and the absence of structural T6SS genes (Figure 2.3).

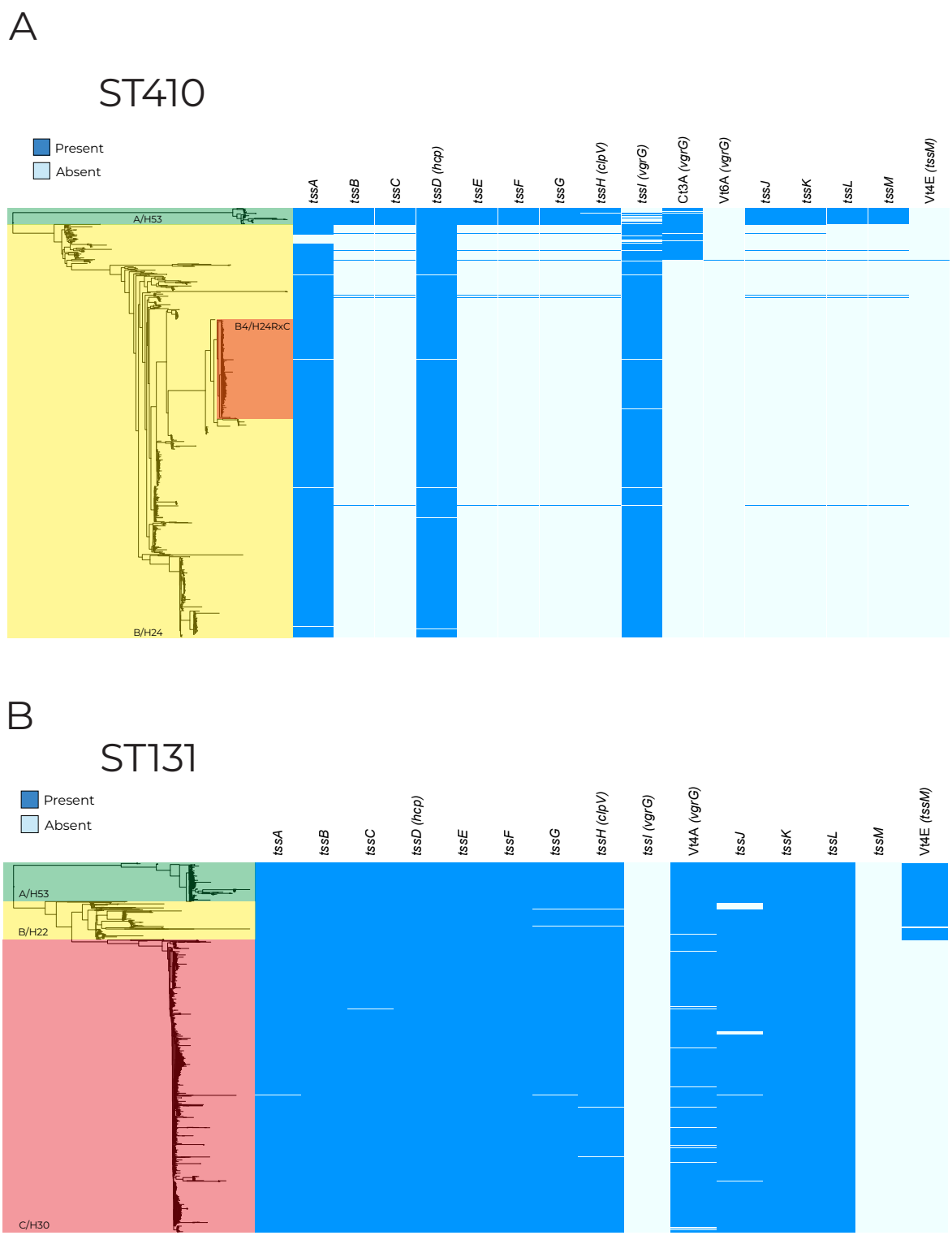


Figure 2.3: Presence of structural experimentally validated type VI secretion genes taken from the SecReT6 database [104, 213] across (A) 1,006 ST410 genomes and (B) 3,186 ST131 genomes. Phylogenetic clades are labelled clade name/*fimH* allele.

2.3.2 Genomic analysis of ST410 T6SS

Presence of structural T6SS genes in ST410

Within ST410, *tssABCDEFGHIJKLM*, (henceforth referred to as the *tss* region) was present in 91.89% (n = 34) of clade A/H53 (Figure 2.3a). The organisation of genes in this region classifies the system as T6SS-2. Of the three genomes that did not contain a full *tss* region, one genome was missing *tssH* due to poor sequence coverage (coverage < 10x for the three contigs spanning *tssH*). In the two other genomes *vgrG* was missing, caused by a combination of low coverage and the presence of multiple *vgrG* genes, which is known to lead to sequence fragmentation. We conclude that the native T6SS of ST410-A/H53 is conserved across the clade.

The *tss* region was present in just 0.83% (n = 8) of clade B/H24 genomes. Two ST410-B/H24 genomes did not contain *tssI/vgrG* and four genomes did not contain *tssLM*, but these otherwise contained all *tss* region genes. While isolates from clade A/H53 are largely drug susceptible, clade B/H24 contains the globally-distributed multi-drug resistant clone ST410-B4/H24RxC, which is highlighted in Figure 2.3a [60, 167]. In the eight rare instances where a full *tss* region was present within the B/H24 clade, five genomes contained *tss* regions most closely related (≥ 98% nucleotide identity) to that of the F-type plasmid pSTEC299_1 (GenBank accession: CP022280) when comparing to the NCBI non-redundant database. The *tss* region of two of the remaining three genomes were closely related (≥ 98% nucleotide identity) to chromosomal T6SS-1 regions from other *E. coli* (GenBank accessions: CP062855, CP091020, AP023237). These findings together suggest that, although rare, independent acquisitions of distinct *tss* regions have occurred within ST410 clade B/H24.

Characterisation of the ST410 type VI secretion locus

We resolved the structure of the complete 28.3 kb T6SS locus of ST410-A/H53 using a draft ST410-A/H53 assembly (EnteroBase assembly barcode: LB4500AA_AS), partially scaffolded (Figure 2.4) using the complete genome of ST88 *E. coli* RHBSTW-00313 (GenBank accession: CP056618). ST88 belongs to the same clonal complex as ST410 (CC23) and was used because there is no complete reference genome for ST410-A/H53 and we were unable to access a strain to generate a closed genome. The complete T6SS region contained 18 open reading frames (ORFs), 16 of which could be assigned functions (Figure 2.5). Among these were representatives of all 13 *tss* genes required for synthesis of a type VI secretion apparatus (*tssA* to *tssM*), along with two additional different *tssA* and *tssD* genes. The recombination hotspot *rhsG* gene, which is known to encode a toxin [116], is located downstream of *vgrG* and is the likely determinant for this systems' effector protein. The locus also contains a gene for an FHA-domain containing protein, which is known to be involved in T6SS activation in *Pseudomonas aeruginosa* [183]. The structure of this region is indicative of a T6SS-2 type system [84] and we conclude that this region of the ST410-A/H53 chromosome includes all determinants necessary for the production of a functional T6SS.

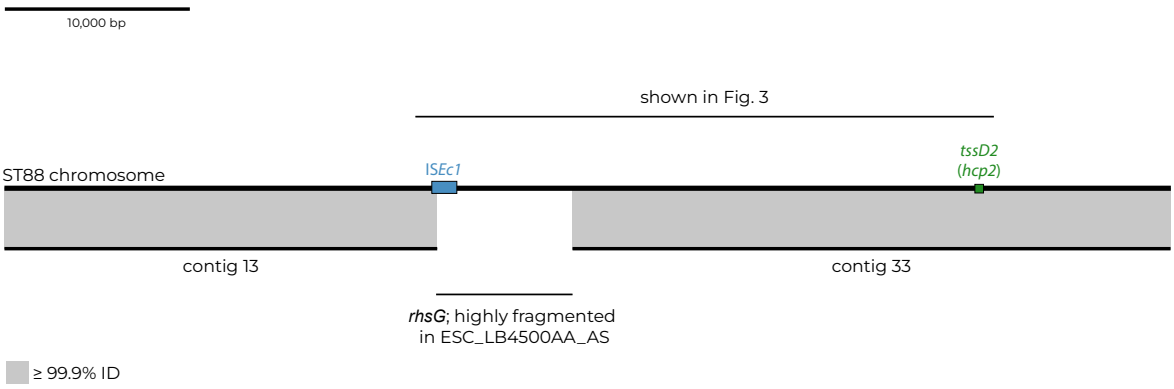


Figure 2.4: Scaled schematic showing the alignment of tss region sequences in the draft ST410 genome (EnteroBase assembly barcode LB4500AA_AS and the complete ST88 genome (GenBank accession CP056618) that was used as a scaffold when determining the structure of the region in ST410 clade A. Horizontal lines represent sequences, connected by grey shading indicative of alignments with ≥ 99.9% nucleotide identity. Contig numbers in LB4500AA_AS are indicated below, and the highly fragmented 'recombination hot spot' region is labelled. The boundaries of the tss region are marked with IS*Ec1* on the left and *tssD2* on the right, and the extent of the sequence shown in Figure 2.5 is indicated above.

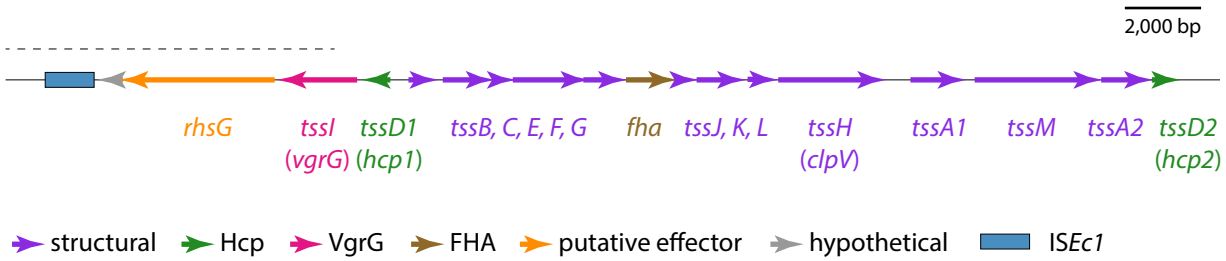


Figure 2.5: The ST410-A/H53 chromosomal T6SS locus. The extents and orientations of open reading frames are indicated by labelled arrows, with colours corresponding to protein types as outlined in the key below. The part of the sequence for which sequence from ST88 was used as a scaffold is indicated by the dotted black line above. FHA: forkhead-associated. Drawn to scale from EnteroBase assembly barcode: LB4500AA_AS and GenBank accession: CP056618.

Deletion events account for the loss of T6SS in ST410-B/H24

To explain the absence of T6SS genes in ST410-B/H24, we annotated the chromosomal region of reference strain SCEC020026 (GenBank accession CP056618) that corresponds to where the *tss* region characterised in ST410-A/H53 is located. Adjacent to the *tssA2* and *tssD2* genes, which are present in the majority of ST410-B/H24 genomes (Figure 2.6a), we found remnants of *rhsG* and *tssM* (Figure 2.6b). Relative to the complete *tss* region in clade A/H53, clade B/H24 has lost 21.0 kb between *rhsG* and *tssM* in a deletion event. As there are no mobile genetic elements present at the deletion site, which is located within a recombination hotspot, we believe the deletion event was most likely the result of recombination. To determine whether this deletion event was responsible for the loss of the *tss* region in other ST410-B/H24 isolates, we queried their genomes with the unique 100 bp sequence found at the *rhsG-tssM* junction in SCEC020026. The junction sequence (G-M) was present in the majority of ST410-B/H24 genomes (n = 941 out of 969; Figure 2.6a), indicating that this deletion event likely occurred as a single event early in the evolution of ST410-B/H24 and not as multiple independent events within this clade.

Of the 28 ST410-B/H24 genomes that did not contain the *rhsG-tssM* junction sequence, 20 were clustered in the phylogeny and also lacked *tssA* (Figure 2.6a). This suggests that a different or additional deletion event may have occurred in a small sub-clade of ST410-B/H24. To determine this, we annotated a contig from one of these 20 genomes (EnteroBase assembly barcode: ESC_HA8479AA_AS), which contained *tssD2* and exhibited a different *tss* region configuration. To confirm the structure of this region, we used the ESC_HA8479AA_AS *tss* contig to query GenBank and found an identical sequence in a complete ST410 genome (strain E94, accession: CP199740) that had been published after our dataset was assembled. This revealed the absence of a 19.8 kb segment of the *tss* region between *tssD1* and *tssA2* (Figure 2.6c).

The presence of the same *tssD-tssA* junction sequence (D-A) in all 20 clustered genomes indicated that this deletion event was responsible for the absence of *tss* genes in this sub-clade of ST410-B/H24.

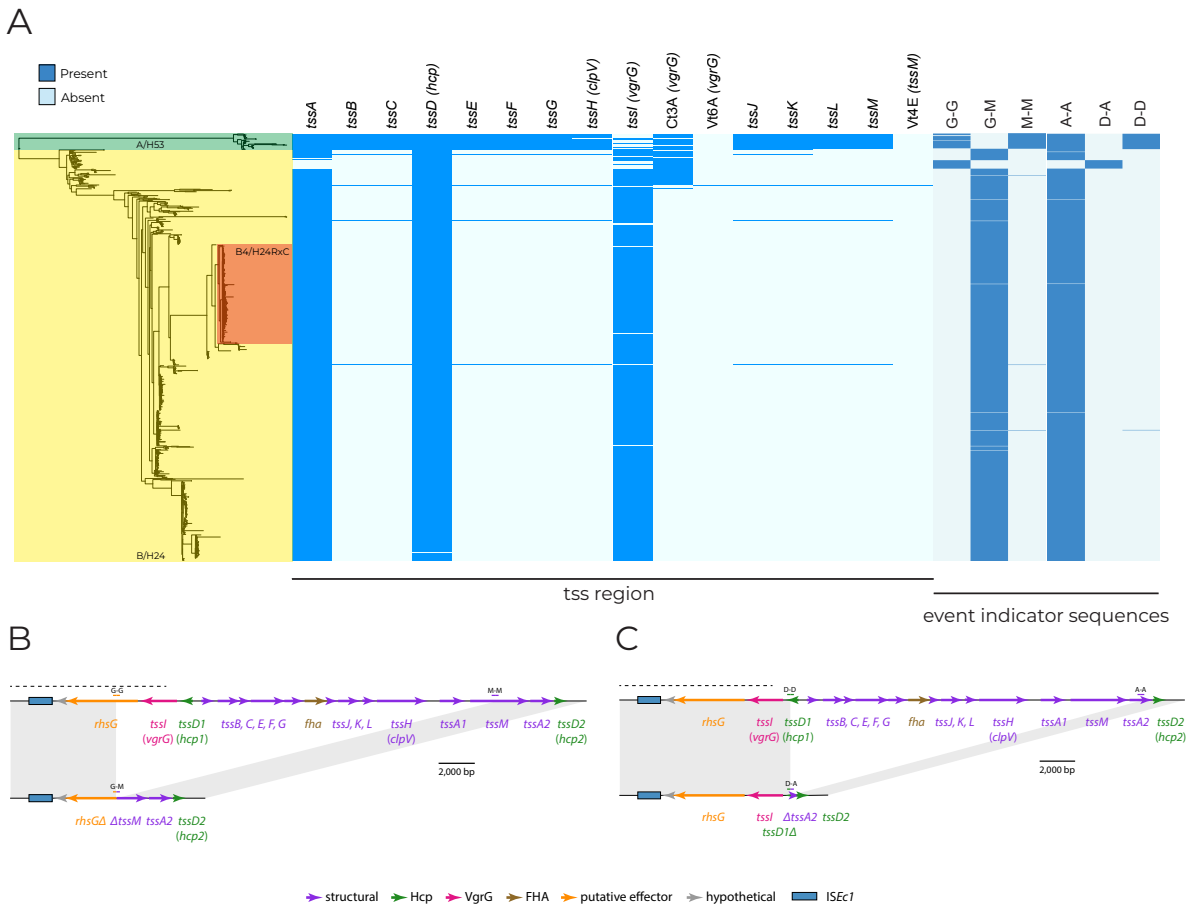


Figure 2.6: (A) ST410 phylogeny. Genomes are shaded according to their clade and *fimH* allele. The presence of *tss* region genes and 100bp deletion event indicator sequences is shown by blue shading to the right of the phylogeny. (B) Deletion of 21.0 kb segment of *tss* region between *rhsG* and *tssM*. (C) Deletion of 19.8 kb segment of *tss* region between *tssD1* and *tssA2*. Chromosomal sequence is shown as thin black lines, with open reading frames indicated by arrows beneath. Dotted black lines indicate sections of region where sequence from ST88 was used as a scaffold.

2.3.3 Genomic analysis of ST131 T6SS

Presence of structural T6SS genes in ST131

Within ST131, the ancestral and more drug-susceptible clades A/H41 and B/H22 possess a gene labelled Vt4E from the SecReT6 database, which is not present in clade C/H30 (Figure 2.3b). To deduce the function of Vt4E, we used BLASTn to search [28] the GenBank non-redundant nucleotide database, which revealed that Vt4E shared 100% sequence identity with *vasK*. The *vasK* gene is a homologue of *tssM/icmF* [96, 114] which is a component of the inner membrane complex of T6SSs [84]. The absence of a functional VasK protein has been shown to render the T6SS non-functional in *Vibrio cholerae* [96].

Characterisation of the ST131 type VI secretion locus

To facilitate comparative analyses, we characterised the T6SS-determining region of the ancestral ST131 lineage, ST131-A/H41. First, we mapped the complete and uninterrupted T6SS locus found in a 39,211 bp region of the USVAST219 chromosome (GenBank accession: CP120633) which was flanked by copies of a perfect 41 bp repeat sequence (Figure 2.7). The region contained 27 ORFs, 22 of which could be assigned functions (Figure 2.7). Among these were representatives of all 13 *tss* genes required for synthesis of a type VI secretion apparatus (*tssA* to *tssM*), along with the structural gene *tagL* and a second, different *tssA* gene. Genes for Hcp (*tssD*) and PAAR-domain containing proteins, along with three different *vgrG* (*tssI*) genes, encode the core and spike of the secretion apparatus in ST131-A/H41. Downstream of each *vgrG*, we found ORFs that encode putative T6SS effector proteins; a M23-family peptidase, a lytic transglycosylase and proteins with DUF4123 and DUF2235 domains, which have been associated with T6SS effectors [169]. This aligns with prior research

by Ma and colleagues that identified diverse effector/immunity modules within *vgrG* modules [114]. Together, the presence and order of these components classify this locus as encoding a T6SS-1 type system [84]. We conclude that this region of the ST131 chromosome includes all determinants sufficient for production of a functional T6SS.

Reference strain EC958 (GenBank accession: HG941718) [63] was used to represent ST131-C2/H30Rx in comparisons with ST131-A/H41. The T6SS regions of USVAST219 and EC958 were almost identical, differing by just seven single nucleotide polymorphisms (SNPs) (shown in Figure 2.7). The only SNP located in a structural gene (*tssH*) was silent. We therefore conclude that this chromosomal region encodes the native T6SS of ST131, and is conserved in clades A, B, and C, consistent with the presence/absence data shown in Figure 2.3b. However, the EC958 T6SS region does not encode TssM.

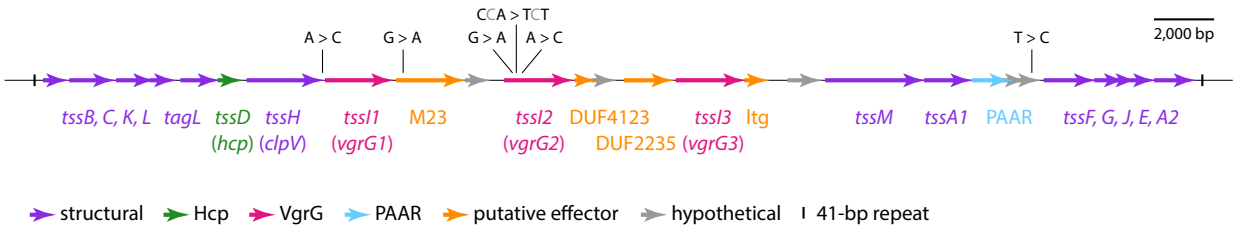


Figure 2.7: The ST131 chromosomal T6SS locus. The extents and orientations of open reading frames are indicated by labelled arrows, with colours corresponding to protein types as outlined in the key below. SNPs that differentiate clade C/H30 strain EC598 from clade A/H41 strain USVAST219 are shown in black text above. M23: M23-family peptidase, ltg: lytic-transglycosylase. Drawn to scale from GenBank accession CP120633.

tssM is interrupted by IS*Ec12* in ST131-C/H30

To determine the cause of *tssM* loss in ST131-C/H30, we examined the complete genome of reference strain EC958 [63] in further detail. We found that *tssM* had been interrupted by a copy of IS*Ec12*, which was flanked by the 5 bp target site

duplication ACTGC (Figure 2.8). The *tssM* reading frame was split approximately in half by the insertion, with 1,570 bp located to the left of IS*Ec12*, and 1,768 bp to the right. To deduce whether other *tssM* homologues were present in EC958, and could encode replacements for the interrupted *tssM*, we used tBLASTn to query the complete EC958 chromosome with the TssM sequence encoded by USVAST219. This query did not return any hits, apart from the fragmented sequences produced by the IS*Ec12*-interrupted *tssM*. It has been proven that the deletion of *tssM* prevents T6SS activity in *E. coli* [114] and the importance of *tssM* homologues for T6SS functionality in other species is well described [118, 214]. The IS*Ec12* insertion has rendered *tssM* non-functional in EC958, and in the absence of *tssM* redundancy, it appears most likely that the entire T6SS itself is also non-functional.

From the EC958 genome, we generated 100 bp sequences that span the left and right ends of the IS*Ec12* insertion (marked M-I and I-M in Figure 2.8b), and from the USVAST219 genome a 100 bp sequence that represents an ancestral *tssM* uninterrupted by this insertion (marked M-M in Figure 2.8b). To determine whether the same insertion event was responsible for the absence of *tssM* in all clade C/H30 genomes, we screened the ST131 collection for the three 100 bp sequences. The naïve M-M sequence was absent in just 2.97% (n = 21) of clade A/H41 and B/H22 genomes, confirming that almost all of them contained uninterrupted *tssM* (Figure 2.8a). 2,462 out of the 2,478 (99.48%) clade C/H30 genomes lacked the M-M sequence but contained one or both of the left and right IS*Ec12* junction sequences M-I and I-M (Figure 2.8a). Of the 16 remaining clade C/H30 genomes, three genomes contained M-M, and 13 genomes did not include either M-M or either of the IS*Ec12* junction sequences (Figure 2.8a). Manual inspection of the 13 genomes that lacked indicator sequences revealed that in six the IS*Ec12* insertion was present, but low assembly quality had resulted in truncated contigs that were too short for recognition with the strict threshold used for junction sequence detection. The remaining seven

genomes were excluded from further analysis, as low coverage appears to have resulted in highly-fragmented assemblies such that the structure of the T6SS region was impossible to determine accurately. Thus, while *tssM* is uninterrupted in 97.03% (n = 687) of ST131 clade A/H41 and B/H22 genomes, it has been interrupted by IS*Ec12* at exactly the same position in 99.78% (n = 2465) of clade C/H30 genomes. We conclude that the loss of *tssM* in ST131-C/H30 was the result of a single IS*Ec12* insertion event that occurred in an ancestral strain.

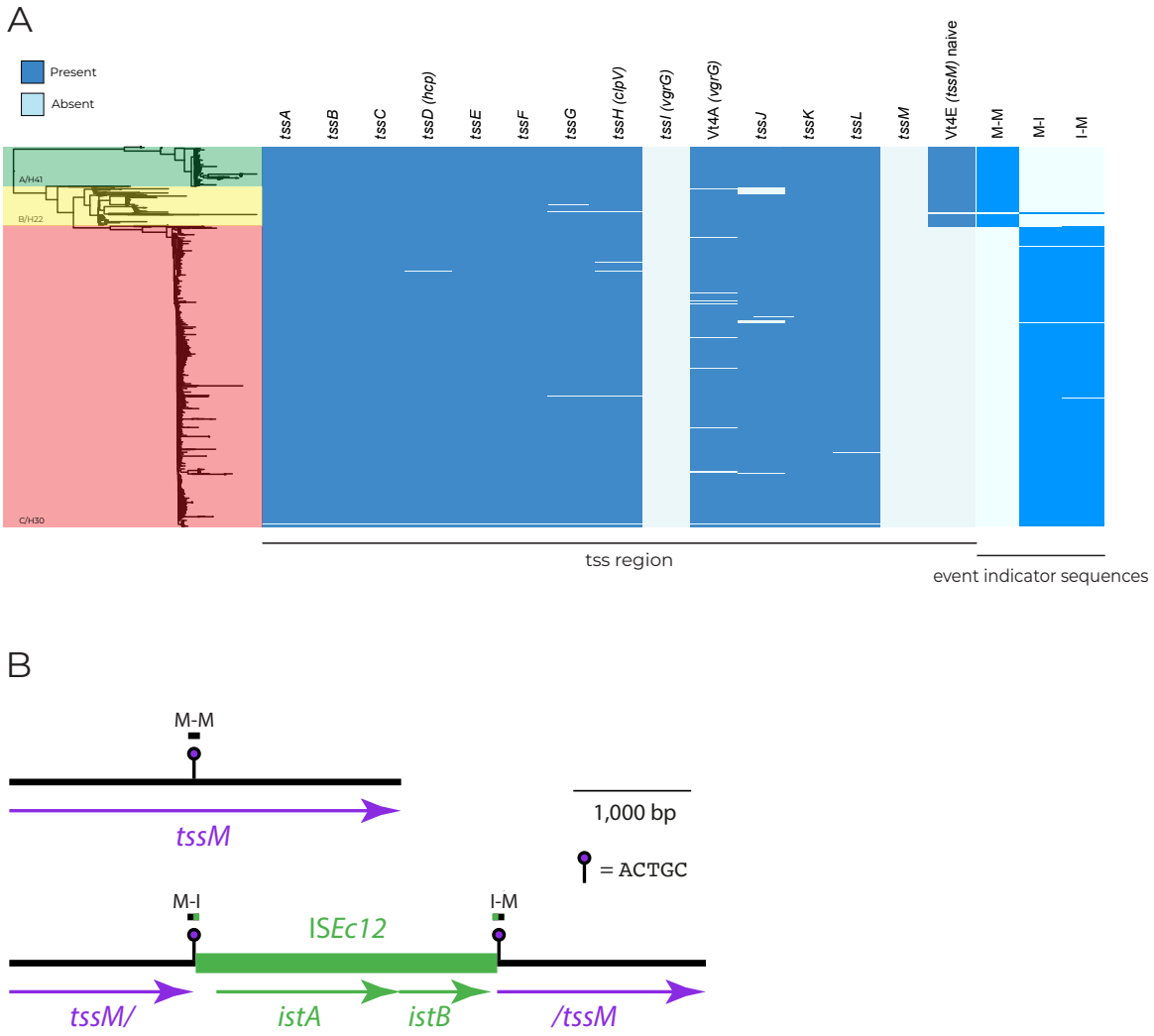


Figure 2.8: (A) ST131 phylogeny. Genomes are shaded according to their clade and *fimH* allele. The presence of the *tss* region and 100 bp IS*Ec12* insertion event indicator sequences is dictated by blue shading to the right of the phylogeny. (B) Interruption of *tssM* by IS*Ec12*. Chromosomal sequence is shown as a thin black line, and IS*Ec12* as a thicker green line, with open reading frames indicated by arrows beneath. The positions of target site duplication sequences and extents of indicator sequences are shown above. Drawn to scale using sequence obtained from GenBank accessions: CP120633 and HG94718.

The *tss* region is conserved in other phylogroup B2 lineages

Other pathogenic lineages of *E. coli* were examined to investigate whether an incomplete *tss* region was specific to clades containing MDR clones or a general feature of successful ExPEC. ST73 and ST95 were selected as comparators due to their documented prevalence in clinical settings in the UK [86, 108]. They and ST131 belong to phylogroup B2, which would also allow us to address any phylogenetic signal influencing the presence of the *tss* region. The *tss* region was found to be consistently present in both lineages (Figure 2.9a and b), demonstrating that T6SS presence varies between pathogenic *E. coli* lineages within phylogroup B2.

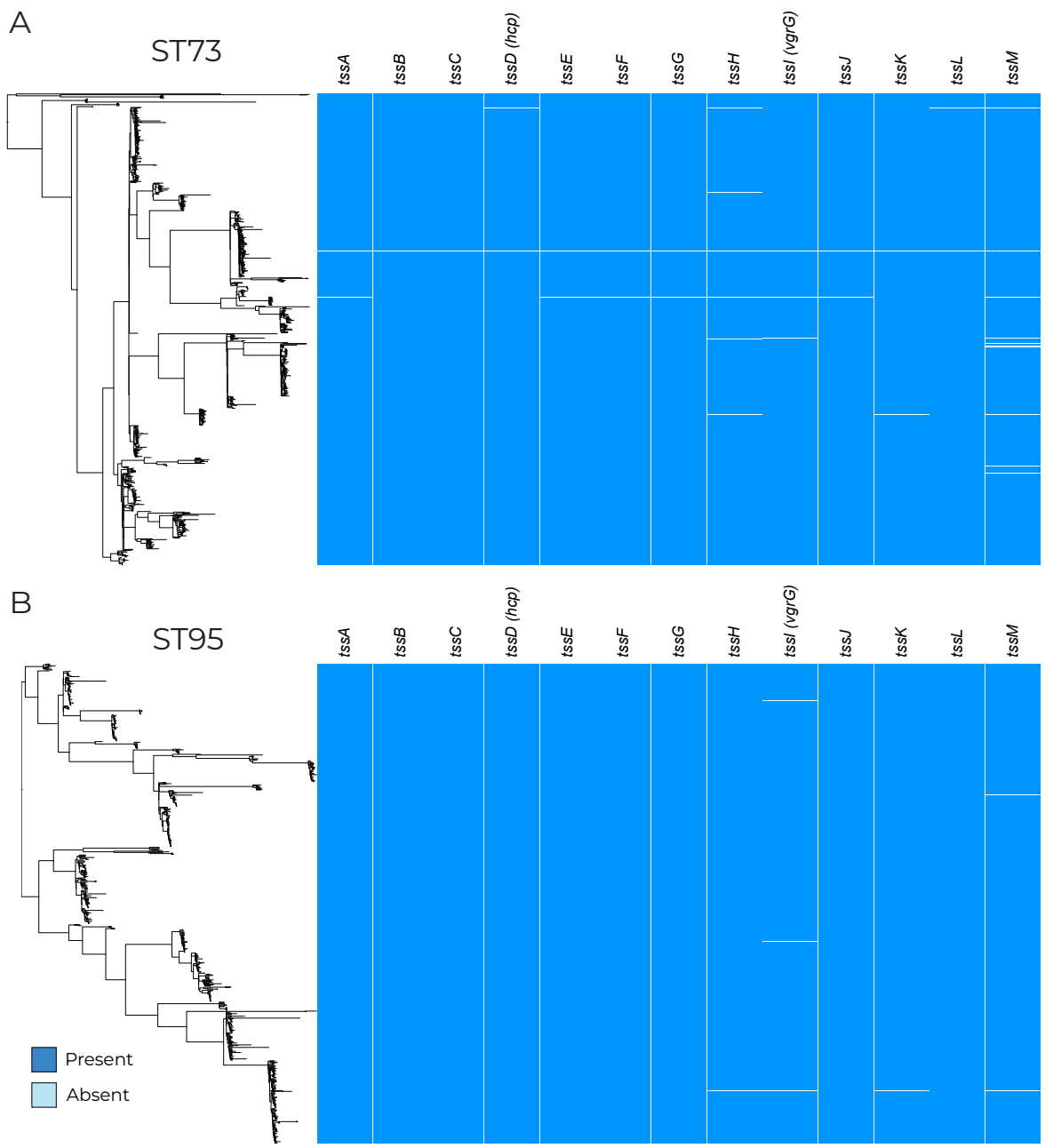


Figure 2.9: Presence of structural experimentally validated type VI secretion genes taken from the SecReT6 database [104, 213] across (A) 873 ST73 genomes and (B) 758 ST95 genomes.

The *tss* region is conserved in other MDR lineages

Other MDR lineages of *E. coli* were also examined to investigate whether an incomplete *tss* region was consistent across lineages and clades containing MDR clones . ST167 and ST648 were selected as comparators due to their documented high levels of multi-drug resistance [175, 220]. The *tss* region was found to be generally present in both lineages (Figure 2.10a and b), with no obvious correlation between phylogenetic structure and gene absence. This demonstrates that T6SS presence does not consistently dissociate with multi-drug resistance.

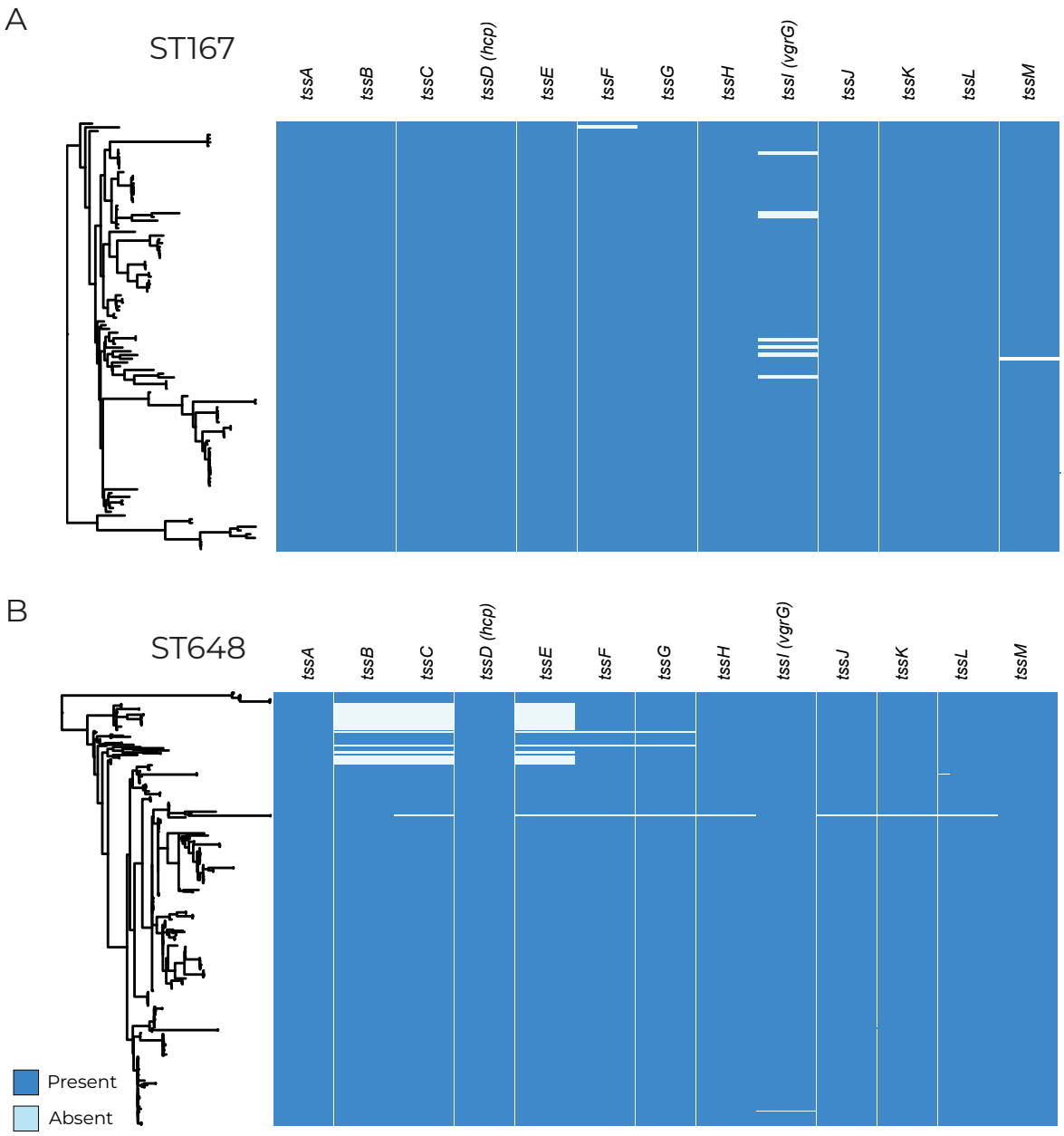


Figure 2.10: Presence of structural experimentally validated type VI secretion genes taken from the SecReT6 database [104, 213] across (A) 115 ST167 genomes and (B) 382 ST648 genomes.

2.3.4 Competitive growth of strains with and without a type VI secretion system

ST131-H30Rx out-competes without a complete T6SS

To determine whether USVAST219 (ST131-A/H41, T6SS+) and UTI62 (ST131-C2/H30Rx, T6SS-) exhibit differential T6SS-mediated antibacterial activity, we used a kanamycin-resistant K-12 MG1655 strain as prey in bacterial killing assays. Growth during co-incubation was compared to growth in monoculture for each strain in all experiments. A significant reduction in growth of MG1655 was observed after a four-hour co-incubation with UTI62 compared to monoculture (Figure 2.11a). Co-incubation of MG1655 with USVAST219 did not show a significant reduction in growth when compared to monoculture of MG1655 (Figure 2.11b). Our results show that the growth of ST131-C2/H30Rx strain UTI62 (T6SS-) out-competes MG1655 (T6SS-) *in vitro* but the ancestral ST131-A/H41 strain USVAST219 (T6SS+) does not.

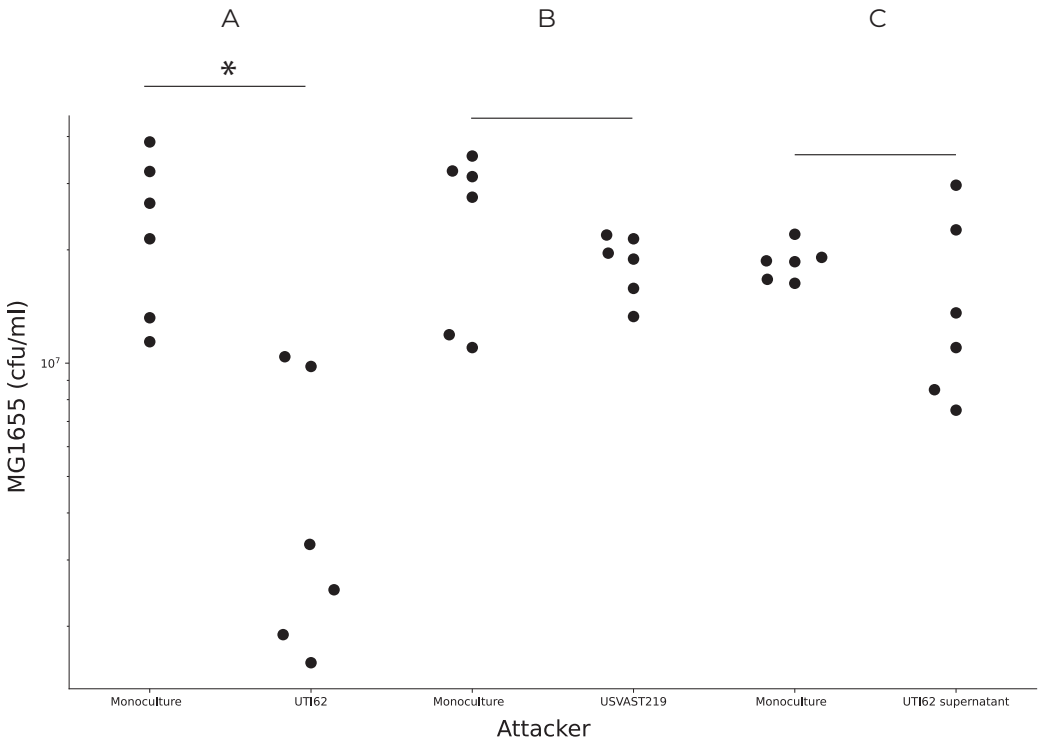


Figure 2.11: Pairwise competition assays. Measurements taken after 4 hours incubation. (A) Growth of MG1655 in monoculture (mono.) and in the presence of UTI62 (T6SS-). (B) Growth of MG1655 in monoculture and in the presence of USVAST219 (T6SS+). (C) Growth of MG1655 in monoculture and in the presence of liquid supernatant of UTI62 (T6SS-). Significance determined by one-way ANOVA, $*$: p = 0.003.

UTI62 does not produce lytic phage or bacteriocins

To further deduce the biological mechanism allowing UTI62 to out-compete MG1655, we grew MG1655 to mid-exponential phase and then challenged with the filter sterilised supernatant from a UTI62 liquid culture to test for the production of lytic phages or bacteriocins by UTI62 that could be responsible for the reduction in growth of MG1655. Growth during co-incubation was compared to growth during incubation of MG1655 in monoculture. Co-incubation of MG1655 with the supernatant of UTI62 did not result in a significant reduc-

tion in growth of MG1655 (Figure 2.11c). This contrasts with our previous result, where the growth of MG1655 was significantly reduced in the presence of UTI62 itself (Figure 2.11a). This finding suggests the production of lytic phages or bacteriocins is not responsible for UTI62 out-competing MG1655 *in vitro*. However, phages or toxins may be produced under different growth conditions. To investigate this possibility, the genome of UTI62 was screened for colicin/bacteriocin genes using the BACTIBASE database [72], and for phages with PHASTER [5, 216]. No bacteriocins were detected and none of the n = 12 prophage regions detected by PHASTER contained lysis genes. These results suggest that proteinaceous elements do not play a role in the observed reduction of MG1655 growth in the presence of UTI62.

2.4 Discussion

Extensive work has been done to characterise the T6SS in terms of its structure, function, variability and prevalence in various genera. However, this is the first study, to the best of our knowledge, to determine the prevalence of the *tss* region within MDR ExPEC lineages. We have particularly focused on ST410 and ST131 as they are known to cause infections worldwide [60, 151, 167, 185]. We have identified the evolutionary events that led to the loss of key structural T6SS components in these lineages *via* deletion or insertion events. Our study has yielded new insights on the evolution of T6SS functionality in two of the most clinically-relevant MDR *E. coli* lineages, and has provided a broader overview of T6SS determinants in other lineages.

In both ST410 and ST131, we have observed a loss of T6SS that appears to coincide with the acquisition of multi-drug resistance. This suggests we may have uncovered a common evolutionary trajectory in two MDR lineages. We have previously described the important role of potentiating mutations involved

in colonisation factor determinants, anaerobic metabolism genes, and intergenic regions as being common evolutionary trajectories in MDR clone formation [43]. The loss or degradation of a T6SS could perhaps integrate into the stepwise evolutionary course of ST410-B4/H24RxC and ST131-C/H30Rx. The structural components for a complete T6SS were chromosomal and conserved within the ancestral clades ST410-A/H53, ST131-A/H41 and ST131-B/H22. We found that the T6SS in ST410-B/H24 was inactivated by deletion events and that the T6SS in ST131-C/H30 was inactivated by an insertion event. In ST131, the rare instances where *tssM* is interrupted in clade A/H41 and B/H22, and uninterrupted in clade C/H30 (Figure 2.8a), might be explained by inter-lineage horizontal gene transfer and homologous recombination, but the T6SS region is too well-conserved (> 99.99% ID between lineages) to accurately detect recombinant sequences. In both ST131 and ST410, the inactivation of T6SS occurring prior to or during the formation of MDR clades implies that these events may have contributed to their successful emergence and expansion, but we cannot generalise to all MDR lineages as further analyses of the common MDR lineages ST167 and ST648 showed intact T6SS regions.

It has previously been suggested that acquiring multi-drug resistance plasmids is a vital step in the evolution of pandemic clones [43]. We speculate that the lack of a functional T6SS allowed both ST410-B/H24 and ST131-C/H30 to become more receptive to cell-to-cell contact, and therefore conjugative transfer. The deleterious effect of T6SS on plasmid conjugation has been demonstrated in *Acinetobacter baumanii*, where a conjugative plasmid has been shown to repress the host T6SS in order to increase conjugation rates [199]. While the *A. baumanii* experiments involved inhibition of T6SS in donor cells, the absence of T6SS in potential recipient cells seems likely to increase their permissiveness to plasmid transfer. The acquisition of conjugative multi-drug resistance plasmids by ST410-B/H24 and ST131-C/H30 might therefore have been facilitated by their loss of functional T6SS.

Our genomic analyses, combined with existing literature, suggest that the MDR clone ST131-C2/H30Rx does not have a functional T6SS due to the interruption of a single gene, *tssM*. IS-mediated insertion is known to deactivate T6SS in *E. coli* [3]. However, the functionality of the uninterrupted *tss* region in the ancestral ST131 clades A/H41 and B/H22 has not been verified experimentally. Further work to determine the functionality of the T6SS in all three ST131 clades is therefore required. Final experimental validation would require complementation of *tssM* to restore function of the T6SS in ST131-C/H30.

A caveat of our experimental work is that appropriate positive and negative controls were not included. A positive control to demonstrate that a known T6SS was capable of killing MG1655 and a negative control of a T6SS mutant should have been included to improve the robustness of our findings. The additional supernatant experiment could be improved by replacing the monoculture supernatant of the attacker with the supernatant from the co-culture of attacker with prey as it is unlikely that the supernatant from the monoculture would have the same components as those from a co-culture. An *in vivo* experiment would be better suited to capture the complex host-associated microbial community of a mammalian gut, as numerous external factors cannot be applied under standard laboratory conditions. Interactions with other species within the gut microbiome could not be replicated here, although recent research has shown that the displacement of commensal *E. coli* with MDR lineages does not disrupt the wider gut microbiome [45]. *In vivo* work may also more accurately replicate the diversity of nutrients available within the gut and any potential interactions with host cells. These factors would need to be considered to draw further conclusions on the superior displacement and colonisation capabilities that could be associated with the loss of T6SSs in MDR lineages.

The data presented here does not support the hypothesis that the ability of drug-resistant *E. coli* to displace resident commensal *E. coli* is due to the

production of a T6SS. The possible role of phage, colicins, or other diffusible elements in strain displacement are alternative explanations that should also be considered. By focusing on T6SS, we have uncovered parallel evolutionary outcomes in ST410-B4/H24RxC and ST131-C2/H30Rx, where T6SS determinants have been lost by MDR clones. Our work suggests that the loss of functional T6SS might have contributed to the evolution and success of pandemic MDR *E. coli* lineages.

Chapter Three

Comparative pangenomics identifies distinct evolutionary trajectories within sequence types

The work described in this chapter is taken from the following publication:

Cummins EA, Hall RJ, Connor C, McInerney JO, McNally A. *Distinct evolutionary trajectories in the Escherichia coli pangenome occur within sequence types*. Microbial Genomics. 2022;8(11).

3.1 Introduction

The increasing levels of resistance in *E. coli* has been well documented over recent years [68, 186], and there are now many successful lineages of MDR *E. coli* [81, 151, 167, 174, 177, 185, 190]. Multiple longitudinal studies, however, have suggested that AMR is not the driving factor of these successful MDR lineages. Kallonen and colleagues observed that the two dominant lineages of *E. coli* causing bloodstream infections in the UK, ST73 and ST131, had contrasting antibiotic resistance profiles [86]. ST73 is largely drug susceptible whilst ST131 carries numerous antibiotic resistance genes conferring resistance to multiple classes of antibiotics. Gladstone and colleagues conducted a similar longitudinal study in Norway and found that ST131 was responsible for the largest proportion of ESBL+ bloodstream infections, but that resistant and susceptible clades within this lineage were both present, again adding evidence for drug resistance not being an essential phenotype for success [68]. These studies emphasise that there are multiple strategies to achieving prevalence that is reflected by the variation in phenotypes of successful lineages.

Coinciding with rich phenotypic heterogeneity, there is no gene pool barrier within *E. coli* [69] meaning that genetic material can be freely exchanged between pathogens and commensals. Therefore, some *E. coli* STs could conceivably act as vital genetic repositories in the development of important characteristics such as pathogenicity or AMR within other STs. The extent to which any given ST, or group of STs, acts as a reservoir (genetic source) or recipient (genetic sink) in the exchange of genetic information is currently not well known. The relatively recent development of pangenomics [192] provides a useful perspective we can use to interrogate the genetic contents of different *E. coli* STs and gain further insight into how these gene collections are structured and evolve.

The pangenome represents the set of all genes present in a given pop-

ulation [24, 192]. Pangenomic studies have been performed to understand *E. coli* at the species level [69, 71, 76, 77, 162] but comparative pangenomic analyses between STs within this species can potentially add to our understanding of the evolution of the species. The influence of population structure [77] and the presence of complex epistatic relationships [207] are increasingly being acknowledged to have a major effect on the evolution of prokaryote pangenomes. Whelan and colleagues, for instance, noted that asymmetrical gene dependencies (e.g., the presence of *geneX* first requiring the presence of *geneY*, but not vice versa) cannot be uncovered by the consideration of coincident gene patterns alone. Conditional gene relationships can exist between genes, between sequence variants, or between genes and variants [52]. Inter-pangenome analysis - comparative analysis of closely related pangenomes - provides an excellent mechanism for generating prioritised lists of putative dependencies between genes. Inter-pangenome analysis can show whether a gene's classification as core in one ST is also core in a different ST. Inter-pangenome analysis can also assess differences in functional composition between closely related pangenomes. The functional contents of a pangenome (whether species-level or ST-level) reflects the biological processes occurring within the given population, such as niche adaption [105, 179], or the evolution of important phenotypes, such as drug resistance [131].

An in-depth study of an ST131 pangenome revealed clade-specific diversity in colonisation and metabolism genes in the accessory genome of the globally dominant MDR sub-lineage of ST131, clade C (H30Rx) [131]. The reported diversity was not due to the presence of unique genes, but rather the presence of unique alleles. Allelic diversity as a signature of selection has now also been observed in ST167 [220] and ST410 [60]. Allelic variation in metabolic genes has been described as an early warning sign of multi-drug resistance, potentially being a key trait of MDR clones [43]. This allelic variation could be perceived as metabolic flexibility or conversely, a signature of degradation of metabolic path-

ways. ST131 is one of the few *E. coli* STs to have undergone detailed pangenome analysis [47, 130, 131]. Understanding of *E. coli* STs on a comparative pangenome level is currently limited in terms of comparative analyses, with little known about how ST-level pangenome evolution is occurring.

In this chapter we further develop our understanding of *E. coli* pangenome dynamics and evolution by splitting *E. coli* into its constituent STs and comparing and contrasting the fates of these STs in the context of their respective pangenomes. We introduce an ST-focused approach to investigating evolutionary trends of pangenomic structure and contents, including the presence of sequence variants of metabolism genes, within *E. coli* and hypothesise that *E. coli* ST pangenomes do not evolve in a uniform way. We address the following objectives: (i) to assess whether ST pangenomes vary in structure, (ii) to evaluate the potential for STs to act as genetic sources or sinks, (iii) to establish a conservative *E. coli* core genome, (iv) to assess the link between antimicrobial resistance and metabolic gene carriage, (v) to assess whether some ST pangenomes are enriched for specific biological processes, and (vi) to assess the level of metabolic variation across ST pangenomes (given the potential link to multi-drug resistance).

3.2 Methods

3.2.1 Sampling diversity of dataset

Sample source information was collected when available, but the majority of isolates in the dataset had no source of isolate sampling data available on EnteroBase. The source sampling diversity of each ST was determined by Shannon diversity index. The Shannon diversity index for each lineage was calculated us-

ing the python package `ecopy` [ecopy.readthedocs.io/en/latest/diversity.html].

3.2.2 Genome annotation, AMR Identification

Genomes were annotated using Prokka (v1.12) [176] with the default parameters and the compliant flag. AMR gene content in individual genome assemblies was assessed using ABRicate (v1.0.1) [github.com/tseemann/abricate] and the default NCBI database (May 2023) [59]. For the purposes of this work strains were classified as MDR if their genome possessed at least one extended-spectrum beta-lactamase, carbapenemase or mobile colistin resistance gene. It is well known that ESBL-resistant *E. coli* can carry multi-drug resistance plasmids harbouring carbapenemases and *mcr* genes [89, 154, 155] and carbapenem-resistant and extended-spectrum beta-lactamase (ESBL)-producing *E. coli* are World Health Organisation priority pathogens [208]. This definition was chosen to capture human clinically relevant resistance because of the urgent need to develop new antibiotics targeting these pathogens [208], rather than the standard definition of multi-drug resistance which is not necessarily clinically relevant.

3.2.3 Pangenome assembly

Genomes were grouped by ST (using the EnteroBase classification) for individual ST-level pangenome analyses using Panaroo (v1.1.2) [194] with a 0.95 sequence identity threshold to determine homologous genes and a 0.99 core-genome sample threshold to allow the inclusion of unique core gene alleles in the accessory genome. We use the term 'gene cluster' to refer to an orthologous gene group identified by Panaroo. Linear regression was performed on the pangenome size and ST core genome size data (see Table 3.1) using Python `scikit-learn` (sklearn) [150] LinearRegression module and `statsmodels.api.OLS`

(ordinary least squares).

3.2.4 Functional pangenome annotation

The linear reference genome provided by Panaroo for each ST pangenome was split into two lists of its respective core and accessory gene clusters. The nucleotide sequence for each gene cluster was translated using a custom Python script [github.com/C-Connor/GeneralTools] to obtain a protein sequence for each cluster. These protein sequences were used to characterise gene function. Gene clusters were assigned Clusters of Orthologous Groups (COG) functional categories [189] using eggNOG-mapper (v2.0.8) bestOG assignment [79] and the eggNOG database [78] with sequence searches performed by DIAMOND (v2.0.7) [26]. Gene clusters that did not return a match within the eggNOG database were categorised under '?'. Linear regression and multiple testing corrections on the proportion of the core and accessory genomes assigned to a COG category against the average number of AMR genes per genome or the proportion of that lineage classified as MDR were performed using Python `statsmodels.api.OLS` and `stats.multitest.multipletests`.

Heatmaps and clustermaps displaying distribution of COG categories across STs were made with seaborn (v0.11.2). An ST was labelled as enriched in COG category 'X' if the percentage of ST-core genome designated to category 'X' lay above the upper quartile plus 1.5 times the inter-quartile range for all ST-core genomes in that category. Functional domain annotation was performed with InterProScan (v5) [17, 83].

3.2.5 Distribution of core genomes

Custom ABRicate databases were made for the core genome of each ST using the representative gene cluster nucleotide sequences from Panaroo and the `--setupdb` option in ABRicate (v0.8.7) [github.com/tseemann/abricate]. The bottom 5th percentile of the average coverage distribution for each set of ST-core genes was removed to ensure any incorrectly called core genes were not included in our analyses. Mass screening across all 20,577 assemblies was carried out for each ST core database with ABRicate (21 searches in total). The results were summarised and partial hits (instances where a gene hit was split over multiple contigs) were accounted for and processed with a custom Python script [github.com/lillycummins/InterPangenome]. The average proportion of gene covered for each core gene cluster per ST was calculated.

3.2.6 Core metabolic reconstructions

Metabolic models were constructed for the core genome of each ST using CarveMe [119]. Representative core gene cluster nucleotide sequences for each ST were used as input and the CarveMe algorithm was executed using the default settings. The number of metabolic reactions and metabolites in each ST core metabolic profile were counted using the Python COBRA package [57].

3.2.7 Identifying unique core metabolic genes

Metabolic reactions uniquely present in a single ST core metabolic reconstruction were extracted using a custom Python script. Unique reaction names were manually searched on the BiGG database website [92] to find the related gene names for each reaction. These related genes of interest (GOIs) were searched

for in the descriptors of the ST-core gene sequences for the ST the related reaction was uniquely present in. These sequences were combined to construct a custom ABRicate database to perform a mass screening for the GOIs across all 20,577 assemblies with ABRicate. ABRicate results were summarised and processed using the same method previously described for the distribution of core genomes. A clustermap displaying the varying presence of the GOIs across 21 STs was made with seaborn (v0.11.2).

3.3 Results

3.3.1 Dataset properties

Selected lineages span the diversity of *Escherichia coli*.

The initial dataset was curated so that multiple highly resistant lineages were included to ensure any observations were common to all resistant populations and not unique to a specific lineage [40]. The STs including MDR lineages are the well documented ST131, ST167, ST38, ST648 and additionally ST410 due to its recent documented emergence and success as an MDR lineage [167, 174]. Comparative ExPEC lineages that are not associated with multi-drug resistance but share pathogenic characteristics (ST73 and ST95), STs of additional pathotypes (EHEC, EPEC), and the commensal and generalist lineage ST10 are also included to capture the diversity of the species. In total, our dataset of assemblies retrieved from EnteroBase covers 21 STs that span six distinct phylogroups (Table 3.1).

Genomic assemblies are primarily from human sources.

The range of sampled sources per ST varied from one known sample source (ST14) to nine known sample sources (ST10, ST11, ST410). Isolates with no source data made up the majority of samples in all STs, with the exceptions of ST11, ST167, and ST410 (Fig. 3.1). The most common sampled source for these three lineages was human, which was the second largest sampled source for all other lineages bar ST117. ST117 belongs to the poultry-associated phylogroup G. Phylogroup G is an intermediate group between B2 and F that was characterised in 2019 by Clermont and colleagues [36] and is not included in the current analysis because it was unknown at the time of data collection. The Shannon Diversity Index was calculated as a measure of sampled source diversity for each ST (Table 3.1). Pangenome size ($R^2 = 0.351$, Ordinary Least Squares) nor ST core genome size ($R^2 = 0.214$, Ordinary Least Squares) were a function of sample source diversity, indicating that sampling source bias should not strongly affect our pangenome analyses.

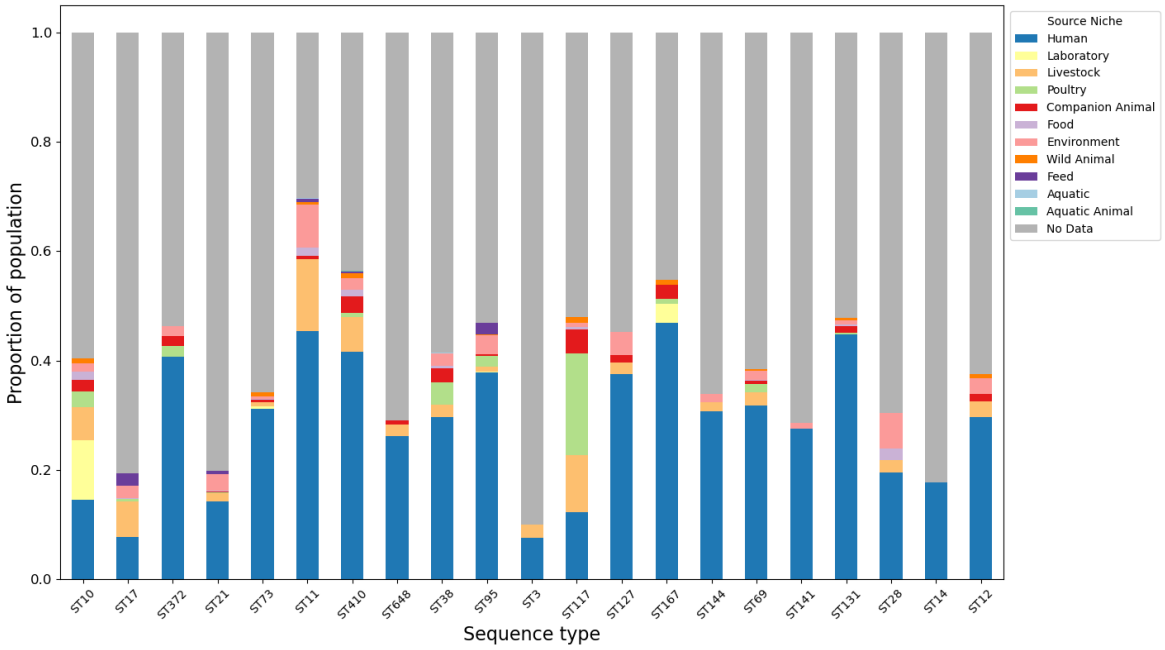


Figure 3.1: Breakdown of sampled source information available from EnteroBase for 21 *E. coli* sequence types. Graph coloured by sample source location.

High rate of AMR gene carriage is concentrated in five lineages.

The number of antibiotic resistance genes present in each genome was assessed using ABRicate and the NCBI database (May 2023). The ST167 lineage possessed the highest average of 11.79 AMR genes per genome and a high overall carriage rate with 100.00% of genomes carrying at least two AMR genes. The lowest level of AMR gene carriage was seen in ST11, with 85.73% of sampled genomes possessing only the native resistance gene bla_{EC} with an average carriage of 1.75 AMR genes per genome for this lineage. AMR gene carriage did not align with phylogrouping. We define a genome as being multi-drug resistant if it possesses at least one extended-spectrum beta-lactamase, carbapenemase, or mobile colistin resistance (*mcr*) gene. ESBL carriage co-occurring with multi-drug resistance plasmids that can carry carbapenemases and *mcr* genes is well

described [154, 155]. This definition was chosen to capture human clinical relevant resistance rather than the standard three classes definition of multi-drug resistance which is not always clinically relevant. Five STs were classified as MDR with our definition: ST131, ST38, ST648, ST410, and ST167. The threshold for multi-drug resistance is shown in Fig. 3.2 and is not phylogroup specific.

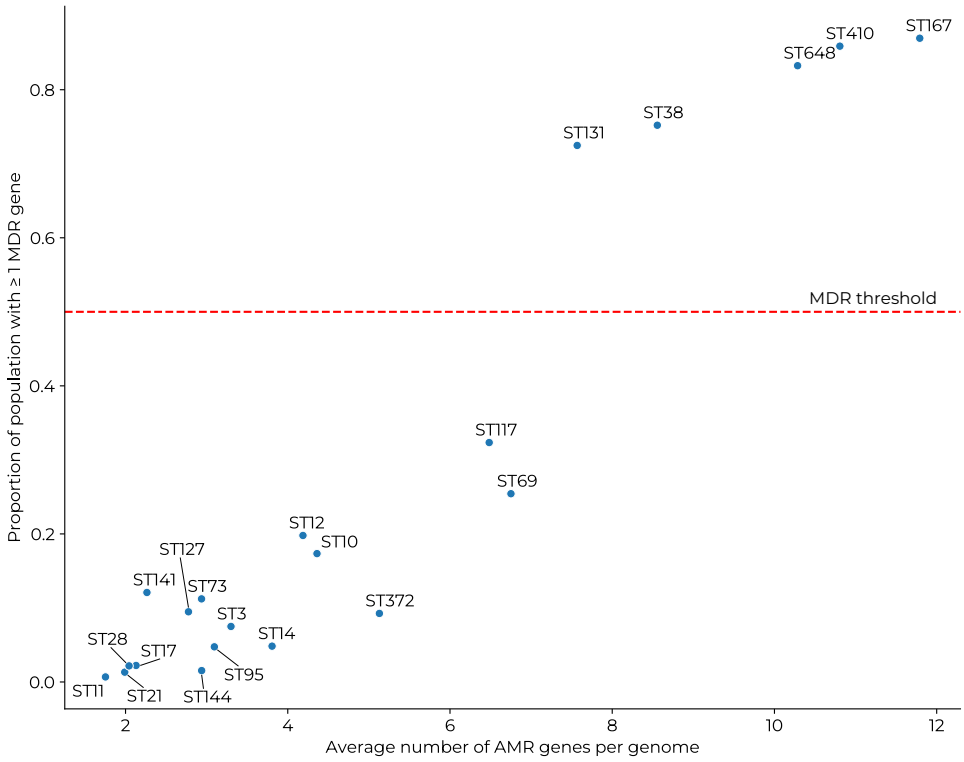


Figure 3.2: Proportion of population with ≥1 multi-drug resistance gene (extended-spectrum beta-lactamase, carbapenemase, or *mcr*) against the average number of antimicrobial resistance genes per genome. Red dotted line indicates the threshold for a lineage to be classed as multi-drug resistant.

3.3.2 Pangenome properties

Pangenome structure varies between *Escherichia coli* sequence types.

To assess the level of variation between ST-level pangenomes we first considered variation in the context of structure. We assembled the pangenomes of 21 STs using Panaroo (v1.1.2). Pangenome sizes ranged from 6,825 to 27,634 gene clusters (Table 3.1), with an average size of 11,653 and core genome sizes ranged from 3,066 to 4,058 with an average of 3,738 clusters per ST pangenome. *E. coli* is known to have an 'open' pangenome [162, 188] meaning that more genes will be added to the pangenome for every genome that is sampled, due to the diverse nature of the species. Therefore, we need to take the effect of ST sample size into account. Neither core gene number (R^2 = 0.005, Ordinary Least Squares), nor total gene number (R^2 = 0.249, Ordinary Least Squares) were a linear function of ST sample size. Consequently, ST pangenomes exhibited variation in the core gene number as a percentage of the total pangenome size, which suggests that there is no uniform pangenome structure within *E. coli*. Variation in this percentage was highest for those STs with the fewest genomes, but even when the STs with greater than 100 genomes were considered, the variation in core gene number as a percentage of the total pangenome size extended from 11.1% (ST10) to 56.86% (ST14) (Table 3.1).

Table 3.1: Summary of sampling metrics and pangenome analyses. (Pangenome size does not include paralogs). ST: Sequence type, DI: Shannon's Diversity Index.

Phylogroup	ST	Genomes	DI	Core genes	Pan. size	Core/Pan (%)
A	ST10	2,370	3.88	3,066	27,634	11.10
	ST167	115	2.74	3,675	9,035	40.68
	ST410	1,006	3.49	3,272	16,223	20.17
B1	ST17	1,884	2.14	4,003	11,870	33.72
	ST21	2,411	1.98	4,058	10,671	38.03
	ST3	40	1.46	3,749	7,933	47.26
B2	ST12	283	2.59	3,816	10,531	36.24
	ST127	232	2.64	3,891	9,696	40.13
	ST131	3,186	2.34	3,460	15,665	22.09
	ST14	62	1.60	3,881	6,825	56.86
	ST141	91	1.91	3,879	8,217	47.21
	ST144	65	2.15	3,908	6,938	56.33
	ST28	46	2.50	3,630	7,346	49.41
	ST372	54	2.51	3,752	7,514	49.93
	ST73	873	2.22	3,789	11,865	31.93
	ST95	758	2.89	3,815	11,933	31.97
D	ST38	617	3.02	3,780	14,443	26.17
	ST69	696	2.56	3,771	14,808	25.47
E	ST11	5,137	3.84	3,944	12,904	30.56
F	ST117	269	4.02	3,722	11,055	33.67
	ST648	382	2.04	3,674	11,610	31.65

ST10 has the potential to be a genetic source for other *Escherichia coli* sequence types.

Evidence for a phylogroup or ST acting as a genetic source for other *E. coli* may arise in the form of an ST pangenome possessing a low (or no) amount of unique ST-specific core genes (genes that are classified as core in only one specific ST). To this end, gene clusters that were uniquely core to a specific phylogroup were examined first. The three ST-pangenomes in phylogroup A considered together had no unique core genes (Table 3.2). Alleles including those of flagellar genes *fliDS* are present amongst the 52 B1 unique core genes. B2 ST pangenomes

possessed 99 uni-quely core gene clusters, including alleles of genes involved in central meta-bolism, *sucABCD*, fructose metabolism, *fruA*, and the decarboxylase *tabA*. The 14 unique core genes of the two phylogroup D ST pangenomes included alleles of a putative fimbrial protein, *yadNMV*. All phylogroup unique core genes with annotation are described in Appendix Table 1. Extending this analysis from phylogroup to ST, we also considered unique ST-specific core genes. The number of alleles uniquely core to a single ST varied within and between phylogroups, with ST11 (phylogroup E, n = 99), ST14 (phylogroup B2, n = 83), and ST21 (phylogroup B1, n = 74) encoding the largest number of unique core genes (Table 3.2). This is most likely due to the low sample sizes for these lineages, where sampling of greater depth would move these unique core genes into the accessory genome. Whilst phylogroup A has no unique core genes, within this phylogroup only the ST10 pangenome had no reported unique ST-specific core genes; the ST167 and ST410 pangenomes were found to have 28 and six unique core genes, respectively. The ubiquity of the ST10 core genome across all other STs may be an indicator that this ST is likely to be capable of acting as a genetic source within *E. coli*.

Table 3.2: Numbers of genes uniquely core to a single phylogroup and sequence type.

Phylogroup	Unique core genes	Sequence Type	Unique core genes
A	0	ST10	0
		ST167	28
		ST410	6
B1	52	ST3	68
		ST17	14
		ST21	74
B2	99	ST131	21
		ST28	44
		ST14	83
		ST12	10
		ST95	8
		ST144	56
		ST141	38
		ST73	26
		ST127	14
		ST372	23
D	14	ST38	23
		ST69	43
E	99	ST11	99
F	1	ST117	24
		ST648	40

3.3.3 Functional pangenomics

A 2,172 gene cluster *Escherichia coli* core genome.

The size and content of the *E. coli* core genome has been estimated in previous studies [31, 77, 85, 113], but not explicitly using a collection of genomes as large as the dataset considered in this work. Here, we provide a representative *E. coli* core gene list. There were n = 2,172 gene clusters identified that had a mean percentage coverage above 98% across all 20,577 assemblies. All pangenomes were functionally annotated with eggNOG e-mapper [78] to assign functional COG categories. Grouping these core genes by COG category showed that genes

of unknown function (category S) were the largest functional group (18.7%). A breakdown of the functional composition of the core genes can be seen in Fig. 3.3. This large percentage of species-level core genes with unknown functions highlights that, despite extensive study and characterisation, there is still a great deal of information to be uncovered regarding core genes of *E. coli*.

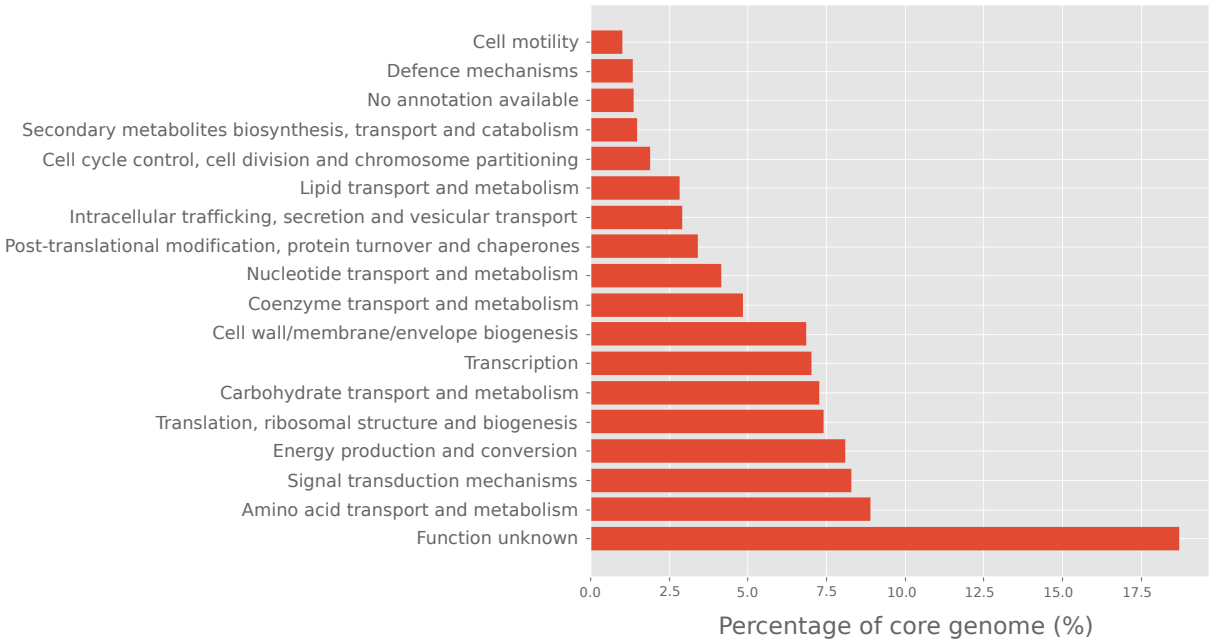


Figure 3.3: Functional breakdown of 2,172 core *E. coli* genes. Functional classes are based on COG categories.

Multiple accessory genome functions display significant correlations with AMR gene carriage and multi-drug resistance.

To determine whether any specific biological processes influence the MDR phenotype we correlated the proportion of the core and accessory genomes assigned to a COG category against the average number of AMR genes per genome and the proportion of that lineage classified as MDR (possessing at least one multi-drug resistance gene defined in §3.2.2). Significant correlations were de-

termined by performing linear regression for all COG categories with multiple testing correction.

There was no correlation between the core genome and carriage of AMR genes or multi-drug resistance for any COG category (Tables 3.3 and 3.5). The accessory genome held multiple significant correlations between COG categories and either AMR gene carriage or multi-drug resistance (Tables 3.4 and 3.6). Positive correlations were seen for genes pertaining to inorganic ion transport and metabolism (P) (Fig. 3.4a and 3.5a), signal transduction mechanisms (T) (Figures 3.4b and 3.5b), and coenzyme transport and metabolism (H) (Fig. 3.4c and 3.5c), whilst genes of unknown function (S) displayed a negative correlation with AMR gene carriage and multi-drug resistance (Fig. 3.4d and 3.5d). Additionally, amino acid transport and metabolism (E) displayed a significant positive correlation with carriage of AMR genes (Fig. 3.4e) and translation, ribosomal structure and repair (J) displayed a positive correlation with the proportion of lineage possessing at least one MDR gene (Fig. 3.5e).

Table 3.3: COG category correlation with carriage of AMR genes for the lineage core genome. Summary statistics for the correlation between the proportion of the core functional pangenome assigned to a specific COG category and carriage of AMR genes. Rows ordered by ascending p value. Rows highlighted in bold face are significant after Bonferroni correction (0.002 cut off). Correlations were assessed using a linear regression.

COG	Category	R^2 value	p value
M	Cell wall/membrane/envelope biogenesis	0.43358241	0.00234516
C	Energy production and conversion	0.3834215	0.00552213
J	Translation, ribosomal structure and biogenesis	0.14867834	0.16859644
H	Coenzyme transport and metabolism	0.14543825	0.17610001
T	Signal transduction mechanisms	0.11540568	0.26379366
L	Replication, recombination and repair	0.10602468	0.29951779
I	Lipid transport and metabolism	0.09786731	0.33470093
G	Carbohydrate transport and metabolism	0.08160357	0.41875963
P	Inorganic ion transport and metabolism	0.07225503	0.47741516
N	Cell motility	0.04837784	0.67610645
V	defence mechanisms	0.04766584	0.6834465
Q	Secondary metabolites biosynthesis, transport and catabolism	0.04372776	0.72595313
O	Post-translational modification, protein turnover, and chaperones	0.04065199	0.76159413
A	RNA processing and modification	0.03864408	0.78613861
E	Amino acid transport and metabolism	0.03479031	0.83639548
U	Intracellular trafficking, secretion, and vesicular transport	0.03140736	0.88436363
?	No functional annotation available	0.02793729	0.93792349
S	Function unknown	0.00027696	1
K	Transcription	0.02388828	1
F	Nucleotide transport and metabolism	0.00040124	1
D	Cell cycle control, cell division, chromosome partitioning	0.00942343	1
W	Extracellular structures	0.00013103	1

Table 3.4: COG category correlation with carriage of AMR genes for the lineage accessory genome. Summary statistics for the correlation between the proportion of the accessory functional pangenome assigned to a specific COG category and carriage of AMR genes. Rows ordered by ascending p value. Rows highlighted in bold face are significant after Bonferroni correction (0.002 cut off). Correlations were assessed using a linear regression.

COG	Category	R^2 value	p value
P	**Inorganic ion transport and metabolism**	**0.62137073**	**4.39×10^{-5}**
T	**Signal transduction mechanisms**	**0.60509965**	**6.62×10^{-5}**
H	**Coenzyme transport and metabolism**	**0.59179169**	**9.16×10^{-5}**
S	**Function unknown**	**0.54057051**	**0.00029264**
E	**Amino acid transport and metabolism**	**0.53295863**	**0.00034415**
G	Carbohydrate transport and metabolism	0.42485379	0.00273468
J	Translation, ribosomal structure and biogenesis	0.39215955	0.00477797
M	Cell wall/membrane/envelope biogenesis	0.35121604	0.00927926
F	Nucleotide transport and metabolism	0.32795714	0.01332518
V	Defence mechanisms	0.29847629	0.02079058
K	Transcription	0.1858726	0.1020498
?	No functional annotation available	0.15339729	0.15823227
U	Intracellular trafficking, secretion, and vesicular transport	0.14466434	0.17794088
C	Energy production and conversion	0.08482257	0.40045474
Z	Cytoskeleton	0.05560602	0.60691353
B	Chromatin structure and dynamics	0.05335077	0.62751698
O	Post-translational modification, protein turnover, and chaperones	0.05272017	0.63343148
I	Lipid transport and metabolism	0.0488452	0.67134286
N	Cell motility	0.03881647	0.78398933
L	Replication, recombination and repair	0.0360216	0.81986141
Q	Secondary metabolites biosynthesis, transport, and catabolism	0.02357216	1
D	Cell cycle control, cell division, chromosome partitioning	0.0012945	1
W	Extracellular structures	0.02172172	1
A	RNA processing and modification	0.00293211	1

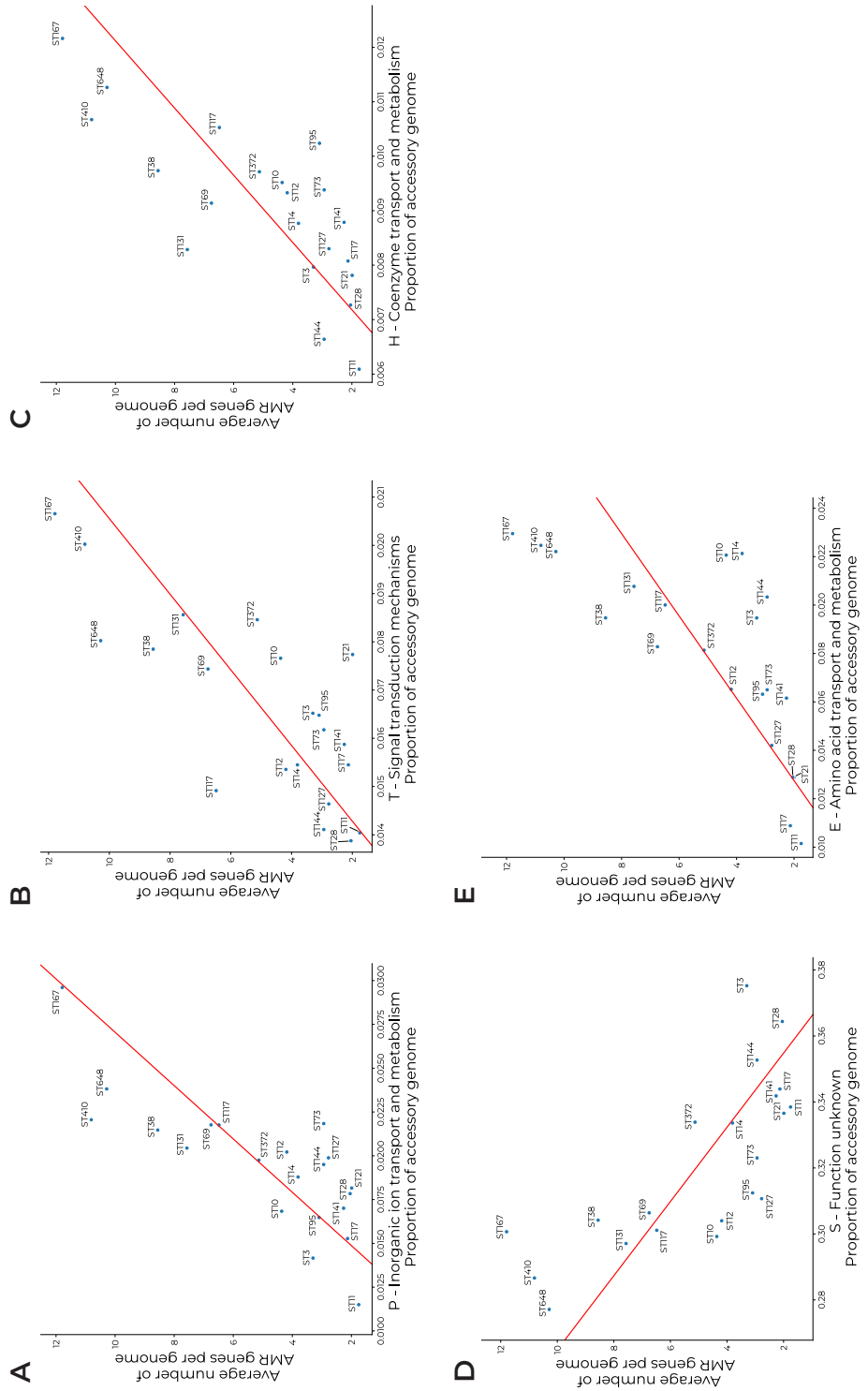


Figure 3.4: Correlation between the carriage of AMR and the proportion of accessory genome assigned to A) 'Inorganic ion transport and metabolism' (R^2 value of 0.621, P value of $4.39x10^{-5}$). B) 'Signal transduction mechanisms' (R^2 value of 0.605, P value of $6.62x10^{-5}$). C) 'Coenzyme transport and metabolism'(R^2 value of 0.592, P value of $9.16x10^{-5}$). D) 'Function unknown' (R^2 value of 0.541, P value of 0.0003). E) 'Amino acid transport and metabolism' (R^2 value of 0.533, P value of 0.0003). Red lines indicates linear regressions.

Table 3.5: COG category correlation with multi-drug resistance for the lineage core genome. Summary statistics for the correlation between the proportion of the core genome assigned to a specific COG category and the proportion of the lineage classified as MDR (population possessing at least one multi-drug resistance gene). Rows ordered by ascending p value. Rows highlighted in bold face are significant after Bonferroni correction (0.002 cut off). Correlations were assessed using a linear regression.

COG	Category	R^2 value	p value
M	Cell wall/membrane/envelope biogenesis	0.40648946	0.00375367
C	Energy production and conversion	0.3565935	0.00852179
L	Replication, recombination and repair	0.12968758	0.21761654
H	Coenzyme transport and metabolism	0.12783089	0.22312009
J	Translation, ribosomal structure and biogenesis	0.09132609	0.36608327
K	Transcription	0.08026056	0.42666801
I	Lipid transport and metabolism	0.06057976	0.56430502
T	Signal transduction mechanisms	0.06022103	0.56725534
P	Inorganic ion transport and metabolism	0.06000214	0.56906457
N	Cell motility	0.05955692	0.57276595
G	Carbohydrate transport and metabolism	0.05485834	0.61365095
Q	Secondary metabolites biosynthesis, transport, and catabolism	0.05366878	0.62456027
U	Intracellular trafficking, secretion, and vesicular transport	0.02971888	0.90983073
A	RNA processing and modification	0.02772102	0.94142428
S	Function unknown	0.00152634	1
E	Amino acid transport and metabolism	0.01153222	1
O	Post-translational modification, protein turnover, and chaperones	0.01530754	1
F	Nucleotide transport and metabolism	0.00105921	1
?	No functional annotation available	0.01984373	1
D	Cell cycle control, cell division, chromosome partitioning	0.00091333	1
V	Defence mechanisms	0.02014155	1
W	Extracellular structures	0.00504381	1

Table 3.6: COG category correlation with multi-drug resistance for the lineage accessory genome. Summary statistics for the correlation between the proportion of the accessory genome assigned to a specific COG category and the proportion of the lineage classified as MDR (population possessing at least one multi-drug resistance gene). Rows ordered by ascending p value. Rows highlighted in bold face are significant after Bonferroni correction (0.002 cut off). Correlations were assessed using a linear regression.

COG	Category	R^2 value	p value
T	**Signal transduction mechanisms**	**0.54576004**	**0.00026164**
S	**Function unknown**	**0.5353361**	**0.00032724**
P	**Inorganic ion transport and metabolism**	**0.51220435**	**0.00052897**
H	**Coenzyme transport and metabolism**	**0.45528401**	**0.00158565**
J	**Translation, ribosomal structure and biogenesis**	**0.45153412**	**0.00169826**
E	Amino acid transport and metabolism	0.40239729	0.00402347
V	Defence mechanisms	0.30922855	0.0177066
G	Carbohydrate transport and metabolism	0.28503073	0.02534908
M	Cell wall/membrane/envelope biogenesis	0.2421635	0.04689925
F	Nucleotide transport and metabolism	0.24107295	0.04762491
U	Intracellular trafficking, secretion, and vesicular transport	0.16794093	0.13008456
K	Transcription	0.16215249	0.14064304
B	Chromatin structure and dynamics	0.08827868	0.38177218
?	No functional annotation available	0.07672631	0.44828494
O	Post-translational modification, protein turnover, and chaperones	0.03944816	0.77618224
Z	Cytoskeleton	0.03534936	0.82883052
N	Cell motility	0.03395336	0.84790619
C	Energy production and conversion	0.02914772	0.91869537
I	Lipid transport and metabolism	0.02293709	1
L	Replication, recombination and repair	0.02388672	1
Q	Secondary metabolites biosynthesis, transport, and catabolism	0.00645243	1
D	Cell cycle control, cell division, chromosome partitioning	0.01054305	1
W	Extracellular structures	0.00900226	1
A	RNA processing and modification	0.00101983	1

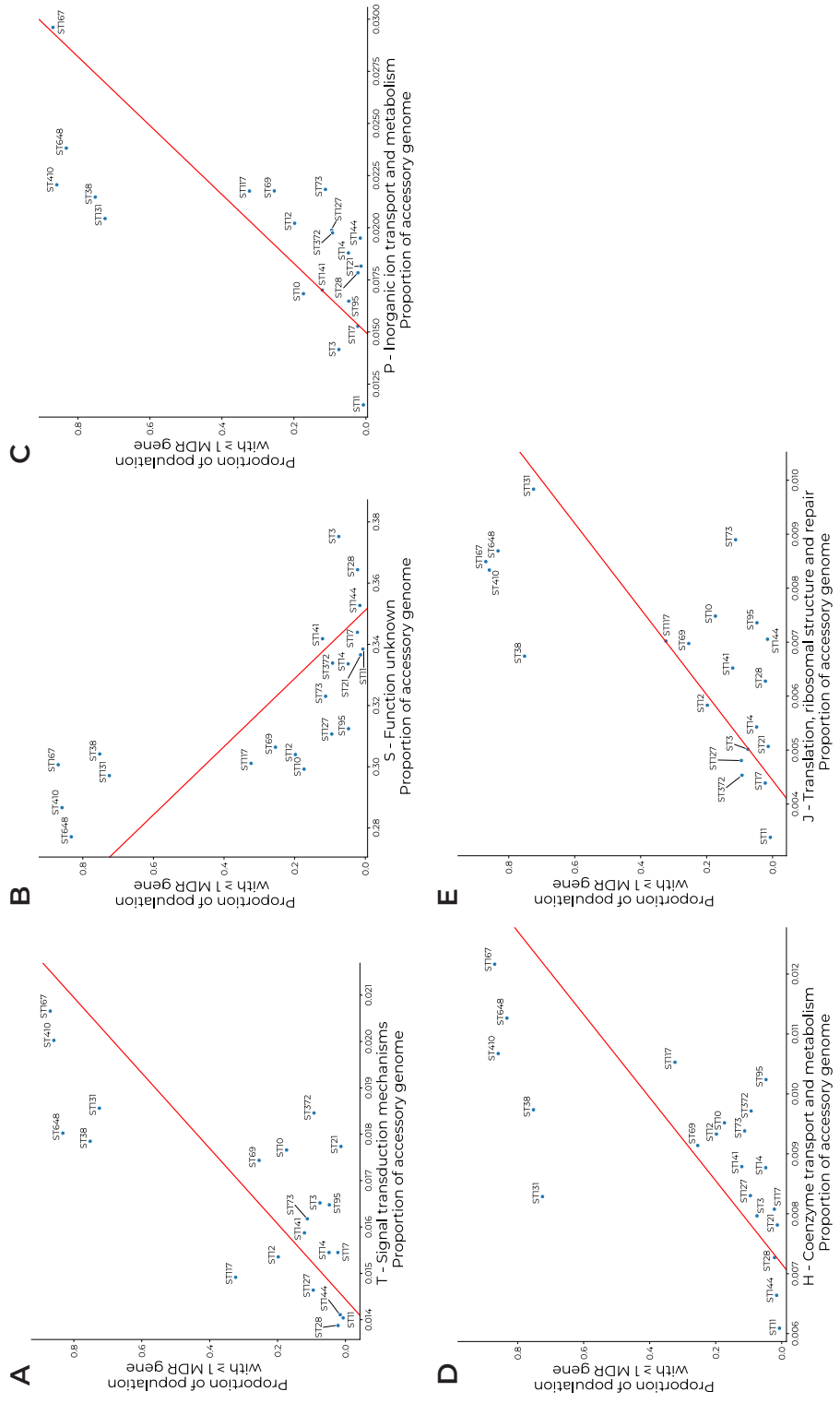


Figure 3.5: Correlation between possessing at least one MDR gene and the proportion of accessory genome assigned to A) 'Signal transduction mechanisms' (R^2 value of 0.546, P value of 0.0003). B) 'Function unknown' (R^2 value of 0.535, P value of 0.0003). C) 'Inorganic ion transport and metabolism' (R^2 value of 0.512, P value of 0.0005). D) 'Coenzyme transport and metabolism' (R^2 value of 0.455, P value of 0.0016). E) 'Translation, ribosomal structure and biogenesis'(R^2 value of 0.452, P value of 0.0017). Red lines indicates linear regressions.

ST-specific core functions vary between sequence types.

For the purposes of this study, we define three pangenome segments that are analysed and discussed throughout this work. Firstly, the 'species core genome' is the set of genes common to all genomes in this study. Next, the 'ST*X*-specific core genome' is the set of all genes considered core to ST*X*, with the species core genome removed. Finally, the 'unique ST*X*-specific core genome' is the set of genes that are found to be core only in ST*X* and not core in any other ST. These pangenome segments are conceptually visualised in Fig. 3.6. We calculated the percentage of each ST-specific core genome that was assigned to each of the COG functional categories. As we were interested primarily in functional differences between STs, category S (function unknown) and '?' (no functional annotation available) were masked from visualisation in Fig. 3.7 as they were always the largest two categories.

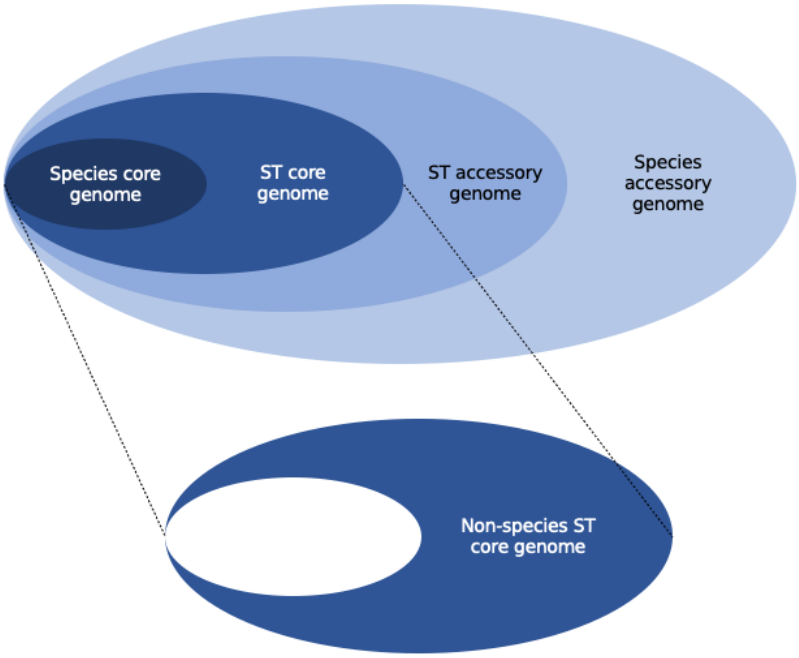


Figure 3.6: Depiction of pangenome segments used in this analysis.

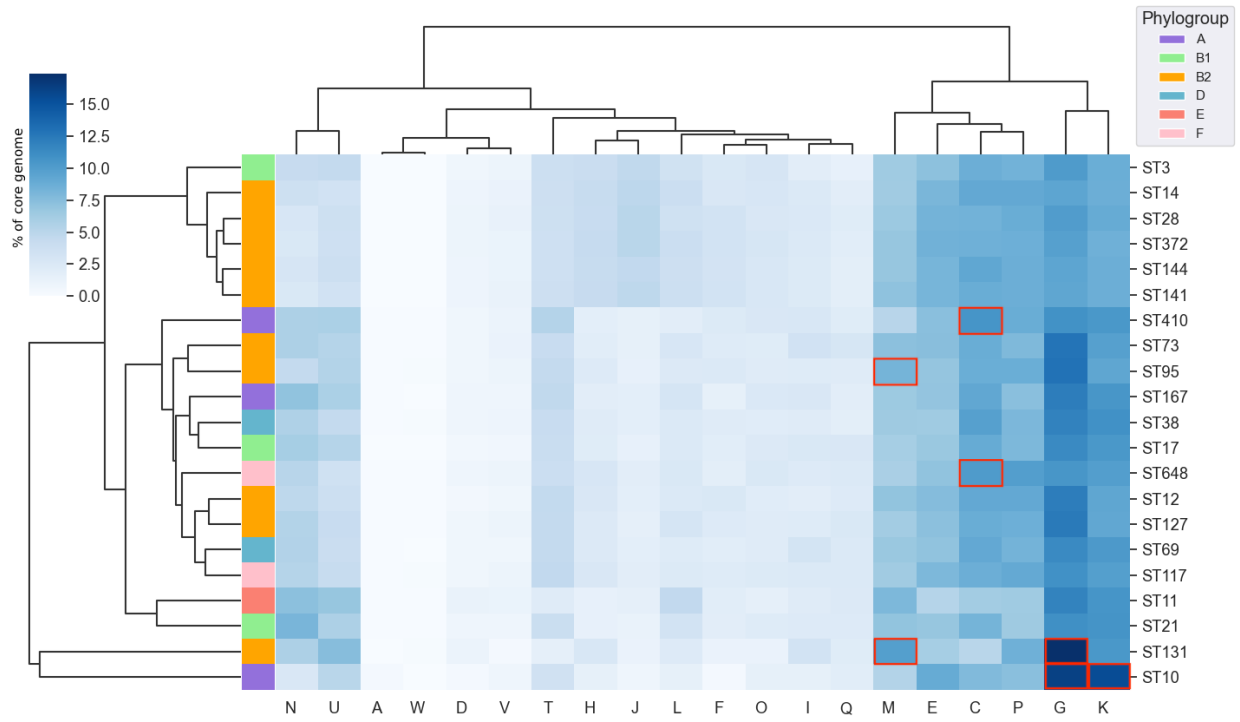


Figure 3.7: Hierarchically clustered heatmap showing percentage of ST-specific core genomes classified into 20 functional COG categories (clustered by percentage presence in core genome row-wise and column-wise using UPGMA clustering). COG categories that are enriched in a sequence type are outlined in red (percentage of ST-core genome designated to a COG category lies above the upper quartile plus 1.5 times the inter-quartile range for all ST-core genomes in that COG category). Functional COG categories: A - RNA processing and modification, C - energy production and conversion, D - cell cycle control and mitosis, E - amino acid metabolism and transport, F - nucleotide metabolism and transport, G - carbohydrate metabolism and transport, H - coenzyme metabolism, I - lipid metabolism, J - translation, ribosomal structure and biogenesis, K - transcription, L - replication and repair, M - cell wall/membrane/envelope biogenesis, N - cell motility, O - post-translational modification, protein turnover, chaperone functions, P - inorganic ion transport and metabolism, Q - secondary metabolites biosynthesis, transport and catabolism, T - signal transduction, U - intracellular trafficking and secretion, V - defence mechanisms, W - extracellular structures.

Hierarchical clustering of the percentage of each ST-core genome assigned to 20 COG functional categories highlighted the ST131- and ST10-specific core genomes as having the most distinct functional profiles, as demonstrated by the dendrogram to the left of Fig. 3.7. The accessory genomes were also functionally

categorised, however in all ST pangenomes the accessory genome was dominated by genes of unknown function (data not shown). The data show that ST pangenomes do not possess uniform core functional profiles and additionally, this observed variation is not heavily influenced by the identity of the phylogroup.

We also examined variation in ST-specific core genomes, and in particular their propensity to be differentially enriched in specific biological processes of a particular function. ST95, ST410, ST648, ST131, and ST10 exhibited functional enrichment in four COG categories in their ST-specific core genomes (see §3.2.4). ST95 was enriched in genes linked to cell membrane biogenesis (category M); ST10 was enriched in genes pertaining to transcription (category K) and carbohydrate metabolism and transport (category G); ST410 and ST648 were enriched in energy production and conversion genes (category C); and ST131 was enriched in genes pertaining to cell membrane biogenesis (category M) and carbohydrate metabolism and transport (category G). These enriched categories are highlighted in Fig. 3.7. ST11 and ST10 were also reported to be notably enriched in categories D (cell cycle control, cell division, chromosome partitioning) and A (RNA processing and modification), respectively, but the percentage levels of these categories across all STs was negligible. The enrichment of certain functional categories in specific STs suggest that genes encoding these functions may be particularly influential in these STs.

The ST131 pangenome has distinct gene presence patterns at phylogroup and species level.

Further investigation of the presence patterns of genes from ST pangenomes enriched in specific COG functional categories (outlined in Fig. 3.7) across other ST pangenomes revealed that the ST131 pangenome displays gene presence and absence patterns that are distinct from other phylogroup B2 ST pangen-

omes. The ‘mhpA’ gene cluster from the ST410 pangenome, or ‘mhpA_1’ gene cluster from the ST648 pangenome, is present in only the ST131 pangenome out of the ten B2 phylogroup ST pangenomes (Fig. 3.8, Fig. 3.9). *mhpA* encodes a 3-(3-hydroxy-phenyl)propionate hydroxylase involved in phenylalanine metabolism [211]. Similarly, the ‘mgnA_mngA_1’, ‘mngB’ (Fig. 3.10) and ‘mngR_1’ (Fig. 3.11) clusters from the ST10 pangenome are only present in the ST131 pangenome out of all B2 phylogroup ST pangenomes. In addition to the phylogroup-unique presence of clusters in the ST131 pangenome there are also gene clusters (‘rspA_1’, ‘hbp’, ‘tsx_1’, ‘fimC_1’) from the ST95 pangenome that are uniquely absent in the ST131 pangenome out of the B2 phylogroup ST pangenomes (Fig. 3.12).

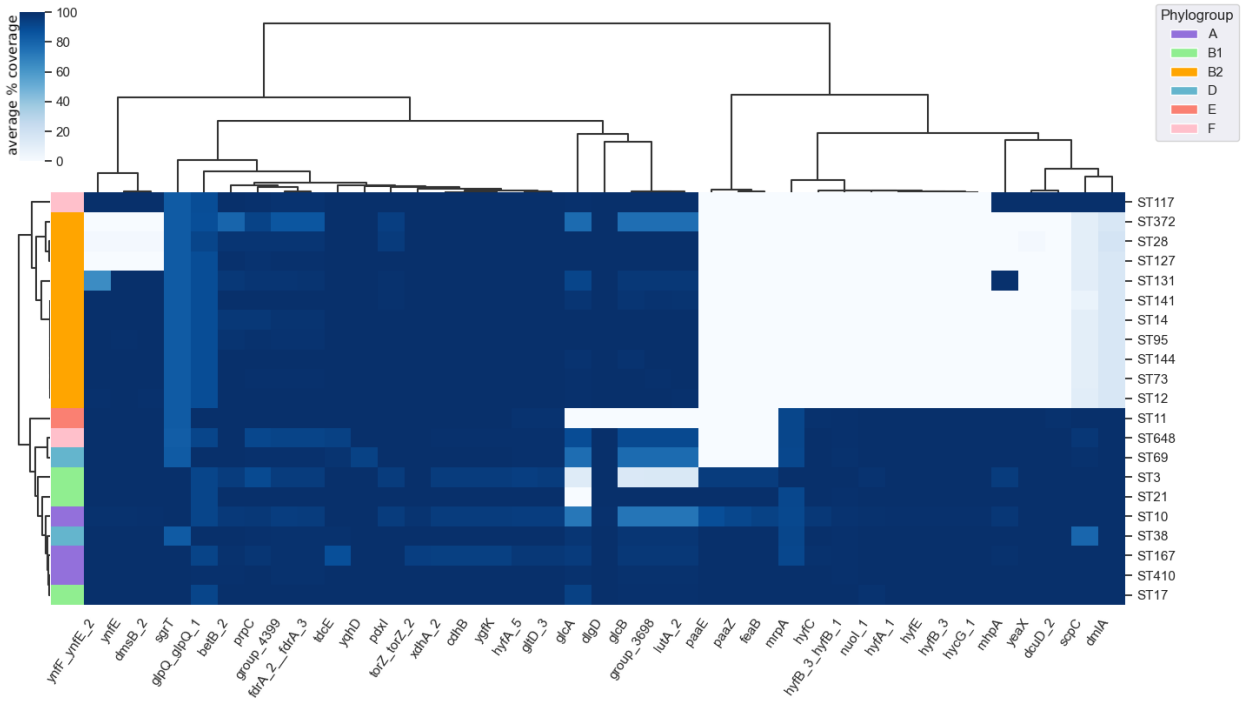


Figure 3.8: Hierarchically clustered heatmap of the average presence of energy production and conversion clusters (COG functional category C) from the ST410 core genome across 21 sequence types of *E. coli*.

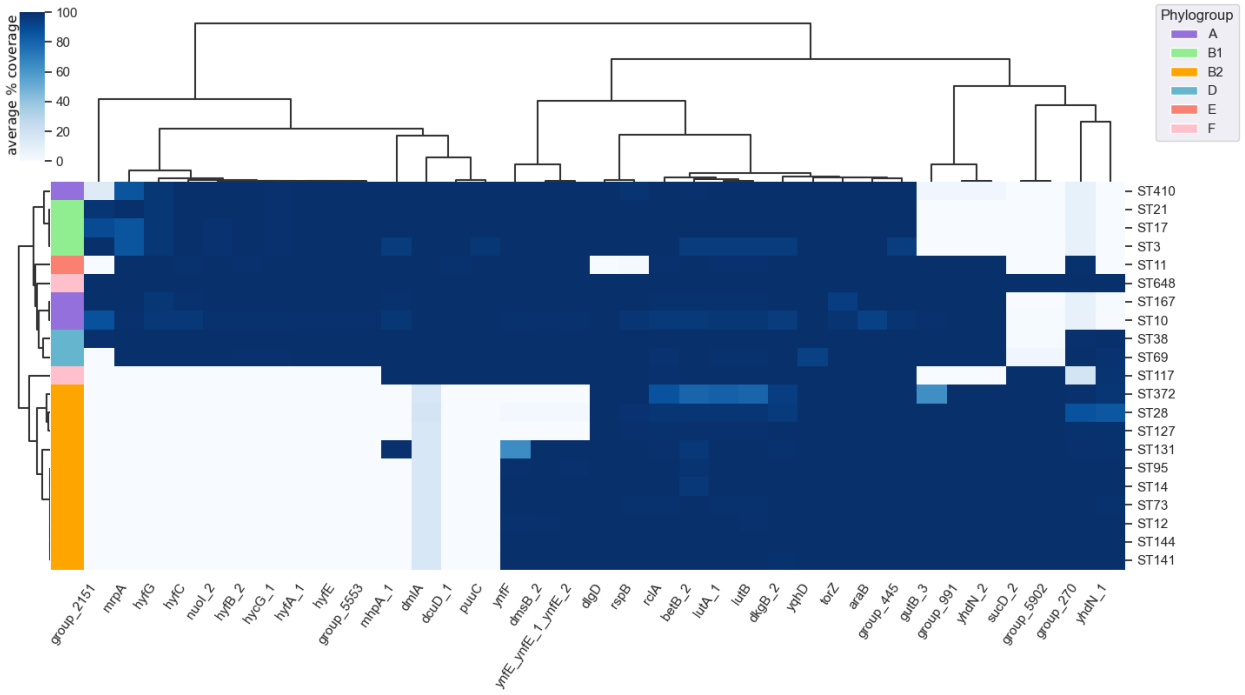


Figure 3.9: Hierarchically clustered heatmap of the average presence of energy production and conversion (COG functional category C) clusters from the ST648 core genome across 21 sequence types of *E. coli*.

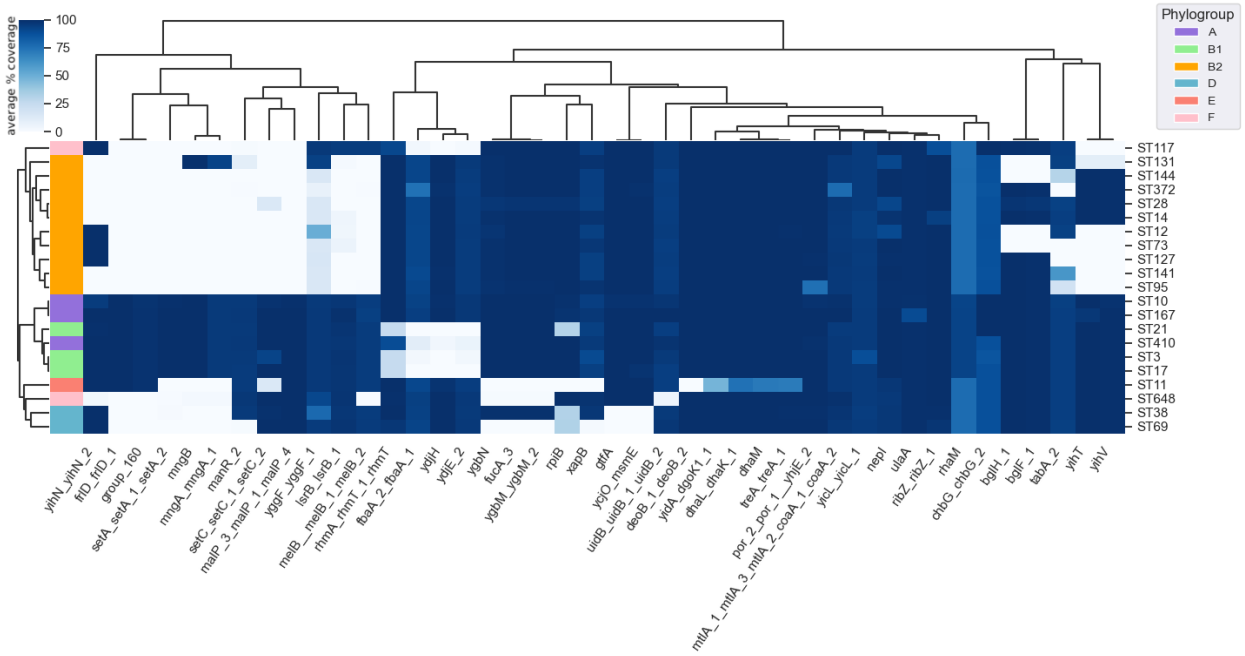


Figure 3.10: Hierarchically clustered heatmap of the average presence of carbohydrate transport and metabolism clusters (COG functional category G) from the ST10 core genome across 21 sequence types of *E. coli*.

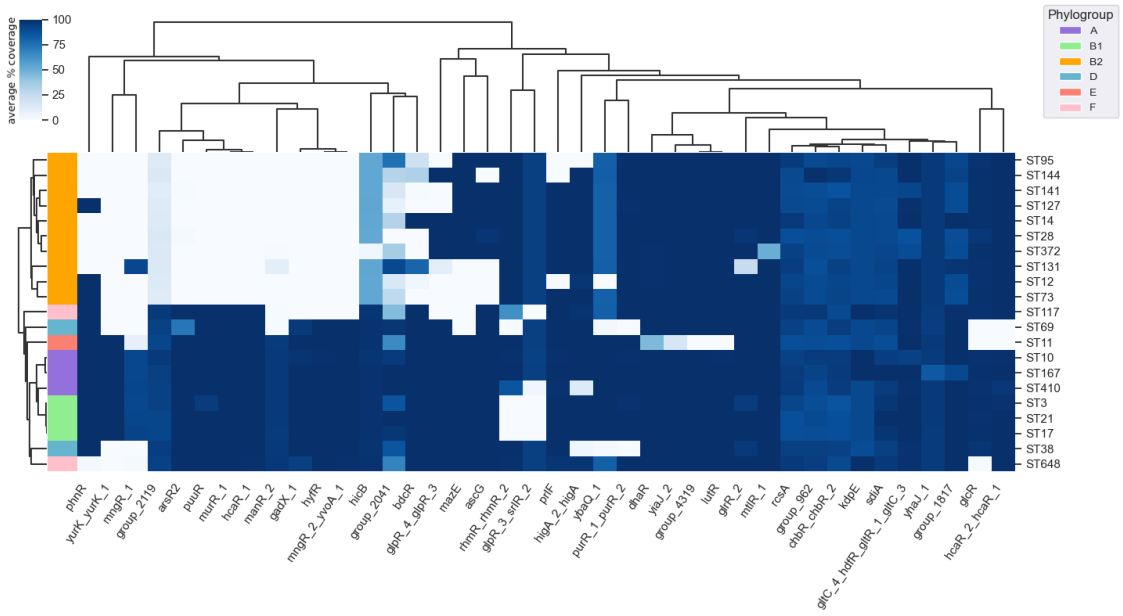


Figure 3.11: Hierarchically clustered heatmap of the average presence of transcription clusters (COG functional category K) from the ST10 core genome across 21 sequence types of *E. coli*.

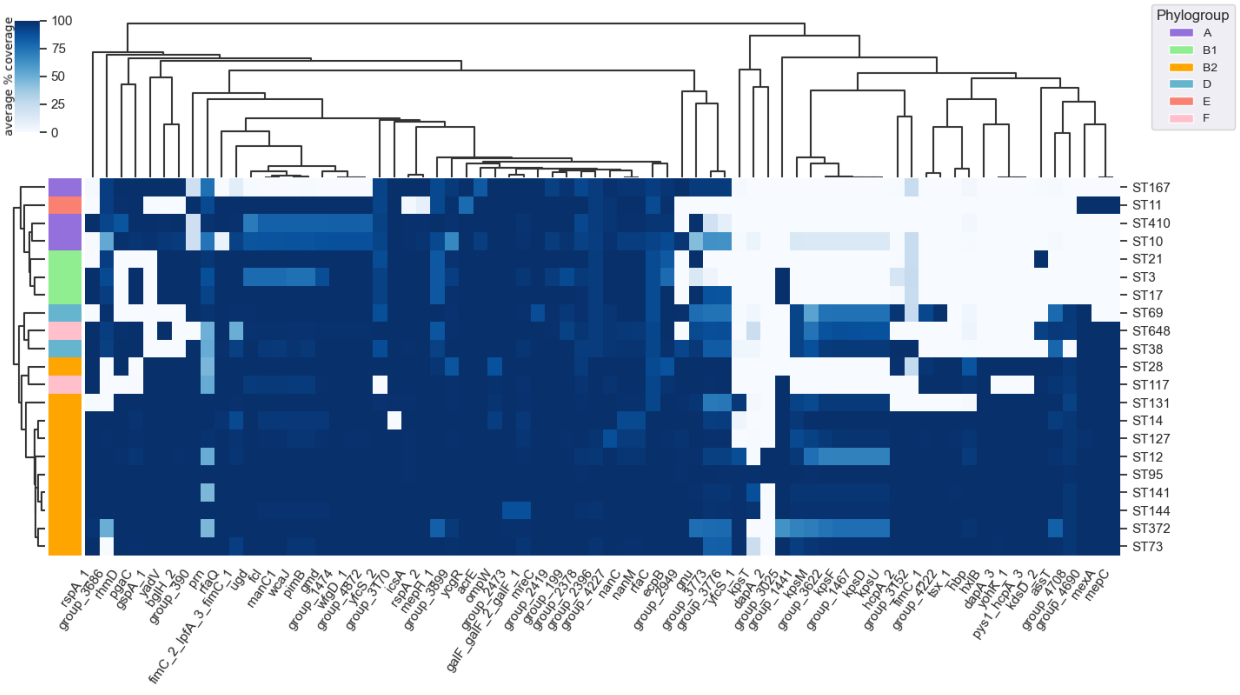


Figure 3.12: Hierarchically clustered heatmap of the average presence of cell wall/membrane/envelope biogenesis clusters (COG functional category M) from the ST95 core genome across 21 sequence types of *E. coli*.

The ST131 pangenome possesses core genes that are not seen in any other *E. coli* considered in this study. Notable presence patterns within ST131's enriched carbohydrate transport and metabolism core genes (Fig. 3.13) include the uniquely present 'group_3501' and 'yihP_yicJ_3_yicJ_1' gene clusters. These clusters were not detected in any other ST pangenome. Functional annotation of 'group_3501' suggests this gene encodes a glycosyl hydrolase and eggNOG provided *xylS* as an annotation for this gene cluster. In the KEGG orthology database [87, 88], *xylS* is synonymous with *yicI* [146]. With this connection, we postulate that these two gene clusters of the *yicJI* operon, uniquely present in the ST131 pangenome, are involved in the same xyloside metabolic pathway. The nucleotide sequence was searched against the uniprot [6] database resulting in a top hit of 85.3% similarity to a putative glycosyl hydrolase from *Citrobacter rodentium* and *Citrobacter freundii*.

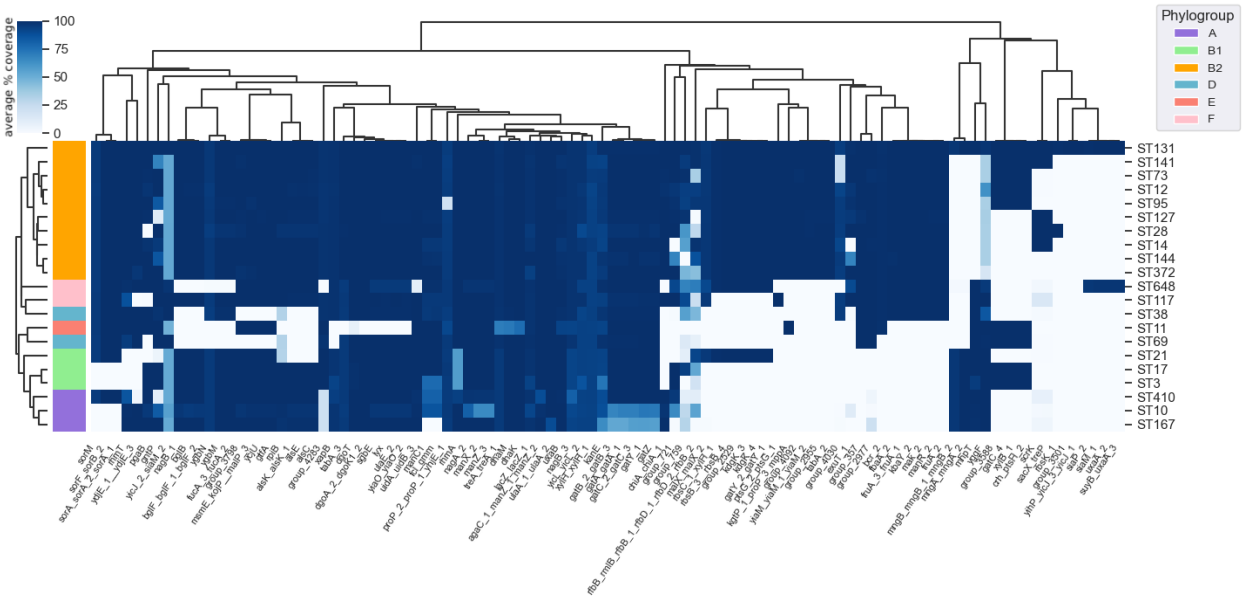


Figure 3.13: Hierarchically clustered heatmap of the average presence of carbohydrate transport and metabolism clusters (COG functional category G) from the ST131 core genome across 21 sequence types of *E. coli*.

3.3.4 Allelic variations of metabolic genes

ST131 and ST10 pangenomes possess multiple alleles related to carbohydrate metabolism and transport.

We have identified two ST-specific core genomes, ST10 and ST131, that are enriched in carbohydrate metabolism and transport (category G) genes. To investigate whether this enrichment was related to metabolic allelic diversity, we explored the presence of alleles within the category G genes in the ST131 and ST10 pangenomes. The ST131-specific core genome includes n = 100 gene clusters linked to carbohydrate metabolism and transport (category G), of which 64% are indicated to have multiple alleles present, as different gene clusters, in the ST131 pangenome (Appendix, Table 2). These include, but are not limited to, *manRXZ*, *sorABFM*, *malPX* and *gatABCYZ*, involved in the mannose, sorbose, maltose, and galactose phosphotranserfase systems [20, 145, 172, 202]. The ST10-specific core genes in category G (Appendix, Table 3) that have multiple alleles present across the ST10 pangenome include *mngAB*, involved in mannose transport and metabolism [171] and sugar efflux transporters *setAC* [109]. A full summary of the number of genes present as multiple alleles per enriched COG category is provided in Table 3.7. Beyond multiple alleles of carbohydrate metabolism genes being present across the pangenome, certain genes were present as multiple alleles within the ST131 core genome. The ST131-specific core genome possessed two alleles of each of the following genes: *fruA*, *gatC*, *kdgK*, *nagB*, *tabA*, and *uxaA*. Multiple alleles of carbohydrate metabolism genes were not present within the ST10-specific core genome, but there were multiple alleles present of three transcription genes: *glpB*, *hcaR*, and *mngR*.

Table 3.7: Numbers of genes and distinct alleles per enriched functional COG category.

Sequence Type	Enriched COG category	Genes	Distinct alleles in ST pangenome	Distinct alleles in ST-core genome
ST648	C	35	15	2
ST410	C	39	16	6
ST10	K	40	20	7
	G	42	28	0
ST131	G	100	64	12
	M	57	21	2
ST95	M	72	18	8

Further investigation revealed that the fructose phosphotransferase system (PTS) gene, *fruA*, beyond being present as two alleles in the ST131-specific core genome (as clusters 'fruA_2' and 'fruA_3_fruA_1') was in fact present as four gene clusters across the ST131 core genome; 'fruA_1' and 'manP_fruA_4_fruA_1' clusters were found to be present in the species core genome. To investigate the functionality of these four gene clusters, functional domain analysis was carried out using InterProScan [83]. Three distinct functional domains corresponding to the PTS system EIIA, EIIB, and EIIC components [168] were identified within the four *fruA* clusters. 'fruA_1' possessed both EIIB and C, whilst 'manP_fruA_4_fruA_1' and 'fruA_3_fruA_1' possessed only EIIC (Fig. 3.14). 'fruA_2' encodes the PTS EIIA component that is associated with *fruB* rather than *fruA* [49]. This discrepancy is likely attributable to the Panaroo cluster naming algorithm rather than being of biological significance. The presence of these four gene clusters is shown in Fig. 3.14. The three correctly named *fruA* clusters are distinct (greater than 5% sequence divergence), highly conserved, seemingly functional, and encode non-truncated peptide sequences which suggests annotation error, degradation, or pseudogenisation are unlikely to be responsible for this multiplicity.

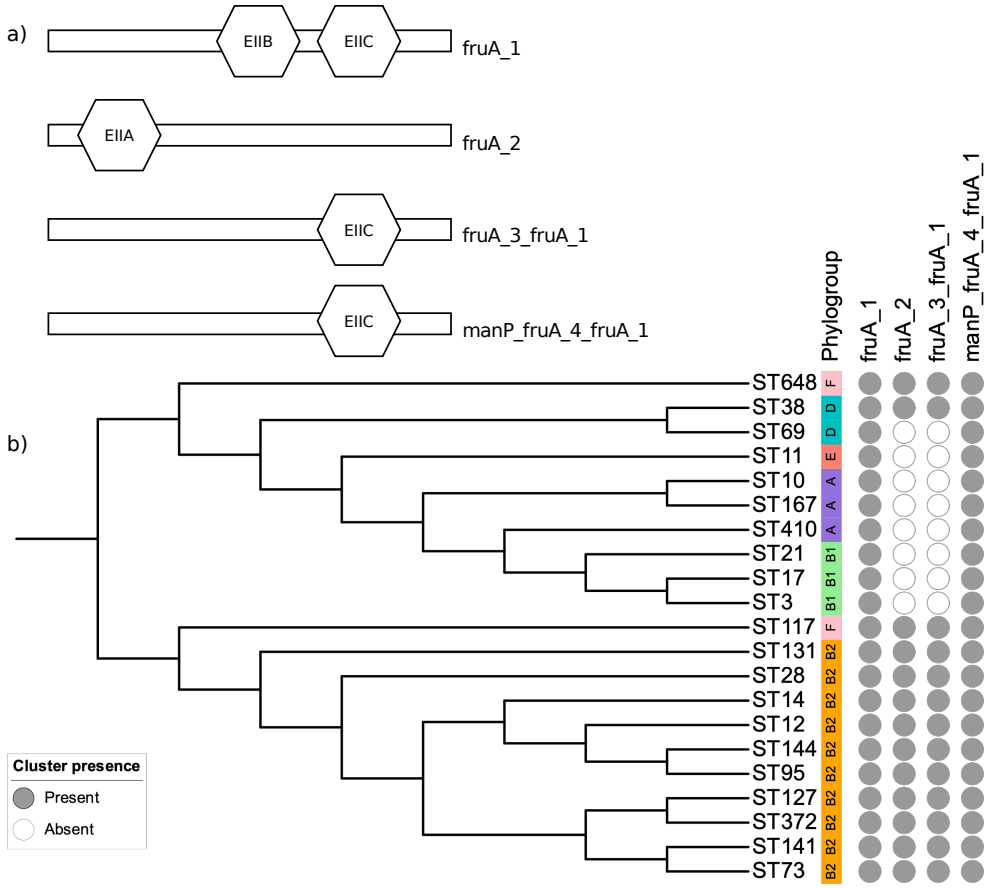


Figure 3.14: a) Phosphotransferase system components encoded by four *fruA* annotated clusters and b) cluster presence across an *E. coli* phylogeny, created using iTOL [103].

Alleles of metabolism genes vary in core status across sequence types.

To examine metabolic diversity within ST-core genomes in more detail, we performed metabolic reconstructions for the core genome of each ST pangenome using CarveMe [119] to create a 'ST core metabolic profile' so that gene information could be extrapolated to utilisation and specific metabolic pathways. Comparison of the 21 ST core metabolic profiles uncovered 825 metabolic reactions that were found in at least one ST core metabolic profile, but not common to all STs (*i.e.* not a species-core reaction). We focused further analysis on metabolic re-

actions that were uniquely present in a single ST core metabolic profile. Tracing uniquely present reactions within the BiGG database [92] back to their related gene names, and then searching for these names within our dataset, led us to a subset of gene clusters with non-ubiquitous presence patterns (Fig. 3.15). These selected clusters were *fhuA* (iron acquisition) [61], *pduCDEF* (propanediol utilisation) [18], *mntH* (manganese transport) [123], and the hydratase *crt* [200].

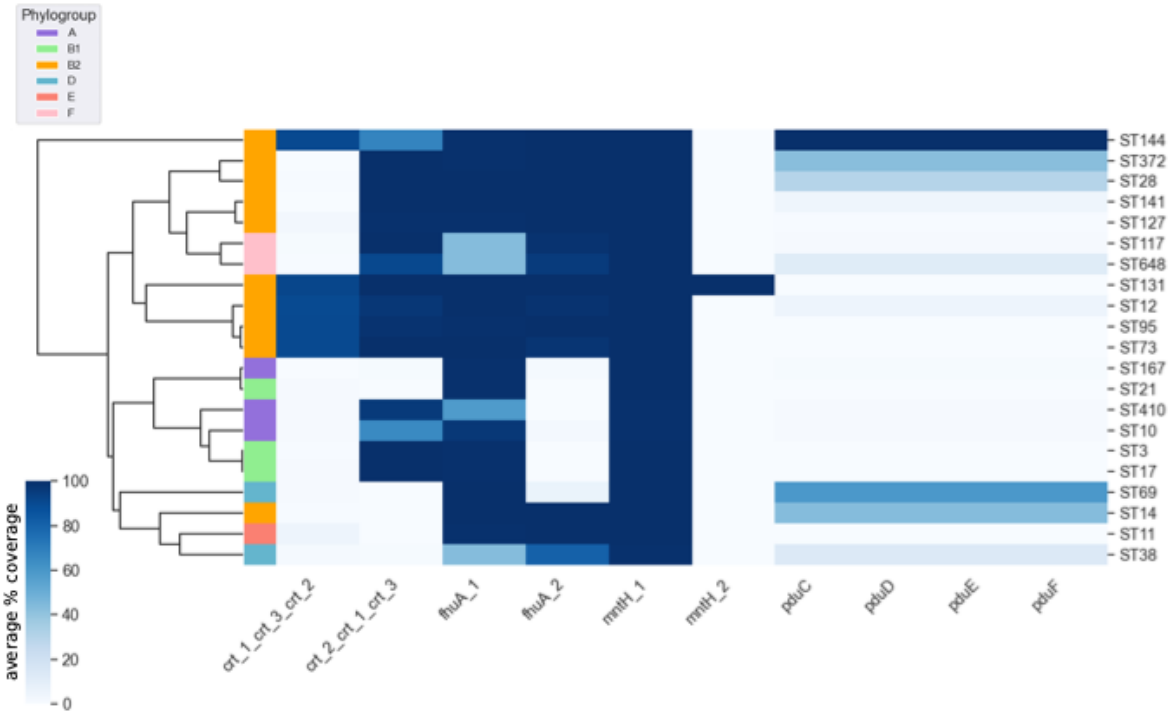


Figure 3.15: Hierarchical row-wise clustering of the average presence of *crt*, *fhuA*, *mntH* and *pduCDEF* gene clusters across 21 sequence types of *E. coli*.

The manganese transporter *mntH* [123] has two alleles present in the core genome of ST131 (Fig. 3.15), which raises the question of why there is a fixed second allele and why this has not happened in another ST pangenome. Functional domain analysis of 'mntH_1' and 'mntH_2' returned the same InterPro an-

notation accession number for both alleles. Similarly, *fhuA* and *crt* have two alleles simultaneously present in the core genomes of multiple STs (Fig. 3.15). The *crt* alleles are also involved in a conditional relationship. From the clustermap in Fig. 3.15, we see that the 'crt_1_crt_3_crt_2' cluster is only present in an ST pangenome when 'crt_2_crt_1_crt_3' is also present, with the possible exception of ST144. Additionally, the 'fhuA_2' cluster is only present, excluding phylogroup F STs, when 'fhuA_1' is also present in the ST, with the possible exception of ST38.

Fixation of the propanediol utilisation operon *pduCDEF* has occurred uniquely in ST144 (Fig. 3.15). These genes are reported in other STs (ST372, ST28, ST141, ST648, ST12, ST69, ST14, ST38) at lower average frequencies, showing that *pduCDEF* are accessory genes intermittently present within these STs. The *pdu* operon is involved in anaerobic respiration which is used by enteropathogenic *Enterobacteriaceae* to out-compete existing intestinal microbiota during infection and is frequently reported in *Yersinia enterocolitica* and *Salmonella* Typhimurium [129]. However, this was considered to be a rare phenotype in *E. coli* [162]. Each of the four genes presented here provide additional evidence for sequence level variation in metabolism between *E. coli* STs.

3.4 Discussion

Extensive phenotypic variation and the existence of diverse STs within *E. coli* is well documented [32, 162]. However, little is known about how the genetic repertoire of each ST varies in terms of pangenome structure and content, and consequently which genes are given core status within different ST pangenomes. We build upon previous work analysing a single *E. coli* species pangenome [71, 162] or *E. coli* ST pangenome [47, 130, 131] by performing large scale comparative analysis on 21 ST pangenomes constructed from over 20,000 genomes. We introduce the concept of comparative pangenomics with a method that interro-

gates ST pangenome content and structure variation across the species. We also classified the ST-specific core genome of each ST pangenome into COG functional groups. Our study revealed variation in pangenome structure and core genome functionality both across and within *E. coli* phylogroups.

Pangenomes can reflect the ecology of an organism [42, 122] so insight may be gained by translating gene presence/absence to, for example, niche occupation. Genes core to an ST, or a group of STs, provide indicators of evolutionary advantages in certain ecological settings and genetic backgrounds [70]. From our dataset, ST10 was the only ST pangenome to have no unique core genes. This weak unique core signature may reflect the heterogeneous nature of ST10 [2]. This aligns with previous work that has identified ST10 as a generalist lineage and a potential genetic reservoir for other *E. coli* lineages [77, 162]. However, it may be possible that other STs in our dataset are less well sampled than ST10 and are therefore less representative of their ecological realities. An underlying caveat of all pangenome analysis is sampling bias. There is almost certainly incomplete coverage of all possible source diversity in our dataset which is impossible to fully capture. This is primarily caused by oversampling of particular niches such as human clinical samples within *E. coli* genomic datasets. Inadequacy in sampling will also affect the classification of a gene as core. For instance, it may be poor sampling that makes a gene appear as core, when denser sampling would have moved it to the accessory category. Ecological under-representation within sampling must always be considered when interpreting any results from pangenomic analysis.

Previous estimates of the size of the *E. coli* core genome fell in the range of 1,000 to 3,000 gene clusters and were extrapolated from small genome collections ranging from 14 to 186 isolates [31, 85, 113]. We build upon this earlier work by estimating a 2,172 *E. coli* core genome from a much larger dataset. The bias within EnteroBase towards human pathogens, due to clinical relevance,

may impact our defined core genome as there could conceivably be a false over-representation of genes relating to, say, virulence and antibiotic resistance as a result of our sampling. Many of the STs sampled in this work are ExPEC lineages, but there are also representative lineages from other pathotypes and commensals, and the lineages represent all possible phylogroups and a range of AMR prevalence, and therefore we believe are a broad representative sampling of *E. coli* from many common genetic backgrounds. Nonetheless, the non-uniformity in structure between *E. coli* ST pangenomes demonstrates the extent of the flexibility within this species and is a valuable lesson gained from comparative pangenomics.

The presence of multiple categories of metabolic genes were seen to correlate with the presence of AMR genes and MDR in the accessory genome. Work by Connor and colleagues [40] has also shown that AMR gene carriage correlates to metabolic gene carriage. How this correlation between metabolism and resistance has occurred we do not know yet, but this observation holds across multiple STs in different phylogroups so may have evolved in parallel in multiple phenotypic settings within the species. The negative correlation between the proportion of the accessory genome of unknown function (COG category S) and carriage of resistance genes does not have an obvious biological implication, but we speculate that this may be influenced by the bias of clinical studies to focus on resistance leading to better characterisation of these isolates and their genomic repertoire.

We also found that the function of genes given core status within an ST pangenome (the ST-specific core genome) varied between STs, with certain ST pangenomes having higher percentages of genes in four functional COG categories: energy production, carbohydrate metabolism, transcription, and cell membrane biogenesis. This enrichment may signpost ST-specific adaptive evolutionary processes as a signature of selection via accumulation of allelic diversity.

Reduced metabolic flexibility has been observed frequently during the process of pathogen niche adaption [130, 136, 193] and this may be what we see occurring in MDR lineages. It is already known that there are large ecological variations in *E. coli*. Isolates have been found as gut commensals in most animals, as well as in environmental samples, and can exist on a spectrum of pathogenicity ranging from complete commensal to strict pathogen.

Our goal was to test whether *E. coli* ST pangenomes are evolving in a uniform way. Our data shows that variation in core functions between ST pangenomes is a clear signal of ST-specificity, and we show that ST pangenomes are distinct in different ways, from structure to alleles of genes varying in core status across ST pangenomes. We have also provided a putative list of core gene clusters from a dataset of over 20,000 *E. coli* genomes. We believe that this comparative pangenomics approach represents a valuable tool in the future analysis of microbial genomics and population genomics.

Chapter Four

MDR *E. coli* do not share a common evolutionary trajectory of gene degradation

4.1 Introduction

The data presented in Chapter 3 show that MDR lineages display increased variation in genes of specific biological processes, including amino acid, inorganic ion, and coenzyme transport and metabolism. We also went on to show that there was allelic metabolic variation, both at the core and accessory gene level, that occurs within and between STs. The prevalent nature of metabolic variation within *E. coli*, and particularly the enrichment of various metabolic genes in MDR lineage core genomes previously presented draws attention to the role that gene level variation, in the form of SNPs and indels, may play in the evolution of multi-drug resistance.

A potentiating mutation is a mutation that permits, or increases the likelihood of, evolving a phenotypic trait or function. These have been described in the context of inter-species cooperation between *E. coli* and *Salmonella en-*

terica [53] and in the evolution of citrate usage under oxic conditions in *E. coli* [16]. Potentiating mutations that permit multi-drug resistance but are not directly linked to AMR genes or antibiotic target sites is a relatively new concept, with increasing evidence for this phenomenon being gathered over the past ten years. Mutations in intergenic regions were identified by McNally and colleagues in 2016 within the MDR ST131-C/H30 that were described as compensatory mutations that offset the fitness cost of resistance plasmid maintenance [130]. A follow up study in 2019 found mutations in core anaerobic metabolism genes and virulence genes relating to iron acquisition, heavy metal resistance, capsule biosynthesis, lipopolysaccharide (LPS) synthesis, and flagella specific to clades B/C of ST131 that resulted in high allelic diversity in these gene categories [131]. This pattern has also been seen in other MDR lineages where mutations in colonisation factors, such as iron acquisition [60], LPS biosynthesis [220], and anaerobic metabolic genes [60, 220], as well as intergenic regions [60, 220] have been shown to correlate with multi-drug resistance. These correlations led Dunn and colleagues to hypothesise that MDR clones of *E. coli* emerge primarily from their evolution towards out-competing other *E.coli* in the intestinal tract via anaerobic metabolic processes, which is subsequently followed by acquisition of plasmids and evolution of AMR during prolonged colonisation [55].

In light of these previous findings, we now take a systematic approach to identify parallelisms between functionally significant non-synonymous changes in protein coding genes, of any function, and multi-drug resistance. To do so we will use the power of machine learning methods. The delta-bitscore method [205] leverages profile hidden Markov models to quantify the potential impact of nonsynonymous mutations on the function of a protein by comparing mutations in a protein to naturally occurring patterns of variation. If a mutation is under-represented in nature, it is likely to be selected against due to a negative impact on protein function. The random forest (RF) algorithm is a 'supervised' machine learning approach that can be used to make predictions about data by

learning from examples. It is well suited to delta-bitscore input data because alternative supervised learning approaches, such as support vector machines and elastic nets do not allow interactions between genes as 'and or' relationships to be modelled. These are relationships where two genes perform the same function and either or both can be present for implementation of the function. These types of genetic interactions may be important in cases where inactivation of genes within a pathway or biological process need to be considered collectively, making alternative methods poorer choices for use on this data. The inclusion of bootstrapping in the random forest algorithm also makes it more robust.

The combination of delta-bitscore and machine learning techniques has already successfully identified signatures of mutational burden that indicate adaption to an invasive lifestyle in *Salmonella enterica* [204] and *Campylobacter jejuni* [206], We aim to apply the same methodology to identify signatures of mutational burden in genes that may be a precursor to the evolution of multi-drug resistance.

In this chapter, we will again interrogate our curated dataset of over 20,000 genomes representing the full phylogenetic diversity of *E. coli* described in §2.2.1, this time with a focus on functionally significant non-synonymous changes in protein coding genes. We quantify the burden of atypical mutations in protein coding genes using the delta-bitscore metric to then apply a machine learning approach to probe this unique dataset. We assess the ability of a RF and an XGBoost classifier to identify atypical mutations in protein coding genes that are signatures of multi-drug resistance in *E. coli* and compare the most important predictive features across both methods. A RF classifier is based on bagging where multiple decision trees votes are combined to make a final prediction whilst XGBoost is a boosting algorithm where decision trees are built sequentially to correct the errors of the previous tree. The latter is more prone to over-fitting. Finally, we will evaluate how well signatures identified by the RF gen-

eralise across distinct *E. coli* lineages.

4.2 Methods

4.2.1 Merging pangenomes

To build a machine learning model, orthology relationships between all proteins from 20,577 isolates are required. To obtain this information, the 21 ST pangenome graphs created in §3.2.3 were merged using the `panaroo-merge` function within Panaroo. This approach is recommended by the creators of Panaroo for working with large or diverse datasets as the algorithm assumes that a dataset is not overly diverse and therefore results will be improved by working on separate clusters of genomes individually before merging.

4.2.2 Bitscore analysis

To assess the functional divergence of proteins within the dataset, bitscores were assigned to each individual orthologous group. Firstly, sequences with internal stop codons were trimmed using a custom Perl script [github.com/Gardner-BinfLab/invasive_salmonella/blob/master/build_random_forest/trimtruncations.pl]. This is done to prevent hmmsearch scoring residues past premature stop codons as if they were translated. The Gammaproteobacteria hidden markov models (HMMs) database was downloaded from eggNOG (v5.0) [78]. All trimmed protein sequences were compared to the database with hmmsearch [hmmer.org] to assign bitscores [205] to 36,766 orthologous groups. These results were then parsed with a custom perl script [github.com/Gardner-BinfLab/deltaBS/blob/master/src/parse_bitscores.pl] to create a tab-delimited file of bitscores that

were a suitable format for analysis. Instances where no bitscore was assigned to a gene cluster were removed from the dataset to exclude samples that did not have an appropriate model match.

4.2.3 Bitscore quality control

After bitscore assignment, quality control was carried out to ensure that all gene clusters included in our analysis had sufficient numbers of bitscores assigned and that appropriate models were chosen to assign the bitscores. The quality control was done in three stages. Firstly, for each orthologous group, samples with a bitscore two standard deviations above the median group bitscore value were removed. This was done to avoid instances where a missing stop codon causes a longer sequence segment than intended to be scored, resulting in an incorrectly inflated bitscore. This artificial score inflation can occur when the extended sequence segment is a high match to related genes in other species where there is no stop codon. Secondly, groups with less than 100 bitscores assigned overall were removed. This was done to reduce potential noise in the dataset as these groups are either genes at extremely low abundance in the dataset, or likely to be incorrect matches between models and protein sequences. Finally, groups with multimodal bitscore distributions that had stationary points > 50 bitscores apart were removed. This was done to exclude gene clusters with more than one top model hit as distinct peaks in a bitscore distribution correlate to bitscores produced from top hits to different profile HMMs.

4.2.4 Phylogenetically-informed subsampling

For the purposes of the machine learning analysis, a smaller set of samples was required to ensure feasible computation times. The initial dataset of 20,577

genomes (Table 3.1) was trimmed using Treemmer [132] with the default parameter of pruning one leaf per iteration. This produced a dataset of 5,000 genomes still spanning the 21 STs that represent the phylogenetic diversity initially captured in the larger dataset.

4.2.5 Delta-bitscore analysis

First, we examined the delta-bitscore distribution of core genes for MDR and non-MDR strains. We defined our delta-bitscore metric with the following equation:

$$DBS_i = \tilde{BS} - BS_i,$$

where DBS_i is the delta-bitscore for the i^{th} sample, $\tilde{BS}$ is the median bitscore value and BS_i is the bitscore of the i^{th} sample for a given orthologous group. In cases where there was no difference between the median and sample bitscore, these zero delta-bitscore values were removed as we were only interested in deviation from the consensus.

To assess the delta-bitscore distributions of the core genome, the 5,000 subsampled dataset was first separated into MDR and non-MDR groups. Then a list of core gene clusters was taken from the pangenome reference output file of Panaroo using a custom python script [github.com/lillycummins/InterPangenome/blob/main/pangenome_splitter.py] with a 0.99 core threshold. Delta-bitscores for all gene clusters in each phenotypic group were compiled. A two-sample Kolmogorov-Smirnov (KS) test was implemented for the MDR and non-MDR core genome comparison using the `ks_2samp` package imported from `scipy.stats`. The same method was used to compute the KS statistic and p values for individual core genes. The core genes were then assigned COG functional categories as previously described in §3.2.4. Fishers exact test was used to identify COG categories significantly enriched with genes with significantly different

delta bitscores in MDR and non-MDR populations. This test was implemented using the `fisher_exact` package imported from `scipy.stats`. To assess similarity of the delta-bitscore distributions of the MDR and non-MDR populations a quantile-quantile plot was created using `pylab`. The quantiles for the larger non-MDR dataset were found by linear interpolation using `interp` from `numpy`. Sequence alignments for genes with significant differences in delta-bitscore distributions were performed using clustalo (v1.2.4) with the default settings.

4.2.6 Machine learning

The python package `sklearn` [150] was used to build machine learning classifiers and balanced accuracy was used to measure the performance of models. A random forest classifier was built with specified hyperparameter settings. The maximum number of features to consider for each split in the decision trees within the random forest could be 10, 50 or 80% of the total features. The number of decision trees to include in the random forest ensemble could be 10, 100, 500, 1,000, or 10,000. The input data was divided into testing and training data using `test_train_split` from `sklearn.model_selection` with a 20% test set. The models were evaluated using grid search for hyperparameter optimisation. Grid search cross-validation was performing using `GridSearchCV` 3-fold cross-validation from `sklearn.model_selection` to identify the best hyperparameters. Model accuracy was plotted across the different hyperparameter values to evaluate the suitability of the hyperparameter space. The features from the initial model training were trimmed by removing any zero feature importance values. After this initial cycle, model retraining with feature selection was implemented five times. During the five further training cycles, iterative feature trimming removed features below the median feature importance value.

An XGBoost classifier for binary classification with hinge loss, a 0.6 column

subsampling ratio and a maximum depth of two for each tree was built with specified hyperparameter settings. The boosting process could have a rate of 0.01 or 0.1, the regularisation parameter (alpha) could be 0.00001 or 0.0001, the minimum loss reduction required to make a further partition on a leaf node (gamma) could be 0.05 or 0.1, and the subsampling ratios of the training instances could be 0.2, 0.4, or 0.6. The same pattern of feature trimming and retraining cycles was implemented as previously described for the random forest classifier.

4.3 Results

4.3.1 A 54,597 gene cluster *E. coli* pangenome

E. coli pangenomes have been estimated and analysed in numerous contexts [85, 113, 137, 162] but not from a dataset as vast as the collection of genomes considered in this work. We constructed an *E. coli* pangenome from 20,577 genomes by merging the 21 individual ST pangenomes analysed in the previous chapter. This resulted in a total of 54,597 orthologous groups being identified (Figure 4.1), which were classified by Panaroo as 48,227 cloud genes (0% $\leq$ strains < 15%), 3,055 shell genes (15% $\leq$ strains < 95%), 123 soft core genes (95% $\leq$ strains < 99%), and 3,192 core genes (99% $\leq$ strains $\leq$ 100%).

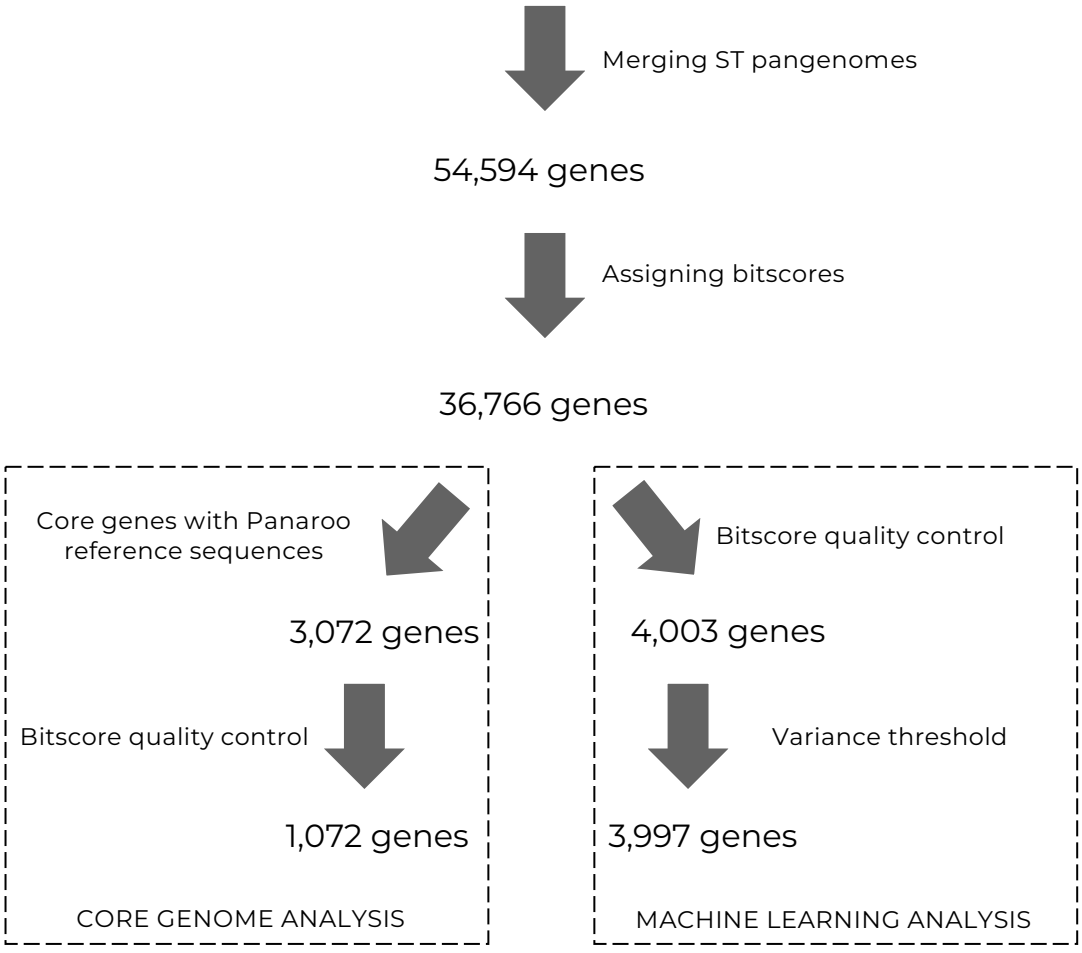


Figure 4.1: Workflow of the analysis steps discussed in this chapter. Number of genes kept at the end of every step is shown below each arrow. Steps pertaining to §4.3.2 are shown in the left dashed box and steps pertaining to §4.3.3 are shown in the right dashed box.

4.3.2 Core genome bitscore analysis

Significant difference in core genome functional sequence variation between MDR and non-MDR phenotype

The delta-bitscore metric was used to quantify the functional significance of gene sequence variation. Of the 3,162 core genes taken from the pangenome ref-

erence Panaroo output file, 1,072 could be assigned appropriate bitscores that passed quality control after comparing them to the eggNOG Gammaproteobacteria profile HMMs. The dataset was then reduced to 5,000 genomes for computational efficiency and split into MDR (n = 1,738) and non-MDR (n = 3,262) populations. We found a significant difference in delta-bitscore distributions between these two groups (KS two sample test, $p = 3.581 \times 10^{-98}$). A quantile-quantile (Q-Q) plot was used to assess differences in the delta-bitscore distributions for the two populations. The blue line in Figure 4.3 represents the quantile levels of the two datasets plotted alongside a black reference line representing an equal distribution. Figure 4.3 shows that the two populations do not share a common distribution of delta-bitscores, with the non-MDR population displaying significantly higher positive values. The non-overlapping mean 95% confidence intervals (MDR: 0.578-0.598, non-MDR: 0.688-0.706) also infer a positive shift in delta-bitscore distribution for the non-MDR population. Together, these statistical results show increased positive delta-bitscore values for the non-MDR group indicating greater divergence from the profile models for protein coding genes from the non-MDR population when compared to the MDR population. This suggests that there are more deleterious mutations occurring within the non-MDR population, which may be resulting from genetic drift or niche specialisation.

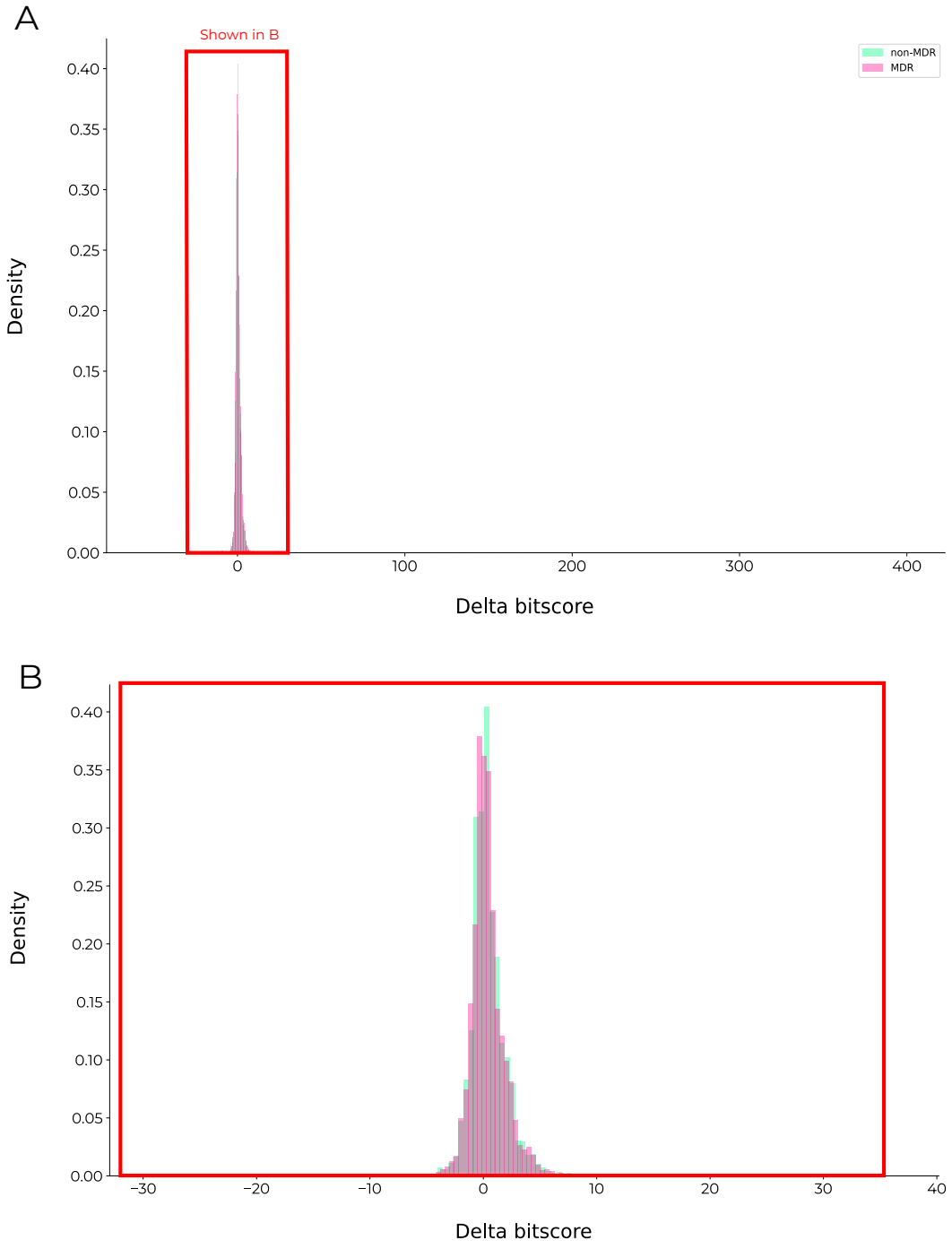


Figure 4.2: (A) Density distribution plot of delta-bitscores of core genes. (B) Zoomed segment of (A). Only non-zero values are plotted to focus on deviation from consensus sequences. MDR density range: [-42.9, 357.4], non-MDR density range: [-55.5, 400.0]. KS test statistic: 0.016, p-value: 3.581×10^{-98}.

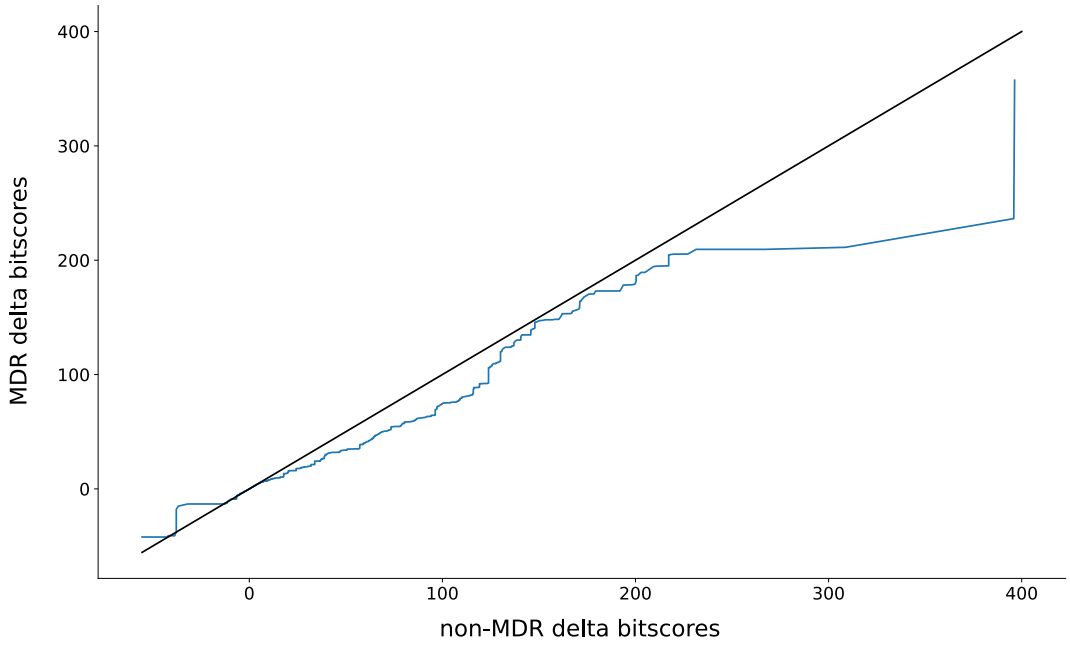


Figure 4.3: Quantile-quantile plot of the delta-bitscore distributions of the non-MDR and MDR populations, shown in blue. Black reference line indicates an equal distribution.

Significant sequence variation differs by gene function

The functional significance of gene sequence variation was then assessed on a gene by gene basis for the core genome of the 5,000 *E. coli* isolates. Of the 1,072 core genes assigned appropriate bitscores, 72% (n = 775) of genes displayed a significant difference in delta-bitscore distribution between the MDR and non-MDR isolates (KS two sample test, $p < 0.05$). Classification into functional COG categories revealed that genes with significant differences in delta-bitscores were not evenly distributed across functional categories (Figure 4.4). Fishers exact test was used to identify COG categories significantly enriched in genes with or without significant differences in delta-bitscores between the MDR and non-MDR populations. The p values are shown in Table 4.1. COG cate-

gories that were significantly enriched with core genes that had significant differences in their delta-bitscores between the MDR and non-MDR populations were cell motility (N), three metabolic groups pertaining to energy conversion, amino acids, and coenzymes (C, E, H), and genes of unknown function (S). Two COG categories had significantly higher levels of genes that did not have significant differences in their delta-bitscores between the MDR and non-MDR populations: translation, ribosomal structure and biogenesis (J) and cell cycle control, cell division, and chromosome partitioning (D). Translation and cell division are both essential biological processes which makes it unlikely that significant genetic variation in genes of related functions would be tolerated.

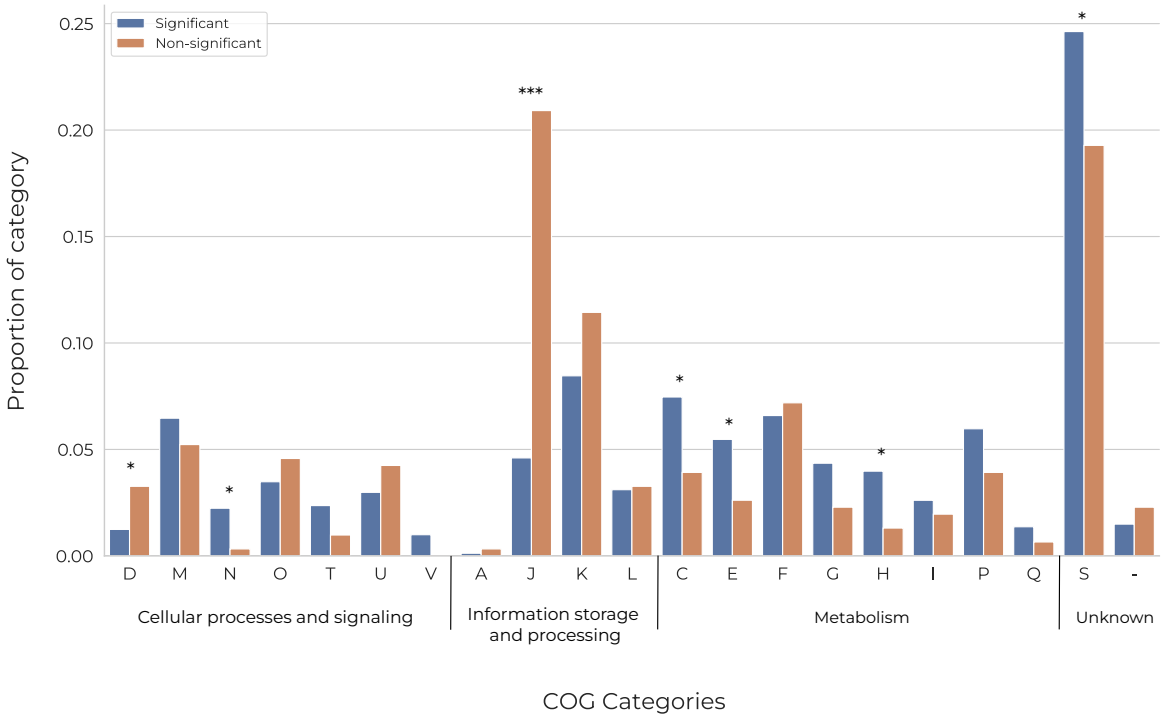


Figure 4.4: Proportion of genes per COG category with significant and non-significant delta-bitscore differences between MDR and non-MDR populations. * : $p < 0.05$, *** : $p < 0.001$.

Table 4.1: COG category correlation with number of genes with a significant difference in delta-bitscore distributions between MDR and non-MDR populations. Categories where genes with significant sequence variation were the majority are indicated in the final column. Rows ordered by ascending p value. Rows highlighted in bold are significant ($p < 0.05$). Correlations were assessed using Fishers exact test.

COG Category		p value	Sig. seq. variation
J	**Translation, ribosomal structure and biogenesis**	**$5.88034501 \times 10^{-15}$**	No
H	**Coenzyme transport and metabolism**	**0.02233998**	Yes
C	**Energy production and conversion**	**0.02910626**	Yes
N	**Cell motility**	**0.03491367**	Yes
E	**Amino acid transport and metabolism**	**0.03947017**	Yes
D	**Cell cycle control, cell division, chromosome partitioning**	**0.04083329**	No
S	**Function unknown**	**0.04627162**	Yes
G	Carbohydrate metabolism	0.11450751	Yes
V	Defence mechanisms	0.11470415	Yes
T	Signal transduction mechanisms	0.15475307	Yes
K	Transcription	0.1661356	No
P	Inorganic ion transport and metabolism	0.18352007	Yes
U	Intracellular trafficking, secretion, and vesicular transport	0.3518905	No
-	No functional annotation available	0.43929156	Yes
A	RNA processing and modification	0.48141264	No
O	Post-translational modification, protein turnover, and chaperones	0.48290359	No
M	Cell wall/membrane/envelope biogenesis	0.48544924	Yes
Q	Secondary metabolites biosynthesis, transport, and catabolism	0.53387507	Yes
I	Lipid transport and metabolism	0.66472361	Yes
F	Nucleotide transport and metabolism	0.79034686	No
L	Replication, recombination and repair	1	No

Sequence variation is not indicative of the MDR genotype

To inform our selection of genes within COG categories that have a significant difference in sequence variation between the MDR and non-MDR populations,

genes were selected from the literature that have been shown to be upregulated in response to multi-drug resistance plasmid acquisition that were present in our dataset. The *glp*, *hyb*, *hyc*, and *nar* operons have been identified previously by Dunn, Carrilero, and colleagues as genomic regions that are altered in response to multi-drug resistance plasmid acquisition [30, 54]. The components of these operons; *glpQ*, *hybE*, *hycI*, and *narJ*; were present within COG category C in our dataset. The bitscore distributions of these four genes all show a common signal that could indicate increased functional divergence in the non-MDR population, demonstrated by the pronounced trail of lower bitscores in each plot (Figure 4.5). Manual inspection of protein alignments for each gene revealed that the lower scores present in the non-MDR group were attributed to single amino acid changes at conserved sites. The substitutions occurred at multiple different sites within each gene. These findings imply that there is no common mutation across lineages that can be associated with an MDR genotype.

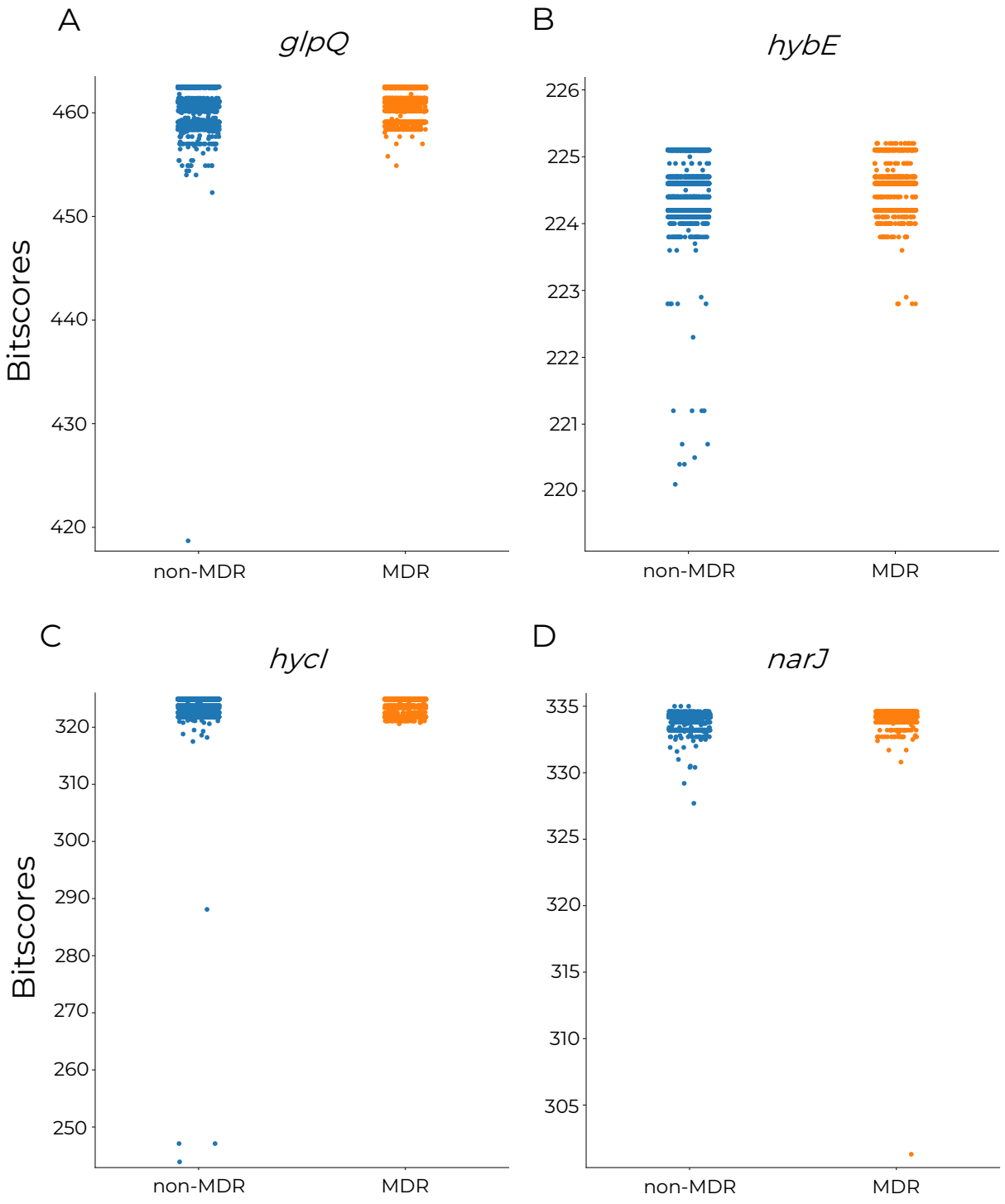


Figure 4.5: Bitscore plots for A) *glpQ* (n = 4968), B) *hybE* (n = 4986), C) *hycI* (n=4997), and D) *narJ* (n = 4994).

4.3.3 Predicting multi-drug resistance from functionally significant mutations

Moving beyond the core genome, we investigated whether functionally significant mutations occurring in genes anywhere in the pangenome could be used as predictors for the MDR phenotype. To do so, machine learning algorithms were trained on the 5,000 genome dataset. The distribution of bitscores from §4.3.2 was used by the machine learning models to identify genes that showed a difference in adherence to sequence constraints between the MDR and non-MDR populations. The input data consisted of 3,997 gene bitscores, following quality control measures described in §4.2.3 and the removal of genes with variance between 0 and 0.01.

A random forest model predicts MDR phenotype with 75.3% accuracy

A RF model was initially selected for training due to its suitability for biological data. Iterative feature selection was performed to improve the accuracy of the model. Initially, genes with zero importance were removed (n = 572) and the model was retrained on the smaller gene set. Following this, four further rounds of feature selection were performed, where genes below the median feature importance value were removed each time. Balanced accuracy was used to assess performance as we were using an imbalanced dataset. The balanced accuracy of each feature selection stage is shown in Figure 4.6. The first iteration of feature selection provided the most accurate classifier with a balanced accuracy of 75.3% (Figure 4.6).

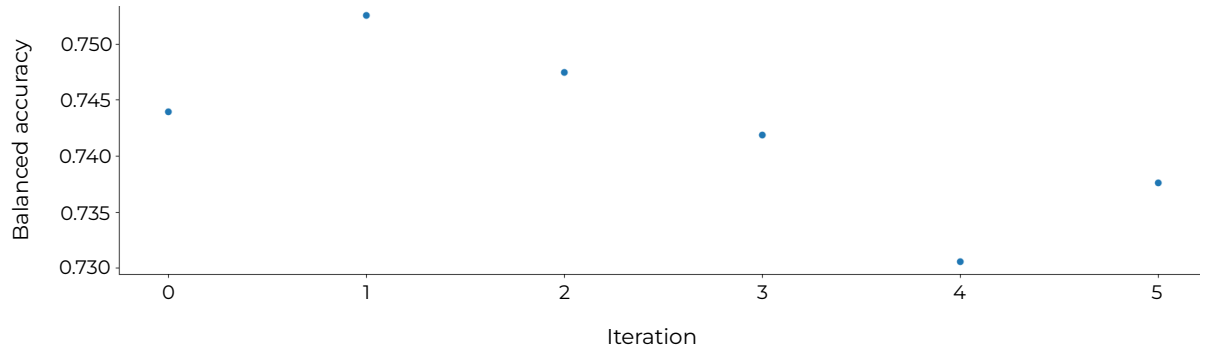


Figure 4.6: Balanced accuracy score after each iteration of feature reduction for the random forest model training.

Metabolic genes account for half of top ten predictive genes in the RF model

The top ten predictive genes, alongside their assigned importance value and COG category are shown in Table 4.2. The gene *ybaA* was deemed the most informative feature for the RF classifier, with a feature importance value of 0.102. The function of this gene is unknown, as it encodes a protein of unknown function (DUF1428). All other predictors had relatively lower feature importance values below 0.03 (Figure 4.7). Half of the ten best predictors had metabolic functions (three energy production and conservation genes, *mioC*, *ynfF*, *hmp*; one inorganic ion transport and metabolism gene, *nrfD*, and the regulator of the operon responsible for 2-keto-3-deoxy gluconate metabolism and transport *kdgR*).

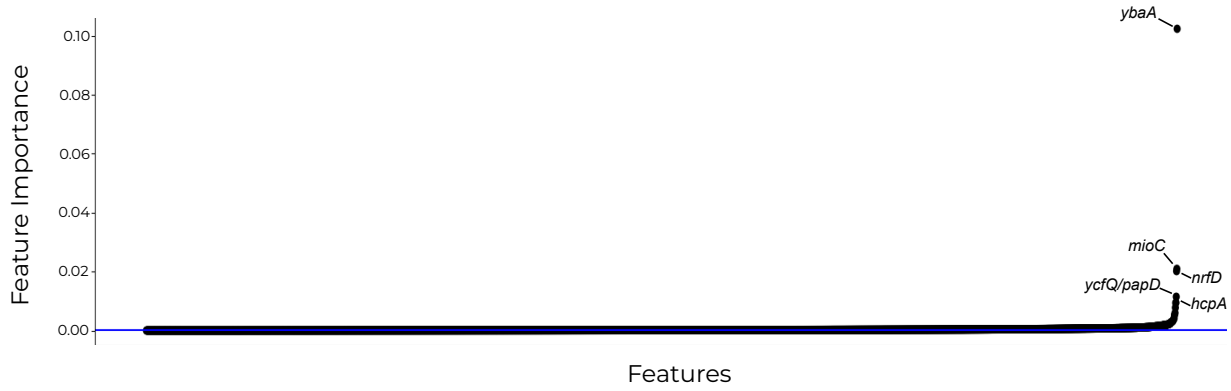


Figure 4.7: Feature importance for each gene in ascending order from the random forest model with the highest balanced accuracy. Blue line indicates the median feature importance value used for iterative feature selection. The top five most important genes are labelled.

Table 4.2: The ten most importance features of the random forest model for predicting multi-drug resistance.

Gene	**Description**	**Feature importance**
ybaA	DUF1428	0.10238156
mioC	electron transporter required for biotin synthase activity	0.0209262
nrfD	Polysulphide reductase	0.02015565
yfcQ	Pili assembly chaperone PapD, C-terminal domain	0.01145689
hcpA	PAAR repeat-containing protein	0.0094587
kdgR	Transcriptional regulator	0.00937971
ynfF	Molybdopterin oxidoreductase	0.0077315
symE	Endoribonuclease activity	0.00598304
hmp	Is involved in NO detoxification in an aerobic process, termed nitric oxide dioxygenase (NOD) reaction that utilizes O(2) and NAD(P)H to convert NO to nitrate, which protects the bacterium from various noxious nitrogen compounds. Therefore, plays a central role in the inducible response to nitrosative stress	0.0055481
lysR	Transcriptional activator protein LysR	0.00514319

An XGBoost model predicts MDR phenotype with 73.4% accuracy

An XGBoost model was then selected for training due to the self-correcting aspect of the algorithm's adaptive learning, to try and improve upon the RF classifier predictions. Iterative feature selection was performed as previously described. The first iteration of feature selection again provided the most accurate classifier with a balanced accuracy of 73.4% (Figure 4.8). This is lower than the RF classifier and is therefore not a better predictor of multi-drug resistance.

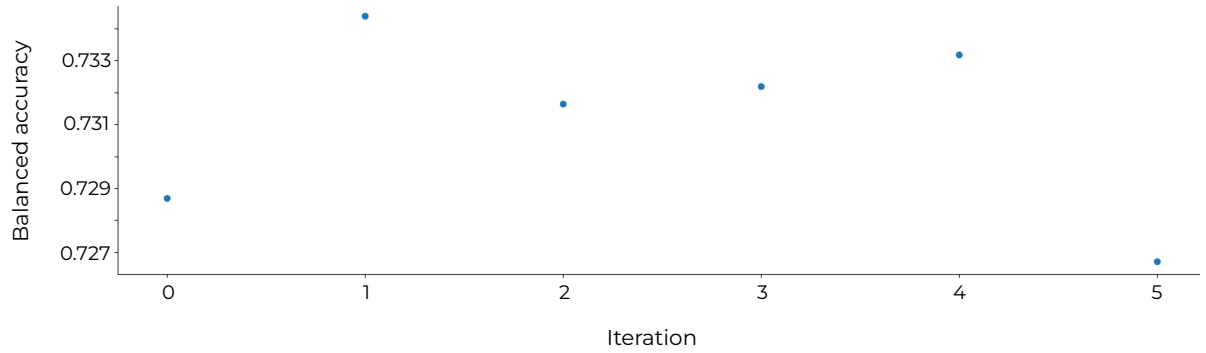


Figure 4.8: Balanced accuracy score after each iteration of feature reduction for the XGBoost model training.

ybaA is the most informative gene for both models

The XGBoost model also selected *ybaA* as the most importance feature, though with a lower feature importance of 0.041 (Figure 4.9) compared to the RF feature importance value of 0.102. Four out of the ten most important features determined by the XGBoost model overlapped with the top ten predictors of the RF model (Table 4.3). These were *ybaA*, a cluster named *hcpA* but annotated as encoding a PAAR repeat-containing protein, *nrfD* encoding a polysulphide reductase, and the transcriptional regulator of transport and catabolism of 2-keto-3-deoxy gluconate *kdgR* [157].

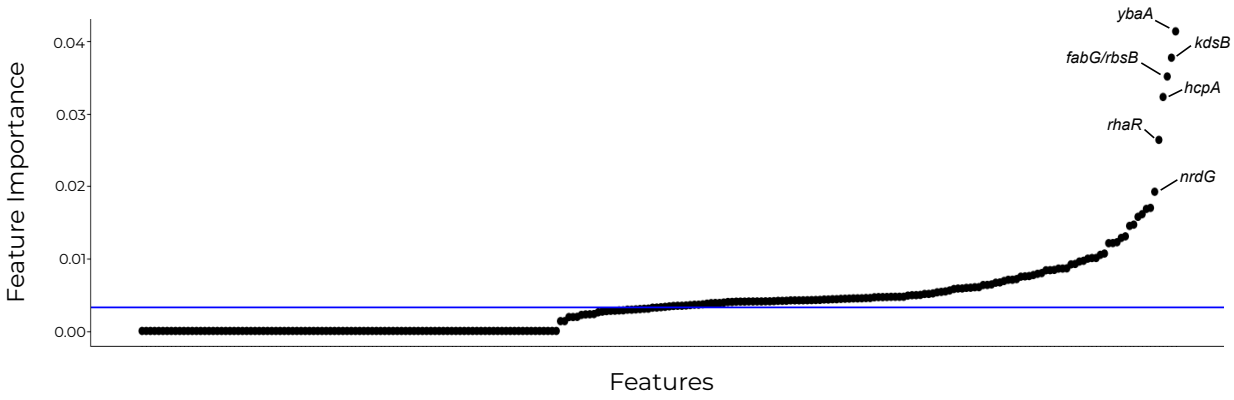


Figure 4.9: Feature importance for each gene in ascending order from the XGBoost model with the highest balanced accuracy. Blue line indicates the median feature importance value used for iterative feature selection. The top six most important genes are labelled.

Table 4.3: The ten most importance features of the XGBoost model for predicting multi-drug resistance.

Gene	**Description**	**Feature importance**
ybaA	DUF1428	0.04138161
kdsB	Activates KDO (a required 8-carbon sugar) for incorporation into bacterial lipopolysaccharide in Gram-negative bacteria	0.03773974
fabG/rbsB	KR domain	0.03514442
hcpA	PAAR repeat-containing protein	0.032314878
rhaR	Activates expression of the rhaSR operon in response to L-rhamnose	0.026388483
nrdG	Activation of anaerobic ribonucleoside-triphosphate reductase under anaerobic conditions by generation of an organic free radical, using S-adenosylmethionine and reduced flavodoxin as cosubstrates to produce 5'-deoxy-adenosine	0.019224498
nrfD	Polysulphide reductase	0.016993467
kdgR	Transcriptional regulator	0.01685254
ydiV	-	0.016128054
group_6486	DUF1158	0.01575974

Predictions are influenced by phylogenetic structure

Both model predictions vary in accuracy between STs (Figure 4.10). No ST410 genome is ever predicted to be MDR, despite the high levels of multi-drug resistance in this lineage (see bottom pale green segment of Figure 4.10). Similarly for ST648, another lineage with high resistance levels, the MDR phenotype is never correctly predicted and one susceptible genome is incorrectly predicted to be MDR (see top orange segment of Figure 4.10). Conversely, some STs, such as ST12, ST17, and ST127, are largely predicted to be MDR when they are not. The percentage of correctly predicted phenotypes per ST in the test dataset is shown in Table 4.4 for both models. The gene *ybaA* was deemed the most important feature by both models. The bitscores for this gene are shown alongside the phylogeny in Figure 4.10. Overall, the XGBoost model provided similar predictions to the RF model, but predicts MDR less frequently (Figure 4.10). An MDR genotype was never predicted within phylogroup A STs, (ST10, ST167, and ST410 - the bottom three lineages shown in Figure 4.10). More generally, XGB predictions cluster together in phylogenetic groups compared to the RF. This suggests that phylogenetic structure may have a greater influence on the results of this model than the RF.

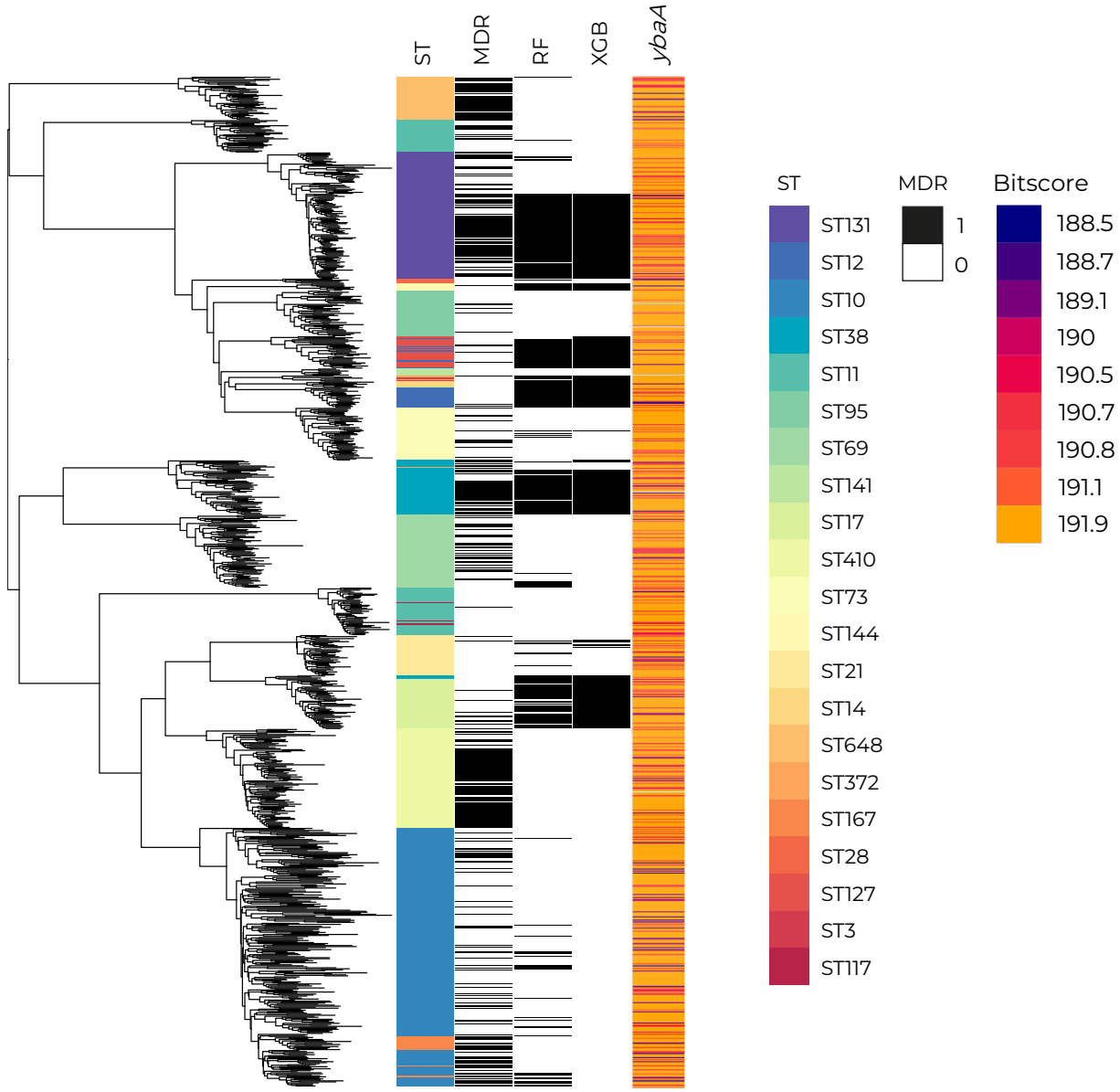


Figure 4.10: Phylogeny of the 1,000 *E. coli* test set. Shading to the right indicates sequence type, actual MDR phenotype (MDR), MDR phenotype predictions of random forest (RF) and XGBoost (XGB) models, and the bitscore values of the top ranked informative gene *ybaA*. Bitscores below 180 not shown for ease of visualisation.

Table 4.4: Distribution of STs in the 1000 genome test dataset alongside the actual MDR prevalence (MDR) and prediction accuracy per ST for each model. RF: random forest, XGB: XGBoost.

ST	Genomes	MDR (%)	RF (%)	XGB (%)
ST10	241	24.90	75.10	75.10
ST11	76	14.47	84.21	85.53
ST117	3	0.00	100.00	100.00
ST12	28	7.14	10.71	7.14
ST127	24	12.50	20.83	12.50
ST131	125	60.00	68.00	68.80
ST14	5	0.00	0.00	0.00
ST141	8	0.00	87.5	87.5
ST144	7	14.29	14.29	14.29
ST167	15	73.33	26.67	26.67
ST17	49	14.29	18.37	14.29
ST21	40	5.00	90.00	85.00
ST28	5	0.00	80.00	100.00
ST3	3	0.00	0.00	0.00
ST372	4	25.00	75.00	50.00
ST38	57	66.67	63.18	64.91
ST410	98	80.61	19.39	19.39
ST648	43	86.05	11.63	13.95
ST69	72	38.89	51.39	61.11
ST73	52	19.23	76.92	82.69
ST95	45	11.11	88.89	88.89

Both models predict ST131-C/H30Rx as being MDR

The predictions of both models for ST131 appear to be heavily influenced by phylogenetic structure. Figure 4.11 depicts the three-clade structure of the ST131 test set against both model predictions. We observe that the RF and XGB predictions almost exclusively predict the MDR clone ST131-C/H30Rx (denoted by a * in Figure 4.11) as being MDR within ST131. Clade B isolates are predicted to be susceptible by both models and a handful of clade A isolates are predicted to be MDR by the RF model only. This is an interesting result, as although the accuracy is not high for these predictions when comparing to the MDR phenotype,

the models may be predicting the 'potential to become MDR' as they accurately detect the MDR clone within ST131.

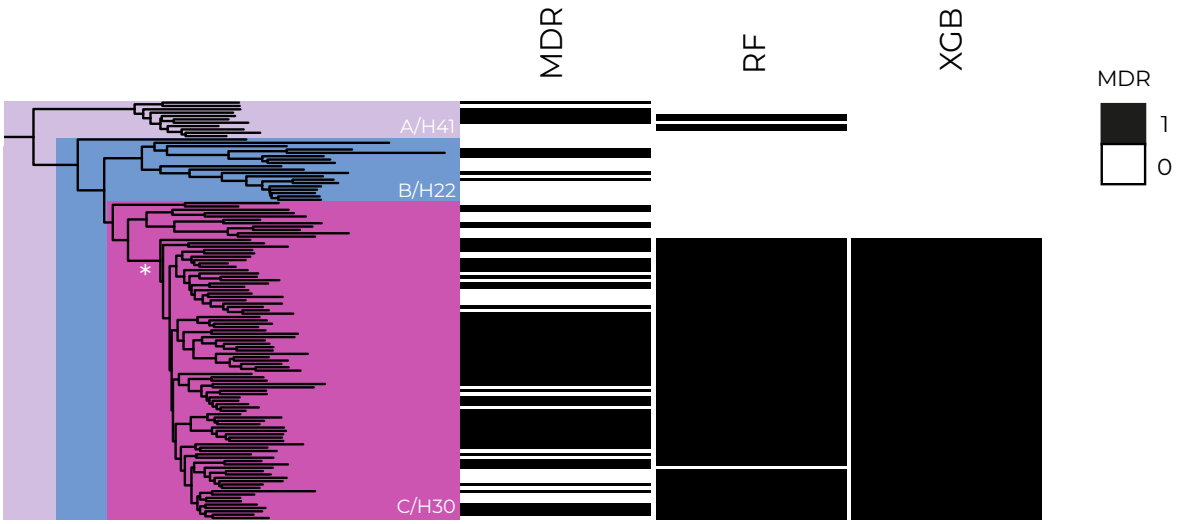


Figure 4.11: Phylogeny of ST131 with clades demonstrated by tonal shading. The MDR clone ST131-C/H30Rx is denoted by *. Shading to the right of the phylogeny indicates MDR phenotype (MDR) and MDR predictions of random forest (RF) and XGBoost (XGB) models. Zoomed in segment from Figure 4.10.

4.4 Discussion

Potentiating mutations have been hypothesised as crucial steps in the evolutionary trajectory of MDR clones. Previous studies have identified these mutations in various genetic contexts; within intergenic regions, anaerobic metabolism loci, and colonisation factors. Here, we took a systematic approach to compare functionally significant genetic variation between MDR and non-MDR populations of *E. coli* at a core gene level. We then extended this to test whether machine learning algorithms could correctly predict whether an isolate was MDR or not based on the functionally significant genetic variation present within its genome.

The *E. coli* pangenome constructed in this chapter consisted of over 50,000

distinct gene clusters, and a core genome size of 3,192 which is in the range of previous estimates [31, 85, 113]. Roughly one third of the *E. coli* core genome was successfully assigned bitscores. The poor coverage of the eggNOG Gammaproteobacterial database for genes that are commonly present in all *E. coli*, of which there are hundreds of thousands of genomes sequenced now, is a surprising limitation of this repository of orthologous groups. Further work to design a custom database of HMMs for *E. coli* core genes would greatly enhance the analysis described in this chapter, but was out-of-scope for the purpose of this thesis. To cover all genes found in > 15% of the sampled *E. coli* population in this work would require 6,370 models to be built. Another caveat of the delta-bitscore method is that epistatic interactions within a protein are not taken into account. Sites are considered as independent units during scoring, therefore any conditional relationships between amino acid residues, such as salt bridges, are overlooked when considering substitutions at a specific site. Potentiating mutations for multi-drug resistance have previously been found in intergenic loci [60, 130, 220], but our chosen method misses the opportunity to identify intergenic signals of multi-drug resistance as scoring can only be applied to protein-coding regions. The changes in gene expression noted in previous work provide evidence this may be a promising future avenue of investigation.

The significant difference found in delta-bitscore distributions between the MDR and non-MDR populations could be attributed to multiple factors. The difference could be resulting from degradation of specific genes, or an overall difference in mutation rate between MDR and non-MDR *E. coli*. A change in mutation rate has been shown to happen during niche adaptation in combination with bottlenecking [133]. The increased prevalence of SNPs in the non-MDR population could also be a reflection on lifestyle. Different lifestyles provide distinct selection pressures that require specific genetic adaptions. All MDR genomes are from ExPEC, whereas the non-MDR population covers a greater variety of niches as it includes pathogens (both InPEC and ExPEC), commensals,

and environmental isolates. The greater variety of lifestyles sampled in the non-MDR group in comparison to the MDR group would therefore be expected to result in greater genetic variation within the population. Overall, the positive shift in core genome delta-bitscore distribution of the non-MDR population imply that the MDR niche isn't resulting in genome reduction or functional specialisation.

We then went on to use bitscores to inform classification models to assess whether we can accurately predict multi-drug resistance from background genetic variation. The balanced accuracy of both models tested here is, in theory, a positive result given the difficulty of the phenotype being predicted. Sequencing genomes is a 'snapshot' of their evolutionary trajectories. Genomes could be about to become MDR, through a SNP or acquisition of a resistance gene *via* HGT, or about to lose multi-drug resistance by loss of a multi-drug resistance plasmid. The rapid nature of bacterial evolution can make it difficult to speculate long term trajectories. Owing to this, it is optimistic to expect a model to accurately predict multi-drug resistance by the nature of this phenotype from the input data available. Further phylogeny-based assessment of the predictions demonstrated that both models were in fact poorly capable of predicting multi-drug resistance. A downfall of many machine learning studies is that accuracy is given as a final result, when in fact digging deeper into the predictions would provide additional and perhaps contrasting findings [203]. When interpreting the results of this study, it is worth bearing in mind that our aim was to predict the 'potential to become MDR'. Therefore, predictions should not perfectly align with the MDR phenotype as a perfect match would mean that no isolates with potential to become MDR, and only isolates that are already MDR are detected.

Upon visualisation, the XGBoost model produced phylogenetically clustered predictions that were spread less widely throughout the phylogeny than the RF predictions, which in turn were still spread less through the phylogeny than the MDR phenotype itself (Figure 4.10). This suggests that both models

were incorrectly using some phylogenetic signals to inform their predictions. However, in the instance of ST131, the phylogenetically-clustered prediction of clades B and C being MDR may actually be very informative of the 'potential to become MDR' within this lineage, given the existing knowledge about the evolution of ST131 [130, 131, 151, 185]. Other lineages where multi-drug resistance is predicted for isolates that are not MDR include ST73, ST95, and ST141, which are all found within phylogroup B2. This phylogroup is known for containing pathogens, and brings into question whether the models are actually making predictions based on the hallmarks of host-adaption or pathogenicity, instead of multi-drug resistance.

Dissection of the models to investigate the most important features highlighted multiple genes and functions of interest. The RF model selected multiple metabolic genes as the most important features, adding evidence to the hypothesis that metabolism plays a key role in niche domination and migration for MDR clones [43, 55]. The colonisation gene *papD*, annotated as *yfcQ* in our dataset, ranked as the fifth most important feature in the RF model, adds to the notion that proficient colonisation may precede the MDR phenotype. The feature *hcpA*/PAAR was common to both models and is linked to the secreted proteins of type VI secretion systems. This is an interesting finding given the previous result of Chapter 2 where structural T6SS components were shown to dissociate with multi-drug resistance in ST131 and ST410. Now we may speculate that non-structural T6SS components may also correlate with multi-drug resistance.

The delta-bitscore method has previously been used to identify sequence differences associated with phenotype. Generally, the phenotype of focus has been pathogenicity or host-adaption [204, 206] so this chapter implemented a new phenotypic application of the method by examining multi-drug resistance. Delta-bitscoring identifies gene degradation which was the evolutionary event associated with the host-adapted phenotype for *Salmonella enterica* [204] and

Campylobacter jejuni [206]. The successes of these previous studies and our inability to accurately predict multi-drug resistance suggests that gene degradation is unlikely to be strongly associated with this phenotype. It is therefore likely that the gene degradation signal we looked for with delta-bitscoring is underrepresented in our dataset, explaining why the model predictions were not sufficiently accurate. An alternative approach could be to consider the genetic background on a pathway level, rather than considering the degradation of individual genes. There are multiple routes for an isolate to become MDR, and we may not be able to capture them all in a single predictive model.

Chapter Five

Conclusions and Future Work

The rise in MDR pathogens is a threat to global health, adding increasing pressure to our health care systems as effective treatment options for infections diminish. Multi-drug resistance is a trait that is spread throughout the *E. coli* phylogeny, as shown by work presented in this thesis, and is not strictly related to antibiotic exposure. Accumulation of resistance determinants is concentrated in specific lineages and some lineages remain largely susceptible to antibiotics despite their prevalence in clinical settings. Documentation of MDR *E. coli* being able to colonise and spread among healthy individuals, in the absence of antibiotic usage, demonstrates that these pathogens are capable colonisers. The combination of high-level resistance and efficient colonisation capabilities makes MDR *E. coli* a double threat in both clinical and community settings. Efforts to better our understanding of how MDR *E. coli* spread and evolve are vital to inform development of new treatment strategies. The overarching aim of this thesis was to improve our understanding of how MDR lineages can be differentiated from the rest of the *E. coli* population, to uncover factors attributed to their accumulation of resistance determinants and rapid colonisation of healthy individuals.

Genomic analysis of globally disseminated lineages that contain MDR clones, namely ST131 and ST410, revealed a pattern where type VI secretion

system genes were present in ancestral clades, but absent from the clades that include MDR clones. We demonstrated that T6SS determinants had been lost in both cases, but through different genomic evolutionary events. Further experimental work could explore the functional repercussions of these loss events. An assessment of firstly, whether the naïve T6SS is functional in these lineages, and secondly, whether the systems are rendered non-functional upon the loss of certain determinants, would greatly inform hypotheses on the role these events have played in MDR clone formation. If the T6SS in MDR ST131 and ST410 is non-functional, it is possible that T6SS determinants remain in the genome due to their co-occurrence with other genomic factors that bear positive fitness effects. Identification of such hypothetical factors could further our understanding of how MDR clones form. Conversely, if the T6SS is shown to be functional and the loss of *tssM* does not alter its functionality, this could suggest that there is a non-homologous alternative to TssM, that is perhaps unique to MDR clones. As we have shown that T6SSs are not responsible for the efficient colonisation by MDR *E. coli*, future work needs to consider alternative hypotheses.

Our genomic analysis in Chapter 3 demonstrated the non-uniform distribution of AMR across *E. coli*. A unique aspect of this work is that a targeted approach specifically considering determinants of resistance to antibiotics of clinical interest in regards to MDR pathogens, such as carbapenems, was implemented. Lineage core genomes were the main focus of Chapter 3 so further work needs to be done to thoroughly interrogate lineage accessory genomes. We identified multiple correlations between carriage of resistance determinants and different functional categories within the accessory genomes of STs. Future studies could aim to implement in-depth comparisons of ST accessory genomes to help identify specific pathways or genes within the identified functional categories that are enriched in MDR *E. coli*.

We identified ST10 as a potential genetic source for other *E. coli* lineages.

This could suggest that there is no outgoing HGT barrier for ST10, or that this lineage is not antagonistic to other *E. coli* and therefore naturally co-occurs more frequently with other lineages, providing greater opportunity for HGT. The implications of this could be important for understanding transmission and formation of drug-resistant *E. coli*. An MDR ST10 being present in a healthy individual's gut flora could increase the risk of an incoming pathogen acquiring resistance determinants from the ST10 reservoir, increasing the chance of an untreatable infection occurring in the future. Resistance determinants being acquired from persisting ST10 strains in healthy individuals has already been documented [138]. A more immediate risk would be posed for immunocompromised patients with MDR ST10 in their gut flora, as the risks of self-infection would be far greater. An MDR strain that can easily share resistance determinants being present in a clinical setting compounds the problem of patient to patient transmission that healthcare settings are already faced with.

We have demonstrated that predicting multi-drug resistance from functionally significant sequence variation in the genomic background is a difficult task. The final results of this thesis show that gene degradation is not an indicator of multi-drug resistance. Phylogeny-based assessment of the final binary predictions revealed large variation in accuracy between STs. It could be interesting to consider the proportion of trees that voted for MDR rather than the final binary classification to provide a more nuanced analysis. Repeating the analysis on a lineage-by-lineage basis may uncover more lineage-specific indicators of multi-drug resistance and untangle the phylogenetic influence, in particular clonality, on the predictive capability of ensemble classifiers.

Multiple results from this work have highlighted the knowledge gaps that still remain in *E. coli* biology, despite being a model microorganism. The breakdown of source niches captured by our dataset reflects the general bias towards human pathogens in sequence databases due to their medical relevance. Greater

efforts to diversify the collection of available genome sequences is necessary to further scientific research. Designing studies to compare community, clinical, and MDR *E. coli* to identify distinguishing features of each group, as well as commonalities between them, will provide greater insight into MDR clone formation beyond the results shown in this thesis. Currently, community *E. coli* genomes are the limiting factor for employing studies of this nature. Expanding the coverage of sequence databases to include more community-isolated samples should be at the forefront of future genomic efforts, including those focusing on MDR *E. coli*.

To conclude, this thesis has shown that T6SSs are unlikely to be responsible for MDR *E. coli* being able to rapidly colonise healthy individuals, but their loss may possibly contribute to the success of specific MDR clones. Multi-drug resistance is concentrated in specific STs of *E. coli* and the accessory genomes of these STs are enriched in genes associated with various metabolic processes. There is no common pangenome size or structure for MDR lineages and we have shown that STs have followed distinct evolutionary trajectories. We then demonstrated that there was no common evolutionary trajectory among MDR *E. coli* that resulted in loss of function mutations. Gene-level sequence variation was not a sufficient predictor for the MDR phenotype. The results of this thesis provide further insights into the evolution of MDR *E. coli* and their ability to colonise the healthy human gut.

References

[1] Sophie S. Abby et al. “Identification of protein secretion systems in bacterial genomes”. In: *Scientific Reports* 6 (Mar. 2016). ISSN: 20452322. DOI: 10.1038/srep23080.

[2] Kaleb Abram et al. “Mash-based analyses of Escherichia coli genomes reveal 14 distinct phylogroups”. In: *Communications Biology* 4.1 (2021). ISSN: 23993642. DOI: 10.1038/s42003-020-01626-5.

[3] Colin T. Archer et al. “The genome sequence of E. coli W (ATCC 9637): Comparative genome analysis and an improved genome-scale reconstruction of E. coli”. In: *BMC Genomics* 12 (Jan. 2011). ISSN: 14712164. DOI: 10.1186/1471-2164-12-9.

[4] Maris S. Arcilla et al. “Import and spread of extended-spectrum β-lactamase-producing Enterobacteriaceae by international travellers (COMBAT study): a prospective, multicentre cohort study”. In: *The Lancet Infectious Diseases* 17.1 (Jan. 2017), pp. 78–85. ISSN: 14744457. DOI: 10.1016/S1473-3099(16)30319-X.

[5] David Arndt et al. “PHASTER: a better, faster version of the PHAST phage search tool”. In: *Nucleic Acids Research* 44.W1 (July 2016), W16–W21. ISSN: 13624962. DOI: 10.1093/nar/gkw387.

[6] Alex Bateman et al. “UniProt: the universal protein knowledgebase in 2021”. In: *Nucleic Acids Research* 49.D1 (2021), pp. D480–D489. ISSN: 13624962. DOI: 10.1093/nar/gkaa1100.

[7] Johann Beghain et al. “ClermonTyping: An easy-to-use and accurate in silico method for Escherichia genus strain phylotyping”. In: *Microbial Genomics* 4.7 (2018), pp. 1–8. ISSN: 20575858. DOI: 10.1099/mgen.0.000192.

[8] Nouri L. Ben Zakour et al. “Sequential acquisition of virulence and fluoroquinolone resistance has shaped the evolution of Escherichia coli ST131”. In: *mBio* 7.2 (2016). ISSN: 21507511. DOI: 10.1128/mBio.00347-16.

[9] Ulfar Bergthorsson and Howard Ochman. *Heterogeneity of Genome Sizes among Natural Isolates of Escherichia coli*. Tech. rep. 20. 1995, pp. 5784–5789.

[10] Patricia Bernal et al. “The Pseudomonas putida T6SS is a plant warden against phytopathogens”. In: *ISME Journal* 11.4 (Apr. 2017), pp. 972–987. ISSN: 17517370. DOI: 10.1038/ismej.2016.169.

[11] Edward R. Bevan et al. “Acquisition and loss of CTX-M-producing and non-producing escherichia coli in the fecal microbiome of travelers to South Asia”. In: *mBio* 9.6 (Nov. 2018). ISSN: 21507511. DOI: 10.1128/mBio.02408-18.

[12] Lewis EH Bingle, Christopher M. Bailey, and Mark J. Pallen. “Type VI secretion: a beginner’s guide”. In: *Current Opinion in Microbiology* 11.1 (Feb. 2008), pp. 3–8. ISSN: 13695274. DOI: 10.1016/j.mib.2008.01.006.

[13] André Birgy et al. “Independent Host Factors and Bacterial Genetic Determinants of the Emergence and Dominance of Escherichia coli Sequence Type 131 CTX-M-27 in a Community Pediatric Cohort Study”. In: (2019). DOI: 10.1128/AAC. URL: https://doi.org/10.1128/AAC.

[14] Grace A. Blackwell et al. “Exploring bacterial diversity via a curated and searchable snapshot of archived DNA sequences”. In: *PLoS Biology* 19.11 (Nov. 2021). ISSN: 15457885. DOI: 10.1371/journal.pbio.3001421.

[15] Jessica M.A. Blair et al. “Molecular mechanisms of antibiotic resistance”. In: *Nature Reviews Microbiology* 13.1 (2015), pp. 42–51. ISSN: 17401534. DOI: 10.1038/nrmicro3380.

[16] Zachary D. Blount, Christina Z. Borland, and Richard E. Lenski. “Historical contingency and the evolution of a key innovation in an experimental population of Escherichia coli”. In: *Proceedings of the National Academy of Sciences of the United States of America* 105.23 (June 2008), pp. 7899–7906. ISSN: 00278424. DOI: 10.1073/pnas.0803151105.

[17] Matthias Blum et al. “The InterPro protein families and domains database: 20 years on”. In: *Nucleic Acids Research* 49.D1 (2021), pp. D344–D354. ISSN: 13624962. DOI: 10.1093/nar/gkaa977.

[18] Thomas A. Bobik et al. “The propanediol utilization (pdu) operon of Salmonella enterica serovar Typhimurium LT2 includes genes necessary for formation of polyhedral organelles involved in coenzyme B12-dependent 1,2-propanediol degradation”. In: *Journal of Bacteriology* 181.19 (1999), pp. 5967–5975. ISSN: 00219193. DOI: 10.1128/jb.181.19.5967-5975.1999.

[19] Hamid Bokhary et al. *Travel-related antimicrobial resistance: a systematic review*. 2021. DOI: 10.3390/tropicalmed6010011.

[20] Winfried Boos and Howard Shuman. “ Maltose/Maltodextrin System of Escherichia coli : Transport, Metabolism, and Regulation ”. In: *Microbiology and Molecular Biology Reviews* 62.1 (1998), pp. 204–229. ISSN: 1092-2172. DOI: 10.1128/mmbr.62.1.204-229.1998.

[21] Michael J. Bottery, A. Jamie Wood, and Michael A. Brockhurst. “Adaptive modulation of antibiotic resistance through intragenomic coevolution”.

In: *Nature Ecology and Evolution* 1.9 (Sept. 2017), pp. 1364–1369. ISSN: 2397334X. DOI: 10.1038/s41559-017-0242-3.

[22] John Bray. "ISOLATION OF ANTIGENICALLY HOMOGENEOUS STRAINS O F BACT. COLI NEAPOLITANUM FROM SUMMER DIARRHCEA O F INFANTS". In: *The Journal of Bacteriology and Pathology* 57.2 (1945), pp. 239–247.

[23] Leo Breiman. "Random Forests". In: *Machine Learning* 45 (2001), pp. 5–32.

[24] Michael A. Brockhurst et al. "The Ecology and Evolution of Pangenomes". In: *Current Biology* 29.20 (2019), R1094–R1103. ISSN: 09609822. DOI: 10.1016/j.cub.2019.08.012. URL: https://doi.org/10.1016/j.cub.2019.08.012.

[25] Yannick R. Brunet et al. "Imaging Type VI Secretion-Mediated Bacterial Killing". In: *Cell Reports* 3.1 (2013), pp. 36–41. ISSN: 22111247. DOI: 10.1016/j.celrep.2012.11.027.

[26] Benjamin Buchfink, Chao Xie, and Daniel H. Huson. "Fast and sensitive protein alignment using DIAMOND". In: *Nature Methods* 12.1 (2014), pp. 59–60. ISSN: 15487105. DOI: 10.1038/nmeth.3176.

[27] Karen Bush and George A. Jacoby. *Updated functional classification of β-lactamases*. Mar. 2010. DOI: 10.1128/AAC.01009-09.

[28] Christiam Camacho et al. "BLAST+: Architecture and applications". In: *BMC Bioinformatics* 10 (Dec. 2009). ISSN: 14712105. DOI: 10.1186/1471-2105-10-421.

[29] Kyle J. Card et al. "Genomic evolution of antibiotic resistance is contingent on genetic background following a long-term experiment with Escherichia coli". In: *Proceedings of the National Academy of Sciences of the United States of America* 118.5 (Feb. 2021). ISSN: 10916490. DOI:

10.1073/pnas.2016886118. URL: https://www.pnas.org/content/118/5/e2016886118.

[30] Laura Carrilero et al. “Evolutionary Responses to Acquiring a Multidrug Resistance Plasmid Are Dominated by Metabolic Functions across Diverse Escherichia coli Lineages”. In: *mSystems* 8.1 (Feb. 2023). ISSN: 23795077. DOI: 10.1128/msystems.00713-22.

[31] Sujay Chattopadhyay et al. “High frequency of hotspot mutations in core genes of Escherichia coli due to short-term positive selection”. In: *Proceedings of the National Academy of Sciences of the United States of America* 106.30 (2009), pp. 12412–12417. ISSN: 00278424. DOI: 10.1073/pnas.0906217106.

[32] Roy R. Chaudhuri and Ian R. Henderson. “The evolution of the Escherichia coli phylogeny”. In: *Infection, Genetics and Evolution* 12.2 (2012), pp. 214–226. ISSN: 15671348. DOI: 10.1016/j.meegid.2012.01.005. URL: http://dx.doi.org/10.1016/j.meegid.2012.01.005.

[33] Xi Chen and Hemant Ishwaran. *Random forests for genomic data analysis*. June 2012. DOI: 10.1016/j.ygeno.2012.04.003.

[34] O. Clermont, S. Bonacorsi, and E. Bingen. “Rapid and simple determination of the Escherichia coli phylogenetic group”. In: *Applied and Environmental Microbiology* 66.10 (2000), pp. 4555–4558. ISSN: 00992240. DOI: 10.1128/AEM.66.10.4555-4558.2000.

[35] Olivier Clermont, David Gordon, and Erick Denamur. “Guide to the various phylogenetic classification schemes for escherichia coli and the correspondence among schemes”. In: *Microbiology (United Kingdom)* 161 (2015), pp. 980–988. ISSN: 14652080. DOI: 10.1099/mic.0.000063.

[36] Olivier Clermont et al. “Characterization and rapid identification of phylogroup G in Escherichia coli, a lineage with high virulence and antibi-

otic resistance potential". In: *Environmental Microbiology* 21.8 (2019), pp. 3107–3117. ISSN: 14622920. DOI: 10.1111/1462-2920.14713.

[37] Olivier Clermont et al. "The Clermont Escherichia coli phylo-typing method revisited: Improvement of specificity and detection of new phylo-groups". In: *Environmental Microbiology Reports* 5.1 (2013), pp. 58–65. ISSN: 17582229. DOI: 10.1111/1758-2229.12019.

[38] S T Cole et al. "Massive gene decay in the leprosy bacillus". In: *NATURE* 409 (2001). URL: www.nature.com.

[39] Aylin Colpan et al. "Escherichia coli sequence type 131 (ST131) subclone h30 as an emergent multidrug-resistant pathogen among US Veterans". In: *Clinical Infectious Diseases* 57.9 (Nov. 2013), pp. 1256–1265. ISSN: 10584838. DOI: 10.1093/cid/cit503.

[40] Christopher H Connor et al. "Multi-drug resistant E. coli displace commensal E. coli from the intestinal tract, a trait associated with elevated levels of genetic diversity in carbohydrate metabolism genes". In: *bioRxiv* (2022). DOI: 10.1101/2022.11.25.517930. URL: https://doi.org/10.1101/2022.11.25.517930.

[41] Gemma Croxall et al. "Increased human pathogenic potential of Escherichia coli from polymicrobial urinary tract infections in comparison to isolates from monomicrobial culture samples". In: *Journal of Medical Microbiology* 60.1 (Jan. 2011), pp. 102–109. ISSN: 00222615. DOI: 10.1099/jmm.0.020602-0.

[42] Elizabeth A. Cummins et al. "Prokaryote pangenomes are dynamic entities". In: *Current Opinion in Microbiology* 66 (2022), pp. 73–78. ISSN: 1369-5274. DOI: 10.1016/j.mib.2022.01.005.

[43] Elizabeth A. Cummins et al. "The role of potentiating mutations in the evolution of pandemic Escherichia coli clones". In: *European Journal of*

Clinical Microbiology & Infectious Diseases (2021). ISSN: 0934-9723. DOI: 10.1007/s10096-021-04359-3.

[44] Adam P. Dale and Neil Woodford. "Extra-intestinal pathogenic Escherichia coli (ExPEC): Disease, carriage and clones". In: *Journal of Infection* 71.6 (2015), pp. 615–626. ISSN: 15322742. DOI: 10.1016/j.jinf.2015.09.009. URL: http://dx.doi.org/10.1016/j.jinf.2015.09.009.

[45] Matthew Davies et al. "Enterobacteriaceae and Bacteroidaceae provide resistance to travel-associated intestinal colonization by multi-drug resistant Escherichia coli". In: *Gut Microbes* 14.1 (2022). ISSN: 19490984. DOI: 10.1080/19490976.2022.2060676.

[46] Fernanda De Pace et al. "The type VI secretion system plays a role in type 1 fimbria expression and pathogenesis of an avian pathogenic Escherichia coli strain". In: *Infection and Immunity* 78.12 (Dec. 2010), pp. 4990–4998. ISSN: 00199567. DOI: 10.1128/IAI.00531-10.

[47] Arun Gonzales Decano and Tim Downing. "An Escherichia coli ST131 pangenome atlas reveals population structure and evolution across 4,071 isolates". In: *Scientific Reports* 9.1 (2019), pp. 1–13. ISSN: 20452322. DOI: 10.1038/s41598-019-54004-5.

[48] Erick Denamur et al. "The population genetics of pathogenic Escherichia coli". In: *Nature Reviews Microbiology* (2020). ISSN: 1740-1526. DOI: 10.1038/s41579-020-0416-x.

[49] Josef Deutscher, Christof Francke, and Pieter W. Postma. "How Phosphotransferase System-Related Protein Phosphorylation Regulates Carbohydrate Metabolism in Bacteria". In: *Microbiology and Molecular Biology Reviews* 70.4 (2006), pp. 939–1031. ISSN: 1092-2172. DOI: 10.1128/mmbr.00024-06.

[50] Maria Chiara Di Luca et al. "Low biological cost of carbapenemase-encoding plasmids following transfer from Klebsiella pneumoniae to Escherichia

coli". In: *Journal of Antimicrobial Chemotherapy* 72.1 (2017), pp. 85–89. ISSN: 14602091. DOI: 10.1093/jac/dkw350.

[51] Ulrich Dobrindt et al. "Genomic islands in pathogenic and environmental microorganisms". In: *Nature Reviews Microbiology* 2.5 (2004), pp. 414–424. ISSN: 17401526. DOI: 10.1038/nrmicro884.

[52] Maria Rosa Domingo-Sananes and James O. McInerney. "Mechanisms That Shape Microbial Pangenomes". In: *Trends in Microbiology* (2021), pp. 1–11. ISSN: 0966-842X. DOI: 10.1016/j.tim.2020.12.004.

[53] Sarah M. Douglas et al. "Identification of the potentiating mutations and synergistic epistasis that enabled the evolution of inter-species cooperation". In: *PLoS ONE* 12.5 (May 2017). ISSN: 19326203. DOI: 10.1371/journal.pone.0174345.

[54] Steven J Dunn et al. "Limited and strain-specific transcriptional and growth responses to acquisition of a multidrug resistance plasmid in genetically diverse Escherichia coli lineages". In: *mSystems* April (2020), pp. 1–14. URL: https://doi.org/10.1101/2020.10.23.351718.

[55] Steven J. Dunn, Christopher Connor, and Alan McNally. "The evolution and transmission of multi-drug resistant Escherichia coli and Klebsiella pneumoniae: the complexity of clones and plasmids". In: *Current Opinion in Microbiology* 51 (2019), pp. 51–56. ISSN: 18790364. DOI: 10.1016/j.mib.2019.06.004. URL: https://doi.org/10.1016/j.mib.2019.06.004.

[56] Marion Duprilot et al. "Success of Escherichia coli O25b:H4 sequence type 131 clade C associated with a decrease in virulence". In: *Infection and Immunity* 88.12 (Dec. 2020). ISSN: 10985522. DOI: 10.1128/IAI.00576-20.

[57] Ali Ebrahim et al. "COBRApy: COnstraints-Based Reconstruction and Analysis for Python". In: *BMC Systems Biology* 7 (2013). ISSN: 17520509. DOI: 10.1186/1752-0509-7-74.

[58] Sean R Eddy. “Profile hidden Markov models”. In: *Bioinformatics* 14.9 (1998), pp. 755–763.

[59] Michael Feldgarden et al. “Validating the AMRFINder tool and resistance gene database by using antimicrobial resistance genotype-phenotype correlations in a collection of isolates”. In: *Antimicrobial Agents and Chemotherapy* 63.11 (2019). ISSN: 10986596. DOI: 10.1128/AAC.00483-19.

[60] Yu Feng et al. “Key evolutionary events in the emergence of a globally disseminated, carbapenem resistant clone in the Escherichia coli ST410 lineage”. In: *Communications Biology* 2.1 (2019), pp. 1–13. ISSN: 23993642. DOI: 10.1038/s42003-019-0569-1.

[61] Andrew D. Ferguson et al. “Siderophore-mediated iron transport: Crystal structure of FhuA with bound lipopolysaccharide”. In: *Science* 282.5397 (1998), pp. 2215–2220. ISSN: 00368075. DOI: 10.1126/science.282.5397.2215.

[62] Robert D Fleischmann et al. “Whole-Genome Random Sequencing and Assembly of Haemophilus influenzae Rd”. In: (1995). URL: https://www.science.org.

[63] Brian M. Forde et al. “The complete genome sequence of escherichia coli EC958: A high quality reference sequence for the globally disseminated multidrug resistant E. coli O25b:H4-ST131 clone”. In: *PLoS ONE* 9.8 (Aug. 2014). ISSN: 19326203. DOI: 10.1371/journal.pone.0104400.

[64] Satoru Fukiya et al. “Extensive genomic diversity in pathogenic Escherichia coli and Shigella strains revealed by comparative genomic hybridization microarray”. In: *Journal of Bacteriology* 186.12 (2004), pp. 3911–3921. ISSN: 00219193. DOI: 10.1128/JB.186.12.3911-3921.2004.

[65] Marco Galardini 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". In: (). DOI: 10.1101/712034. URL: https://doi.org/10.1101/712034.

[66] João Alves Gama et al. "Host dependent maintenance of a bla NDM-1-encoding plasmid in clinical Escherichia coli isolates". In: *Scientific Reports* 10.1 (2020), pp. 1–7. ISSN: 20452322. DOI: 10.1038/s41598-020-66239-8.

[67] Leonor García-Bayona, Michael J. Coyne, and Laurie E. Comstock. "Mobile Type VI secretion system loci of the gut Bacteroidales display extensive intra-ecosystem transfer, multi-species spread and geographical clustering". In: *PLoS Genetics* 17.4 (Apr. 2021). ISSN: 15537404. DOI: 10.1371/journal.pgen.1009541.

[68] Rebecca A Gladstone 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". In: *The Lancet Microbe* 5247.21 (2021), pp. 1–11. DOI: 10.1016/S2666-5247(21)00031-8.

[69] Evgeny N. Gordienko, Marat D. Kazanov, and Mikhail S. Gelfand. "Evolution of pan-genomes of Escherichia coli, Shigella spp., and Salmonella enterica". In: *Journal of Bacteriology* 195.12 (2013), pp. 2786–2792. ISSN: 00219193. DOI: 10.1128/JB.02285-12.

[70] Andrea Gori et al. "Pan-GWAS of streptococcus agalactiae highlights lineage-specific genes associated with virulence and niche adaptation". In: *mBio* 11.3 (2020). ISSN: 21507511. DOI: 10.1128/mBio.00728-20.

[71] Rebecca J. Hall et al. "Gene-gene relationships in an Escherichia coli accessory genome are linked to function and mobility". In: *Microbial Genomics* 7.9 (2021). DOI: 10.1099/mgen.0.000650.

[72] Riadh Hammami et al. “BACTIBASE: A new web-accessible database for bacteriocin characterization”. In: *BMC Microbiology* 7 (2007). ISSN: 14712180. DOI: 10.1186/1471-2180-7-89.

[73] Peter J Herzer et al. “Phylogenetic Distribution of Branched RNA-Linked Multicopy Single-Stranded DNA among Natural Isolates of Escherichia coli”. In: *JOURNAL OF BACTERIOLOGY* 172.11 (1990), pp. 6175–6181. URL: https://journals.asm.org/journal/jb.

[74] M. Indriati Hood and Eric P. Skaar. “Nutritional immunity: Transition metals at the pathogen-host interface”. In: *Nature Reviews Microbiology* 10.8 (Aug. 2012), pp. 525–537. ISSN: 17401526. DOI: 10.1038/nrmicro2836.

[75] David C Hooper. “Mechanisms of Action and Resistance of Older and Newer Fluoroquinolones”. In: *Clinical Infectious Diseases* 31.2 (2000), S24–S28.

[76] Gal Horesh et al. “A comprehensive and high-quality collection of escherichia coli genomes and their genes”. In: *Microbial Genomics* 7.2 (2021), pp. 1–15. ISSN: 20575858. DOI: 10.1099/mgen.0.000499.

[77] Gal Horesh et al. “Different evolutionary trends form the twilight zone of the bacterial pan-genome”. In: *Microbial Genomics* 7.9 (2021). DOI: 10.1099/mgen.0.000670.

[78] Jaime Huerta-Cepas et al. “EggNOG 5.0: A hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses”. In: *Nucleic Acids Research* 47.D1 (2019), pp. D309–D314. ISSN: 13624962. DOI: 10.1093/nar/gky1085.

[79] Jaime Huerta-Cepas et al. “Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper”. In: *Molecular Biology and Evolution* 34.8 (2017), pp. 2115–2122. ISSN: 15371719. DOI: 10.1093/molbev/msx148.

[80] Françoise Jaureguy et al. “Phylogenetic and genomic diversity of human bacteremic Escherichia coli strains”. In: *BMC Genomics* 9 (2008), pp. 1–14. ISSN: 14712164. DOI: 10.1186/1471-2164-9-560.

[81] James R. Johnson et al. “Rapid Emergence, Subsidence, and Molecular Detection of Escherichia coli Sequence Type 1193-fimH64, a New Disseminated Multidrug-Resistant Commensal and Extraintestinal Pathogen”. In: *mSphere* 57.5 (2019), pp. 1–12.

[82] Brian D. Johnston et al. “Activity of cefiderocol, ceftazidime-avibactam, and eravacycline against extended-spectrum cephalosporin-resistant Escherichia coli clinical isolates (2012–20017) in relation to phylogenetic background, sequence type 131 subclones, blaCTX-M genotype, and ”. In: *Diagnostic Microbiology and Infectious Disease* 100.1 (2021). ISSN: 18790070. DOI: 10.1016/j.diagmicrobio.2021.115314.

[83] Philip Jones et al. “InterProScan 5: Genome-scale protein function classification”. In: *Bioinformatics* 30.9 (2014), pp. 1236–1240. ISSN: 14602059. DOI: 10.1093/bioinformatics/btu031.

[84] Laure Journet and Eric Cascales. “The Type VI Secretion System in Escherichia coli and Related Species”. In: *EcoSal Plus* 7.1 (2016). ISSN: 23246200. DOI: 10.1128/ecosalplus.esp-0009-2015.

[85] Rolf S. Kaas et al. “Estimating variation within the genes and inferring the phylogeny of 186 sequenced diverse Escherichia coli genomes”. In: *BMC Genomics* 13.1 (2012). ISSN: 14712164. DOI: 10.1186/1471-2164-13-577.

[86] Teemu Kallonen et al. “Systematic longitudinal survey of invasive Escherichia coli in England demonstrates a stable population structure only transiently disturbed by the emergence of ST131”. In: *Genome Research* 27.8 (2017), pp. 1437–1449. ISSN: 15495469. DOI: 10.1101/gr.216606.116.

[87] Minoru Kanehisa. “KEGG: Kyoto Encyclopedia of Genes and Genomes”. In: *Nucleic Acids Research* 28.1 (2000), pp. 27–30.

[88] Minoru Kanehisa. “Toward understanding the origin and evolution of cellular organisms”. In: *Protein Science* 28.11 (2019), pp. 1947–1951. ISSN: 1469896X. DOI: 10.1002/pro.3715.

[89] Anu Kantele et al. “Dynamics of intestinal multidrug-resistant bacteria colonisation contracted by visitors to a high-endemic setting: a prospective, daily, real-time sampling study”. In: *The Lancet Microbe* 2.4 (2021), e151–e158. ISSN: 26665247. DOI: 10.1016/S2666-5247(20)30224-X. URL: http://dx.doi.org/10.1016/S2666-5247(20)30224-X.

[90] Kumiko Kawamura et al. “Prevalence of CTX-M-type extended-spectrum β-lactamase-producing escherichia coli B2-O25-ST131 H30R among residents in nonacute care facilities in Japan”. In: *Microbial Drug Resistance* 24.10 (Dec. 2018), pp. 1513–1520. ISSN: 19318448. DOI: 10.1089/mdr.2018.0068.

[91] Michael W. Kempnich and Marilou P. Sison-Mangus. “Presence and abundance of bacteria with the Type VI secretion system in a coastal environment and in the global oceans”. In: *PLoS ONE* 15.12 December 2020 (Dec. 2020). ISSN: 19326203. DOI: 10.1371/journal.pone.0244217.

[92] Zachary A. King et al. “BiGG Models: A platform for integrating, standardizing and sharing genome-scale models”. In: *Nucleic Acids Research* 44.D1 (2016), pp. D515–D522. ISSN: 13624962. DOI: 10.1093/nar/gkv1049.

[93] Rob Kitchin. “Big Data, new epistemologies and paradigm shifts”. In: *Big Data and Society* 1.1 (July 2014). ISSN: 20539517. DOI: 10.1177/2053951714528481.

[94] Eugene V. Koonin and Yuri I. Wolf. “Genomics of bacteria and archaea: The emerging dynamic view of the prokaryotic world”. In: *Nucleic Acids*

Research 36.21 (2008), pp. 6688–6719. ISSN: 03051048. DOI: 10.1093/nar/gkn668.

[95] Sergey Koren et al. “Hybrid error correction and de novo assembly of single-molecule sequencing reads”. In: *Nature Biotechnology* 30.7 (July 2012), pp. 693–700. ISSN: 10870156. DOI: 10.1038/nbt.2280.

[96] Benjamin Kostiuk et al. “Type VI secretion system mutations reduced competitive fitness of classical Vibrio cholerae biotype”. In: *Nature Communications* 12.1 (2021), pp. 1–11. ISSN: 20411723. DOI: 10.1038/s41467-021-26847-y.

[97] Anastasia Kottara et al. “Variable plasmid fitness effects and mobile genetic element dynamics across Pseudomonas species”. In: *FEMS Microbiology Ecology* 94.1 (Jan. 2018). ISSN: 15746941. DOI: 10.1093/femsec/fix172.

[98] Chih Horng Kuo and Howard Ochman. “The extinction dynamics of bacterial pseudogenes”. In: *PLoS Genetics* 6.8 (Aug. 2010). ISSN: 15537390. DOI: 10.1371/journal.pgen.1001050.

[99] Pascal Lapierre and J. Peter Gogarten. “Estimating the size of the bacterial pan-genome”. In: *Trends in Genetics* 25.3 (2009), pp. 107–110. ISSN: 01689525. DOI: 10.1016/j.tig.2008.12.004.

[100] Tony Le Gall et al. “Extraintestinal virulence is a coincidental by-product of commensalism in b2 phylogenetic group Escherichia coli strains”. In: *Molecular Biology and Evolution* 24.11 (2007), pp. 2373–2384. ISSN: 07374038. DOI: 10.1093/molbev/msm172.

[101] Tristan Lefébure and Michael J. Stanhope. “Evolution of the core and pan-genome of Streptococcus: Positive selection, recombination, and genome composition”. In: *Genome Biology* 8.5 (2007), pp. 1–17. ISSN: 14747596. DOI: 10.1186/gb-2007-8-5-r71.

[102] Andreas Leimbach, Jörg Hacker, and Ulrich Dobrindt. “E. coli as an all-rounder: The thin line between commensalism and pathogenicity”. In: *Current Topics in Microbiology and Immunology* 358 (2013), pp. 3–32. ISSN: 0070217X. DOI: 10.1007/82{_}2012{_}303.

[103] Ivica Letunic, Peer Bork, and Biobyte Solutions Gmbh. “Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation”. In: *Nucleic Acids Research* (2021), pp. 1–4. ISSN: 0305-1048. DOI: 10.1093/nar/gkab301.

[104] Jun Li et al. “SecReT6: A web-based resource for type VI secretion systems found in bacteria”. In: *Environmental Microbiology* 17.7 (July 2015), pp. 2196–2202. ISSN: 14622920. DOI: 10.1111/1462-2920.12794.

[105] Jingqiu Liao et al. “Nationwide genomic atlas of soil-dwelling Listeria reveals effects of selection and population ecology on pangenome evolution”. In: *Nature Microbiology* (2021). ISSN: 2058-5276. DOI: 10.1038/s41564-021-00935-7. URL: http://dx.doi.org/10.1038/s41564-021-00935-7.

[106] Lidia Moreira Lima et al. *β-lactam antibiotics: An overview from a medicinal chemistry perspective*. Dec. 2020. DOI: 10.1016/j.ejmech.2020.112829.

[107] Sam Lipworth et al. “Antimicrobial resistance genes and clonal success in Escherichia coli isolates causing bloodstream infection”. In: *The Lancet Microbe* 2 (July 2021), e492. ISSN: 26665247. DOI: 10.1016/S2666-5247(21)00031-8.

[108] Samuel Lipworth et al. “Ten-year longitudinal molecular epidemiology study of Escherichia coli and Klebsiella species bloodstream infections in Oxfordshire, UK”. In: *Genome Medicine* 13.1 (Dec. 2021). ISSN: 1756994X. DOI: 10.1186/s13073-021-00947-2.

[109] Jia Yeu Liu et al. “Functional and biochemical characterization of Escherichia coli sugar efflux transporters”. In: *Journal of Biological Chemistry* 274.33 (1999), pp. 22977–22984. ISSN: 00219258. DOI: 10.1074/jbc.274.33.22977.

[110] David M. Livermore. *Defining an extended-spectrum β-lactamase*. 2008. DOI: 10.1111/j.1469-0691.2007.01857.x.

[111] Nicholas J. Loman et al. “High-throughput bacterial genome sequencing: An embarrassment of choice, a world of opportunity”. In: *Nature Reviews Microbiology* 10.9 (Sept. 2012), pp. 599–606. ISSN: 17401526. DOI: 10.1038/nrmicro2850.

[112] Allison J. Lopatkin et al. “Clinically relevant mutations in core metabolic genes confer antibiotic resistance”. In: *Science* 371.6531 (2021). ISSN: 0036-8075. DOI: 10.1126/science.aba0862. URL: https://science.sciencemag.org/content/371/6531/eaba0862?hwshib2=authn%3A1613815442%3A20210219%253A00001618-1c9e-4df6-9475-c4970e641f65%3A0%3A0%3A0%3ASVF78jRccblZjvRtcYxlpQ%3D%3D.

[113] Oksana Lukjancenko, Trudy M. Wassenaar, and David W. Ussery. “Comparison of 61 Sequenced Escherichia coli Genomes”. In: *Microbial Ecology* 60.4 (2010), pp. 708–720. ISSN: 00953628. DOI: 10.1007/s00248-010-9717-3.

[114] Jiale Ma et al. “Diverse toxic effectors are harbored by vgrG islands for interbacterial antagonism in type VI secretion system”. In: *Biochimica et Biophysica Acta - General Subjects* 1862.7 (July 2018), pp. 1635–1643. ISSN: 18728006. DOI: 10.1016/j.bbagen.2018.04.010.

[115] Jiale Ma et al. “Genetic diversity and features analysis of type VI secretion systems loci in avian pathogenic Escherichia coli by wide genomic scanning”. In: *Infection, Genetics and Evolution* 20 (2013), pp. 454–464. ISSN: 15677257. DOI: 10.1016/j.meegid.2013.09.031.

[116] Jiale Ma et al. “PAAR-Rhs proteins harbor various C-terminal toxins to diversify the antibacterial pathways of type VI secretion systems”. In: *Environmental Microbiology* 19.1 (Jan. 2017), pp. 345–360. ISSN: 14622920. DOI: 10.1111/1462-2920.13621.

[117] Jiale Ma et al. “Two functional type VI secretion systems in avian pathogenic Escherichia coli are involved in different pathogenic pathways”. In: *Infection and Immunity* 82.9 (2014), pp. 3867–3879. ISSN: 10985522. DOI: 10.1128/IAI.01769-14.

[118] Lay Sun Ma, Franz Narberhaus, and Erh Min Lai. “IcmF family protein TssM exhibits ATPase activity and energizes type VI secretion”. In: *Journal of Biological Chemistry* 287.19 (May 2012), pp. 15610–15621. ISSN: 00219258. DOI: 10.1074/jbc.M111.301630.

[119] Daniel Machado et al. “Fast automated reconstruction of genome-scale metabolic models for microbial species and communities”. In: *Nucleic Acids Research* 46.15 (2018), pp. 7542–7553. ISSN: 13624962. DOI: 10.1093/nar/gky537.

[120] A Mademidis and W Koster. “Transport activity of FhuA, FhuC, FhuD, and FhuB derivatives in a system free of polar effects, and stoichiometry of components involved in ferrichrome uptake”. In: *Mol Gen Genet* 258 (1998), pp. 156–165.

[121] Leonardos Mageiros et al. “Genome evolution and the emergence of pathogenicity in avian Escherichia coli”. In: *Nature Communications* (2021), pp. 1–13. ISSN: 2041-1723. DOI: 10.1038/s41467-021-20988-w. URL: http://dx.doi.org/10.1038/s41467-021-20988-w.

[122] Oleksandr M. Maistrenko et al. “Disentangling the impact of environmental and phylogenetic constraints on prokaryotic within-species diversity”. In: *ISME Journal* 14.5 (2020), pp. 1247–1259. ISSN: 17517370. DOI: 10.1038/s41396-020-0600-z.

[123] Hortence Makui et al. “Identification of the Escherichia coli K-12 Nramp orthologue (MntH) as a selective divalent metal ion transporter”. In: *Molecular Microbiology* 35.5 (2000), pp. 1065–1078. ISSN: 0950382X. DOI: 10.1046/j.1365-2958.2000.01774.x.

[124] Amee R Manges et al. “Global extraintestinal pathogenic escherichia coli (Expec) lineages”. In: *Clinical Microbiology Reviews* 32.3 (2019), pp. 1–25. ISSN: 10986618. DOI: 10.1128/CMR.00135-18.

[125] Giuseppina Mariano et al. “A family of Type VI secretion system effector proteins that form ion-selective pores”. In: *Nature Communications* 10.1 (Dec. 2019). ISSN: 20411723. DOI: 10.1038/s41467-019-13439-0.

[126] Amy J. Mathers, Gisele Peirano, and Johann D.D. Pitout. “The role of epidemic resistance plasmids and international high- risk clones in the spread of multidrug-resistant Enterobacteriaceae”. In: *Clinical Microbiology Reviews* 28.3 (2015), pp. 565–591. ISSN: 10986618. DOI: 10.1128/CMR.00116-14.

[127] Yasufumi Matsumura et al. “CTX-M-27- and CTX-M-14-producing, ciprofloxacin-resistant Escherichia coli of the H30 subclonal group within ST131 drive a Japanese regional ESBL epidemic”. In: *Journal of Antimicrobial Chemotherapy* 70.6 (Nov. 2015), pp. 1639–1649. ISSN: 14602091. DOI: 10.1093/jac/dkv017.

[128] Yasufumi Matsumura et al. “Global Escherichia coli sequence type 131 clade with blaCTX-M-27 gene”. In: *Emerging Infectious Diseases* 22.11 (2016), pp. 1900–1907. ISSN: 10806059. DOI: 10.3201/eid2211.160519.

[129] Alan McNally et al. “‘Add, stir and reduce’: Yersinia spp. as model bacteria for pathogen evolution”. In: *Nature Reviews Microbiology* 14.3 (2016), pp. 177–190. ISSN: 17401534. DOI: 10.1038/nrmicro.2015.29.

[130] Alan McNally et al. “Combined Analysis of Variation in Core, Accessory and Regulatory Genome Regions Provides a Super-Resolution View into the Evolution of Bacterial Populations”. In: *PLoS Genetics* 12.9 (2016), pp. 1–16. ISSN: 15537404. DOI: 10.1371/journal.pgen.1006280.

[131] Alan McNally et al. “Diversification of colonization factors in a multidrug-resistant escherichia coli lineage evolving under negative frequency- dependent selection”. In: *mBio* 10.2 (2019), pp. 1–19. ISSN: 21507511. DOI: 10.1128/mBio.00644-19.

[132] Fabrizio Menardo et al. “Treemmer: A tool to reduce large phylogenetic datasets with minimal loss of diversity”. In: *BMC Bioinformatics* 19.1 (2018), pp. 1–8. ISSN: 14712105. DOI: 10.1186/s12859-018-2164-8.

[133] Vicky Merhej, Kalliopi Georgiades, and Didier Raoult. “Postgenomic analysis of bacterial pathogens repertoire reveals genome reduction rather than virulence factors”. In: *Briefings in Functional Genomics* 12.4 (July 2013), pp. 291–304. ISSN: 20412649. DOI: 10.1093/bfgp/elt015.

[134] Irene Merino et al. “Emergence of ESBL-producing Escherichia coli ST131-C1-M27 clade colonizing patients in Europe”. In: *Journal of Antimicrobial Chemotherapy* 73.11 (Nov. 2018), pp. 2973–2980. ISSN: 14602091. DOI: 10.1093/jac/dky296.

[135] Alex Mira, Howard Ochman, and Nancy A Moran. “Deletional bias and the evolution of bacterial genomes”. In: *TRENDS in Genetics* 17.10 (2001). URL: http://tig.trends.com0168.

[136] Jonathan M. Monk et al. “Genome-scale metabolic reconstructions of multiple Escherichia coli strains highlight strain-specific adaptations to nutritional environments”. In: *Proceedings of the National Academy of Sciences of the United States of America* 110.50 (2013), pp. 20338–20343. ISSN: 00278424. DOI: 10.1073/pnas.1307797110.

[137] Danesh Moradigaravand et al. “Prediction of antibiotic resistance in Escherichia coli from large-scale pan-genome data”. In: *PLoS Computational Biology* 14.12 (2018), pp. 1–17. ISSN: 15537358. DOI: 10.1371/journal.pcbi.1006258.

[138] Robert A. Moran et al. “Plasmids in antibiotic susceptible and antibiotic resistant commensal Escherichia coli from healthy Australian adults”. In: *Plasmid* 80 (July 2015), pp. 24–31. ISSN: 10959890. DOI: 10.1016/j.plasmid.2015.03.005.

[139] Hamed Mosaei and Nikolay Zenkin. “Inhibition of RNA Polymerase by Rifampicin and Rifamycin-Like Molecules”. In: *EcoSal Plus* 9.1 (Mar. 2020). ISSN: 23246200. DOI: 10.1128/ecosalplus.esp-0017-2019.

[140] Christopher JL Murray et al. “Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis”. In: *The Lancet* 399.10325 (Feb. 2022), pp. 629–655. ISSN: 1474547X. DOI: 10.1016/S0140-6736(21)02724-0.

[141] Fernando Navarro-Garcia et al. *Type VI secretion system in pathogenic escherichia coli: Structure, role in virulence, and acquisition*. 2019. DOI: 10.3389/fmicb.2019.01965.

[142] Marie Hélène Nicolas-Chanoine, Xavier Bertrand, and Jean Yves Madec. “Escherichia coli st131, an intriguing clonal group”. In: *Clinical Microbiology Reviews* 27.3 (2014), pp. 543–574. ISSN: 10986618. DOI: 10.1128/CMR.00125-13.

[143] Marie Hélène Nicolas-Chanoine et al. “Intercontinental emergence of Escherichia coli clone O25:H4-ST131 producing CTX-M-15”. In: *Journal of Antimicrobial Chemotherapy* 61.2 (Feb. 2008), pp. 273–281. ISSN: 03057453. DOI: 10.1093/jac/dkm464.

[144] Karen L. Nielsen et al. “Faecal Escherichia coli from patients with E. coli urinary tract infection and healthy controls who have never had a urinary

tract infection". In: *Journal of Medical Microbiology* 63.PART 4 (2014), pp. 582–589. ISSN: 00222615. DOI: 10.1099/jmm.0.068783-0.

[145] Barbara Nobelmann and Joseph W Lengeler. "Sequence of the gat operon for galactitol utilization from a wild-type strain EC3132 of Escherichia coli". In: *BBA - Gene Structure and Expression* 1262.1 (1995), pp. 69–72. ISSN: 01674781. DOI: 10.1016/0167-4781(95)00053-J.

[146] Masayuki Okuyama et al. "Overexpression and characterization of two unknown proteins, YicI and YihQ, originated from Escherichia coli". In: *Protein Expression and Purification* 37.1 (2004), pp. 170–179. ISSN: 10465928. DOI: 10.1016/j.pep.2004.05.008.

[147] Brian D. Ondov et al. "Mash: Fast genome and metagenome distance estimation using MinHash". In: *Genome Biology* 17.1 (2016), pp. 1–14. ISSN: 1474760X. DOI: 10.1186/s13059-016-0997-x.

[148] Rafael Patiño-Navarrete et al. "Stepwise evolution and convergent recombination underlie the global dissemination of carbapenemase-producing Escherichia coli". In: *Genome Medicine* 6 (2020), pp. 1–18. DOI: 10.1101/446195.

[149] N Pearcy et al. "Genome-Scale Metabolic Models and Machine Learning Reveal Genetic Determinants of Antibiotic Resistance in Escherichia coli and Unravel the Underlying Metabolic Adaptation Mechanisms". In: *mSystems* 6 (2021), pp. 00913–20.

[150] Fabian Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830. URL: http://scikit-learn.sourceforge.net..

[151] Nicola K. Petty et al. "Global dissemination of a multidrug resistant Escherichia coli clone". In: *Proceedings of the National Academy of Sciences of the United States of America* 111.15 (2014), pp. 5694–5699. ISSN: 10916490. DOI: 10.1073/pnas.1322678111.

[152] Laura J V Piddock. *Mechanisms of Fluoroquinolone Resistance: An Update 1994-1998*. Tech. rep. 1999, pp. 11–18.

[153] Johann D.D. Pitout. "Extraintestinal pathogenic Escherichia coli: A combination of virulence with antibiotic resistance". In: *Frontiers in Microbiology* 3.JAN (2012), pp. 1–7. ISSN: 1664302X. DOI: 10.3389/fmicb.2012.00009.

[154] Johann D.D. Pitout et al. "Emergence of Enterobacteriaceae producing extended-spectrum β-lactamases (ESBLs) in the community". In: *Journal of Antimicrobial Chemotherapy* 56.1 (July 2005), pp. 52–59. ISSN: 03057453. DOI: 10.1093/jac/dki166.

[155] Johann DD Pitout and Kevin B. Laupland. *Extended-spectrum β-lactamase-producing Enterobacteriaceae: an emerging public-health concern*. Mar. 2008. DOI: 10.1016/S1473-3099(08)70041-0.

[156] Laurent Poirel, Aurélie Jayol, and Patrice Nordmanna. *Polymyxins: Antibacterial activity, susceptibility testing, and resistance mechanisms encoded by plasmids or chromosomes*. Apr. 2017. DOI: 10.1128/CMR.00064-16.

[157] Jacques Pouyssegur and Francois Stoeber. "Genetic Control of the 2-Keto-3-Deoxy-D-Gluconate Metabolism in Escherichia coli K-12: kdg Regulon". In: *JOURNAL OF BACTERIOLOGY* (1974), pp. 641–651. URL: https://journals.asm.org/journal/jb.

[158] Lance B. Price et al. "The epidemic of extended-spectrum-β-lactamase-producing Escherichia coli ST131 is driven by a single highly pathogenic subclone, H30-Rx". In: *mBio* 4.6 (Dec. 2013). ISSN: 21612129. DOI: 10.1128/mBio.00377-13.

[159] Pere Puigbò et al. "Genomes in turmoil: Quantification of genome dynamics in prokaryote supergenomes". In: *BMC Medicine* 12.1 (2014), pp. 1–19. ISSN: 17417015. DOI: 10.1186/s12915-014-0066-4.

[160] Stefan Pukatzki et al. “Type VI secretion system translocates a phage tail spike-like protein into target cells where it cross-links actin”. In: *Proceedings of the National Academy of Sciences of the United States of America* 104.39 (Sept. 2007), pp. 15508–15513. ISSN: 00278424. DOI: 10.1073/pnas.0706532104.

[161] Amit Ranjan et al. “ESBL-plasmid carriage in E. coli enhances in vitro bacterial competition fitness and serum resistance in some strains of pandemic sequence types without overall fitness cost”. In: *Gut Pathogens* 10.1 (2018), pp. 1–9. ISSN: 17574749. DOI: 10.1186/s13099-018-0243-z. URL: https://doi.org/10.1186/s13099-018-0243-z.

[162] David A. Rasko et al. “The pangenome structure of Escherichia coli: Comparative genomic analysis of E. coli commensal and pathogenic isolates”. In: *Journal of Bacteriology* 190.20 (2008), pp. 6881–6893. ISSN: 00219193. DOI: 10.1128/JB.00619-08.

[163] Sean D Reid et al. *Parallel evolution of virulence in pathogenic Escherichia coli*. Tech. rep. 2000. URL: www.nature.com.

[164] Sandra Reuter et al. “Parallel independent evolution of pathogenicity within the genus Yersinia”. In: *Proceedings of the National Academy of Sciences of the United States of America* 111.18 (May 2014), pp. 6768–6773. ISSN: 10916490. DOI: 10.1073/pnas.1317161111.

[165] Sofia Rigou et al. “Metagenomic survey of the microbiome of ancient Siberian permafrost and modern Kamchatkan cryosols”. In: *microLife* 3 (June 2022). DOI: 10.1093/femsml/uqac003.

[166] L. W. Riley. “Pandemic lineages of extraintestinal pathogenic Escherichia coli”. In: *Clinical Microbiology and Infection* 20.5 (2014), pp. 380–390. ISSN: 14690691. DOI: 10.1111/1469-0691.12646. URL: http://dx.doi.org/10.1111/1469-0691.12646.

[167] Louise Roer et al. “Escherichia coli Sequence Type 410 Is Causing New International High-Risk Clones”. In: *mSphere* 3.4 (2018), pp. 00337–18. ISSN: 1535-9786. DOI: 10.1128/msphere.00337-18.

[168] M. H. Saier and J. Reizer. “Proposed uniform nomenclature for the proteins and protein domains of the bacterial phosphoenolpyruvate: Sugar phosphotransferase system”. In: *Journal of Bacteriology* 174.5 (1992), pp. 1433–1438. ISSN: 00219193. DOI: 10.1128/jb.174.5.1433-1438.1992.

[169] Dor Salomon et al. “Marker for type VI secretion system effectors”. In: *Proceedings of the National Academy of Sciences of the United States of America* 111.25 (June 2014), pp. 9271–9276. ISSN: 10916490. DOI: 10.1073/pnas.1406110111.

[170] Dor Salomon et al. “Type VI Secretion System Toxins Horizontally Shared between Marine Bacteria”. In: *PLoS Pathogens* 11.8 (Aug. 2015). ISSN: 15537374. DOI: 10.1371/journal.ppat.1005128.

[171] Maria Manuel Sampaio et al. “Phosphotransferase-mediated Transport of the Osmolyte 2-O-α -Mannosyl-D-glycerate in Escherichia coli Occurs by the Product of the mngA (hrsA) Gene and Is Regulated by the mngR (farR) Gene Product Acting as Repressor”. In: *Journal of Biological Chemistry* 279.7 (2004), pp. 5537–5548. ISSN: 00219258. DOI: 10.1074/jbc.M310980200.

[172] Per E.J. Saris and E. Tapio Palva. “Regulation of manX (ptsL) and manY (pel) genes required for mannose transport and penetration of λDNA in Escherichia coli K12”. In: *FEMS Microbiology Letters* 44.3 (1987), pp. 377–382. ISSN: 15746968. DOI: 10.1111/j.1574-6968.1987.tb02317.x.

[173] Robert E Schapire. “The Strength of Weak Learnability”. In: *Machine Learning* 5 (1990), pp. 197–227.

[174] Katharina Schaufler et al. “Clonal spread and interspecies transmission of clinically relevant ESBL-producing Escherichia coli of ST410-another successful pandemic clone?” In: *FEMS Microbiology Ecology* 92.1 (2016), pp. 1–9. ISSN: 15746941. DOI: 10.1093/femsec/fiv155.

[175] Katharina Schaufler et al. “Genomic and Functional Analysis of Emerging Virulent and Multidrug-Resistant Escherichia coli Lineage Sequence Type 648”. In: *Antimicrobial Agents and Chemotherapy* 63.6 (2019), pp. 1–12.

[176] Torsten Seemann. “Prokka: Rapid prokaryotic genome annotation”. In: *Bioinformatics* 30.14 (2014), pp. 2068–2069. ISSN: 14602059. DOI: 10.1093/bioinformatics/btu153.

[177] Sabiha Shaik et al. “Comparative genomic analysis of globally dominant ST131 clone with other epidemiologically successful extraintestinal pathogenic escherichia coli (ExPEC) lineages”. In: *mBio* 8.5 (2017), pp. 1–15. ISSN: 21507511. DOI: 10.1128/mBio.01596-17.

[178] Sabiha Shaik et al. “Genome Informatics and Machine Learning-Based Identification of Antimicrobial Resistance-Encoding Features and Virulence Attributes in Escherichia coli Genomes Representing Globally Prevalent Lineages, Including High-Risk Clonal Complexes”. In: *mBio* 13.1 (2022). ISSN: 21507511. DOI: 10.1128/MBIO.03796-21.

[179] Samuel K. Sheppard, David S. Guttman, and J. Ross Fitzgerald. “Population genomics of bacterial host adaptation”. In: *Nature Reviews Genetics* 19.9 (2018), pp. 549–565. ISSN: 14710064. DOI: 10.1038/s41576-018-0032-z.

[180] Stanford T. Shulman, Herbert C. Friedmann, and Ronald H. Sims. “Theodor Escherich: the first pediatric infectious diseases physician?” In: *Clinical infectious diseases : an official publication of the Infectious Diseases Soci-*

ety of America 45.8 (2007), pp. 1025–1029. ISSN: 15376591. DOI: 10.1086/521946.

[181] Hanna E. Sidjabat et al. "Molecular epidemiology of CTX-M-producing Escherichia coli isolates at a tertiary medical center in Western Pennsylvania". In: *Antimicrobial Agents and Chemotherapy* 53.11 (2009), pp. 4733–4739. ISSN: 00664804. DOI: 10.1128/AAC.00533-09.

[182] P. Siguier et al. "ISfinder: the reference centre for bacterial insertion sequences." In: *Nucleic acids research* 34.Database issue (2006). ISSN: 13624962. DOI: 10.1093/nar/gkj014.

[183] Julie M. Silverman et al. "Separate inputs modulate phosphorylation-dependent and -independent type VI secretion activation". In: *Molecular Microbiology* 82.5 (Dec. 2011), pp. 1277–1290. ISSN: 0950382X. DOI: 10.1111/j.1365-2958.2011.07889.x.

[184] Ola Sköld. "Sulfonamide resistance: Mechanisms and trends". In: *Drug Resistance Updates* 3.3 (2000), pp. 155–160. ISSN: 13687646. DOI: 10.1054/drup.2000.0146.

[185] Nicole Stoesser et al. "Evolutionary history of the global emergence of the Escherichia coli epidemic clone ST131". In: *mBio* 7.2 (2016), pp. 1–15. ISSN: 21507511. DOI: 10.1128/mBio.02162-15.

[186] Daniel A. Tadesse et al. "Antimicrobial drug resistance in Escherichia coli from humans and food animals, United States, 1950-2002". In: *Emerging Infectious Diseases* 18.5 (2012), pp. 741–749. ISSN: 10806040. DOI: 10.3201/eid1805.111153.

[187] Lei Tang et al. "Expression, secretion and bactericidal activity of type VI secretion system in Vibrio anguillarum". In: *Archives of Microbiology* 198.8 (Oct. 2016), pp. 751–760. ISSN: 1432072X. DOI: 10.1007/s00203-016-1236-2.

[188] Erwin Tantoso et al. “To kill or to be killed : pangenome analysis of Escherichia coli strains reveals a tailocin specific for pandemic ST131”. In: *BMC Biology* (2022), pp. 1–26. ISSN: 1741-7007. DOI: 10.1186/s12915-022-01347-7. URL: https://doi.org/10.1186/s12915-022-01347-7.

[189] Roman L. Tatusov et al. “The COG database: A tool for genome-scale analysis of protein functions and evolution”. In: *Nucleic Acids Research* 28.1 (2000), pp. 33–36. ISSN: 03051048. DOI: 10.1093/nar/28.1.33.

[190] Veronika L. Tchesnokova et al. “Rapid and Extensive Expansion in the United States of a New Multidrug-resistant Escherichia coli Clonal Group, Sequence Type 1193”. In: *Clinical Infectious Diseases* 68.2 (2019), pp. 334–337. ISSN: 15376591. DOI: 10.1093/cid/ciy525.

[191] Olivier Tenaillon et al. “The population genetics of commensal Escherichia coli”. In: *Nature Reviews Microbiology* 8.3 (2010), pp. 207–217. ISSN: 17401526. DOI: 10.1038/nrmicro2298.

[192] Hervé Tettelin et al. “Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial ’pan-genome’”. In: *Proceedings of the National Academy of Sciences of the United States of America* 102.45 (2005), p. 16530. ISSN: 00278424. DOI: 10.1073/pnas.0508532102.

[193] Hao Chung The et al. *The genomic signatures of Shigella evolution, adaptation and geographical spread*. Mar. 2016. DOI: 10.1038/nrmicro.2016.10.

[194] Gerry Tonkin-Hill et al. “Producing Polished Prokaryotic Pangenomes with the Panaroo Pipeline”. In: *BMC Genome Biology* (2020), pp. 1–21. DOI: 10.1101/2020.01.28.922989.

[195] Todd J. Treangen and Eduardo P.C. Rocha. “Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes”. In: *PLoS*

Genetics 7.1 (2011). ISSN: 15537390. DOI: 10.1371/journal.pgen.1001284.

[196] Vera Troselj et al. "Physiological Heterogeneity Triggers Sibling Conflict Mediated by the Type VI Secretion System in an Aggregative Multicellular Bacterium". In: *mBio* 9.1 (2018). DOI: 10.1128/mBio.01645. URL: https://doi.org/10.1128/mBio.01645.

[197] Rahul Unni et al. "Presence and absence of type VI secretion systems in bacteria". In: *Microbiology* in print (2022), pp. 1–13. DOI: 10.1099/mic.0.001151.

[198] Sergei B. Vakulenko and Shahriar Mobashery. "Versatility of aminoglycosides and prospects for their future". In: *Clinical Microbiology Reviews* 16.3 (July 2003), pp. 430–450. ISSN: 08938512. DOI: 10.1128/CMR.16.3.430-450.2003.

[199] Gisela Di Venanzio et al. "Multidrug-resistant plasmids repress chromosomally encoded T6SS to enable their dissemination". In: *Proceedings of the National Academy of Sciences of the United States of America* 116.4 (2019), pp. 1378–1383. ISSN: 10916490. DOI: 10.1073/pnas.1812557116.

[200] R M Waterson and R S Conway. "Enoyl-CoA hydratases from Clostridium acetobutylicum and Escherichia coli". In: *Methods Enzymol.* 71 (1981), pp. 421–30.

[201] Brent S. Weber et al. "A multidrug resistance plasmid contains the molecular switch for type VI secretion in Acinetobacter baumannii". In: *Proceedings of the National Academy of Sciences of the United States of America* 112.30 (2015), pp. 9442–9447. ISSN: 10916490. DOI: 10.1073/pnas.1502966112.

[202] U. F. Wehmeier, B. Nobelmann, and J. W. Lengeler. “Cloning of the Escherichia coli sor genes for L-sorbose transport and metabolism and physical mapping of the genes near metH and iclR”. In: *Journal of Bacteriology* 174.23 (1992), pp. 7784–7790. ISSN: 00219193. DOI: 10.1128/jb.174.23.7784-7790.1992.

[203] Nicole E. Wheeler. *Tracing outbreaks with machine learning*. May 2019. DOI: 10.1038/s41579-019-0153-1.

[204] Nicole E. Wheeler, Paul P. Gardner, and Lars Barquist. “Machine learning identifies signatures of host adaptation in the bacterial pathogen Salmonella enterica”. In: *PLoS Genetics* 14.5 (2018), pp. 1–20. ISSN: 15537404. DOI: 10.1371/journal.pgen.1007333. URL: http://dx.doi.org/10.1371/journal.pgen.1007333.

[205] Nicole E. Wheeler et al. “A profile-based method for identifying functional divergence of orthologous genes in bacterial genomes”. In: *Bioinformatics* 32.23 (2016), pp. 3566–3574. ISSN: 14602059. DOI: 10.1093/bioinformatics/btw518.

[206] Nicole E. Wheeler et al. “Genomic correlates of extraintestinal infection are linked with changes in cell morphology in Campylobacter jejuni”. In: *Microbial Genomics* 5.2 (Feb. 2019). ISSN: 20575858. DOI: 10.1099/mgen.0.000251.

[207] Fiona J. Whelan, Rebecca J. Hall, and James O. McInerney. “Evidence for Selection in the Abundant Accessory Gene Content of a Prokaryote Pangenome”. In: *Molecular Biology and Evolution* 38.9 (2021), pp. 3697–3708. ISSN: 15371719. DOI: 10.1093/molbev/msab139.

[208] WHO. *PRIORITIZATION OF PATHOGENS TO GUIDE DISCOVERY, RESEARCH AND DEVELOPMENT OF NEW ANTIBIOTICS FOR DRUG-RESISTANT BACTERIAL INFECTIONS, INCLUDING TUBERCULOSIS*. Tech. rep. 2017.

[209] Ryan R. Wick et al. “Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads”. In: *PLoS Computational Biology* 13.6 (June 2017). ISSN: 15537358. DOI: 10.1371/journal.pcbi.1005595.

[210] Thierry Wirth et al. “Sex and virulence in Escherichia coli: An evolutionary perspective”. In: *Molecular Microbiology* 60.5 (2006), pp. 1136–1151. ISSN: 0950382X. DOI: 10.1111/j.1365-2958.2006.05172.x.

[211] Ying Xu and Ning Yi Zhou. “MhpA is a hydroxylase catalyzing the initial reaction of 3-(3-Hydroxyphenyl)propionate catabolism in Escherichia coli K-12”. In: *Applied and Environmental Microbiology* 86.4 (2020), e02385. ISSN: 10985336. DOI: 10.1128/AEM.02385-19.

[212] Huimin Zhang et al. “Genomic insights into the ESBL and MCR-1-producing ST648 Escherichia coli with multi-drug resistance”. In: *Science Bulletin* 61.11 (2016), pp. 875–878. ISSN: 20959281. DOI: 10.1007/s11434-016-1086-y.

[213] Jianfeng Zhang et al. “SecReT6 update: a comprehensive resource of bacterial Type VI Secretion Systems”. In: *Science China Life Sciences* (Nov. 2022). ISSN: 1674-7305. DOI: 10.1007/s11427-022-2172-x.

[214] Liqing Zhang et al. *TssM is essential for virulence and required for type VI secretion in Ralstonia solanacearum*. Tech. rep. 4. 2012, pp. 125–134. URL: https://www.jstor.org/stable/43229327?seq=1&cid=pdf-.

[215] Xueqing Zhang et al. “First identification of coexistence of blaNDM-1 and bla CMY-42 among Escherichia coli ST167 clinical isolates”. In: *BMC Microbiology* 13.1 (2013), pp. 1–7. ISSN: 14712180. DOI: 10.1186/1471-2180-13-282.

[216] You Zhou et al. “PHAST: A Fast Phage Search Tool”. In: *Nucleic Acids Research* 39.SUPPL. 2 (July 2011). ISSN: 03051048. DOI: 10.1093/nar/gkr485.

[217] Zhemin Zhou et al. “The EnteroBase user’s guide, with case studies on Salmonella transmissions, Yersinia pestis phylogeny, and Escherichia core genomic diversity”. In: *Genome Research* 30.1 (2020), pp. 138–152. ISSN: 15495469. DOI: 10.1101/gr.251678.119.

[218] Xiangkai Zhuge et al. “Avian-source mcr-1-positive Escherichia coli is phylogenetically diverse and shares virulence characteristics with E. coli causing human extra-intestinal infections”. In: *Veterinary Microbiology* 239 (Dec. 2019). ISSN: 18732542. DOI: 10.1016/j.vetmic.2019.108483.

[219] Zhiyong Zong and Rujia Yu. “Escherichia coli carrying the blaCTX-M-15 gene of ST648”. In: *Journal of Medical Microbiology* 59.12 (2010), pp. 1536–1537. ISSN: 00222615. DOI: 10.1099/jmm.0.022459-0.

[220] Zhiyong Zong et al. “Complete genomic characterization of two Escherichia coli lineages responsible for a cluster of carbapenem-resistant infections in a Chinese hospital”. In: *Journal of Antimicrobial Chemotherapy* 73.9 (2018), pp. 2340–2346. ISSN: 14602091. DOI: 10.1093/jac/dky210.

Appendix Zero

Supplementary Data

Table 1: Genes assigned as core uniquely to a specified phylogroup. Description and preferred name assigned by eggNOG-mapper.

Cluster name	Description	Preferred name
PHYLOGROUP B1:		
elfA_elfA_1	Fimbrial	elfA
esiB_1	Sel1-like repeats.	-
esiB_3	Sel1-like repeats.	-
fliD	Required for morphogenesis and for the elongation of the flagellar filament by facilitating polymerization of the flagellin monomers at the tip of growing filament. Forms a capping structure, which prevents flagellin subunits (transported through the central channel of the flagellum) from leaking out without polymerization at the distal end.	fliD

Cluster name	Description	Preferred name
fliD_2_fliD_1	Required for morphogenesis and for the elongation of the flagellar filament by facilitating polymerization of the flagellin monomers at the tip of growing filament. Forms a capping structure, which prevents flagellin subunits (transported through the central channel of the flagellum) from leaking out without polymerization at the distal end.	fliD
fliS	flagellar protein FliS	fliS
fliS_2	flagellar protein FliS	fliS
gltC_3	Transcriptional regulator	yiaU
group_1051	Acetyltransferase (GNAT) domain	-
group_1575	Protein of unknown function (DUF2556)	yhdU
group_1579	Enterobacterial EspB protein	espB
group_1631	intracellular signal transduction	marT
group_1712	Protein of unknown function (DUF1778)	-
group_1968	tail protein	-
group_2120	-	-
group_2194	Lambda Phage CIII	-
group_2210	cell adhesion	-
group_2324	Domain of unknown function (DUF1869)	yoaG
group_2397	Acetyltransferase (GNAT) domain	-
group_2925	-	yneK
group_3071	Usher protein	yqiG
group_3183	NAD(P)H-binding	-
group_34	Outer membrane autotransporter	yejO
group_3501	Domain of unknown function (DUF1869)	yoaG
group_3602	Phage minor tail protein U	-
group_3611	Protein of unknown function (DUF1778)	-
group_3807	-	-
group_4393	Protein of unknown function (DUF1778)	-
group_525	Transposase	yadD
group_602	Hydrolase	yafV

Cluster name	Description	Preferred name
group_603	Domain of unknown function (DUF1869)	yoaG
group_999	Acetyltransferase (GNAT) domain	-
icsA_1	outer membrane autotransporter barrel domain protein	-
lpfD	Fimbrial protein	lpfD
manX_1	IIa component	yadI
mepH_2	COG0791 Cell wall-associated hydrolases (invasion-associated proteins)	yafL
nudI	Catalyzes the hydrolysis of nucleoside triphosphates, with a preference for pyrimidine deoxynucleoside triphosphates (dUTP, dTTP and dCTP)	nudI
papD_1	chaperone	yqiH
pezT	zeta toxin	-
prn	outer membrane autotransporter barrel	yaiT
prn__tibA	outer membrane autotransporter barrel	yaiT
prn_tibA	outer membrane autotransporter barrel	yaiT
ssaV	type iii secretion	ssaV
tir	Translocated intimin receptor (Tir) N-terminus	tir
yadC	Fimbrial protein	yadC
yadC_yadC_1 _yadC_2	Fimbrial protein	yadC
ybgD_1_mrpA_2	Fimbrial protein	ygiL
yhjX	Major facilitator superfamily	yhjX
yqeG	transporter	yqeG
group_1099		-
group_2556		-
group_1271		-
PHYLOGROUP B2:		
adh1	alcohol dehydrogenase	dhaT
atsA_2_atsA_3	COG3119 Arylsulfatase A and related enzymes	-
cidA_1	LrgA family	cidA
cidA_2	LrgA family	cidA
cidA_2_cidA_1	LrgA family	cidA

Cluster name	Description	Preferred name
citT_1	transporter	citT
citT_2	transporter	-
dapA_3_dapA_1	Belongs to the DapA family	-
dctB_dctB_2	Signal transduction histidine kinase	-
dctD	COG2204 Response regulator containing CheY-like receiver, AAA-type ATPase, and DNA-binding domains	IV02_08980
dctD_1	COG2204 Response regulator containing CheY-like receiver, AAA-type ATPase, and DNA-binding domains	IV02_08980
dctD_2	COG2204 Response regulator containing CheY-like receiver, AAA-type ATPase, and DNA-binding domains	IV02_08980
dctD_2_afsQ1	COG2204 Response regulator containing CheY-like receiver, AAA-type ATPase, and DNA-binding domains	IV02_08980
ddpC_dppB_1 _ddpC_2	Binding-protein-dependent transport systems inner membrane component	ddpB
ddpD_oppD_3	(ABC) transporter	gsiA_3
flhE	Flagellar protein flhE	flhE
fruA_1	Pts system fructose-specific	-
fruA_3	Pts system fructose-specific	fruA
glcR_1	transcriptional regulator	ygbI
glcR_2	transcriptional regulator	ygbI
group_1065	-	-
group_1144	curli production assembly transport component CsgF	csgF
group_1196	Metal-dependent hydrolase	ygjP
group_1453	-	-
group_1506	-	-
group_2026	Nucleoside-specific channel-forming protein, Tsx	-
group_2243	cell adhesion	-
group_2294	-	-

Cluster name	Description	Preferred name
group_230	ydiK promoter presents a PurR sequence, suggesting that its expression is purine-regulated	ydiK
group_2541	Domain of unknown function (DUF4862)	-
group_2583	Transposase	-
group_2629	Small membrane protein	ygdD
group_2930	Nucleoside-specific channel-forming protein, Tsx	-
group_2953	Transposase	-
group_2955	BNR repeat-like domain	-
group_322	type VI secretion-associated protein	-
group_3624	-	-
group_3900	type III effector	ygbK
group_4031	-	-
group_4092	type III effector	ygbK
group_4435	membrane protein, terc	terC
group_4485	-	-
group_4807	Nucleoside-specific channel-forming protein, Tsx	-
group_4841	BNR repeat-like domain	-
group_5171	BNR repeat-like domain	-
group_5175	type III effector	ygbK
group_5315	-	yohO
group_553	ATP binding	yqeC
group_571	Transposase	yfcI
group_628	Uncharacterised MFS-type transporter YbfB	yjiJ
group_638	BNR repeat-like domain	-
group_86	-	-
group_880	type III effector	ygbK
imm_3	immunity protein	-
ldh	Belongs to the LDH2 MDH2 oxidoreductase family	-
ldh_ldh_2 _ldh_1	Belongs to the LDH2 MDH2 oxidoreductase family	-
lpd	COG1249 Pyruvate 2-oxoglutarate dehydrogenase complex, dihydrolipoamide dehydrogenase (E3) component, and related enzymes	lpdG

Cluster name	Description	Preferred name
lpd_lpd_1	COG1249 Pyruvate 2-oxoglutarate dehydrogenase complex, dihydrolipoamide dehydrogenase (E3) component, and related enzymes	lpdG
mcbR	Transcriptional regulator	mcbR
mscL	Channel that opens in response to stretch forces in the membrane lipid bilayer. May participate in the regulation of osmotic pressure changes within the cell	mscL
nudB	dihydroneopterin triphosphate	nudB
oppF_2	Belongs to the ABC transporter superfamily	dppF
pdxA2	Catalyzes the NAD(P)-dependent oxidation of 4-(phosphohydroxy)-L-threonine (HTP) into 2-amino-3-oxo-4-(phosphohydroxy)butyric acid which spontaneously decarboxylates to form 3-amino-2-oxopropyl phosphate (AHAP)	pdxA2
pdxA2_2_pdxA2_1	Catalyzes the NAD(P)-dependent oxidation of 4-(phosphohydroxy)-L-threonine (HTP) into 2-amino-3-oxo-4-(phosphohydroxy)butyric acid which spontaneously decarboxylates to form 3-amino-2-oxopropyl phosphate (AHAP)	pdxA2
psuK_1	PfkB domain protein	psuK
psuK_2	PfkB domain protein	psuK
pys1_hcpA_3	COG3209 Rhs family protein	yhhZ
pys2	COG3209 Rhs family protein	yhhZ
pys2_hcpA_5	COG3209 Rhs family protein	yhhZ
rpoS	RNA polymerase sigma	rpoS
siaM	transporter large permease	yiaN
sucA_1	Dehydrogenase E1 component	sucA
sucA_2	Dehydrogenase E1 component	sucA
sucA_2_sucA_3	Dehydrogenase E1 component	sucA

Cluster name	Description	Preferred name
sucB_1	The 2-oxoglutarate dehydrogenase complex catalyzes the overall conversion of 2-oxoglutarate to succinyl-CoA and CO(2)	sucB
sucB_1__chbF_1	The 2-oxoglutarate dehydrogenase complex catalyzes the overall conversion of 2-oxoglutarate to succinyl-CoA and CO(2)	-
sucB_2	The 2-oxoglutarate dehydrogenase complex catalyzes the overall conversion of 2-oxoglutarate to succinyl-CoA and CO(2)	-
sucC_1	Succinyl-CoA synthetase functions in the citric acid cycle (TCA), coupling the hydrolysis of succinyl-CoA to the synthesis of either ATP or GTP and thus represents the only step of substrate-level phosphorylation in the TCA. The beta subunit provides nucleotide specificity of the enzyme and binds the substrate succinate, while the binding sites for coenzyme A and phosphate are found in the alpha subunit	sucC
sucC_2_sucC_3	Succinyl-CoA synthetase functions in the citric acid cycle (TCA), coupling the hydrolysis of succinyl-CoA to the synthesis of either ATP or GTP and thus represents the only step of substrate-level phosphorylation in the TCA. The beta subunit provides nucleotide specificity of the enzyme and binds the substrate succinate, while the binding sites for coenzyme A and phosphate are found in the alpha subunit	sucC

Cluster name	Description	Preferred name
sucD_1	Succinyl-CoA synthetase functions in the citric acid cycle (TCA), coupling the hydrolysis of succinyl-CoA to the synthesis of either ATP or GTP and thus represents the only step of substrate-level phosphorylation in the TCA. The alpha subunit of the enzyme binds the substrates coenzyme A and phosphate, while succinate binding and nucleotide specificity is provided by the beta subunit	sucD
sucD_1_sucD_4	Succinyl-CoA synthetase functions in the citric acid cycle (TCA), coupling the hydrolysis of succinyl-CoA to the synthesis of either ATP or GTP and thus represents the only step of substrate-level phosphorylation in the TCA. The alpha subunit of the enzyme binds the substrates coenzyme A and phosphate, while succinate binding and nucleotide specificity is provided by the beta subunit	sucD
sucD_4	Membrane protein FdrA	yahF
tabA_1	Domain of unknown function (DUF386)	yhcH
tabA_2	Domain of unknown function (DUF386)	yiaL
tabA_3	PFAM conserved	yjgK
tabA_4	Domain of unknown function (DUF386)	yhcH
ybhI_2	transporter	-
ybhI_2_ybhI_1	transporter	-
ycaD	Major facilitator superfamily	ycaD

Cluster name	Description	Preferred name
yfcP	Part of the yfcOPQRSUV fimbrial operon. Could contribute to adhesion to various surfaces in specific environmental niches. Increases adhesion to eukaryotic T24 bladder epithelial cells in the absence of fim genes	yfcP
yhaH	Membrane	yhaH
yhaH_yhaH_2	Membrane	yhaH
yiaM	Tripartite ATP-independent periplasmic transporters, DctQ component	-
yiaM_yiaM_1 _yiaM_2	Tripartite ATP-independent periplasmic transporters, DctQ component	-
yjjX	Phosphatase that hydrolyzes non-canonical purine nucleotides such as XTP and ITP to their respective diphosphate derivatives. Probably excludes non-canonical purines from DNA precursor pool, thus preventing their incorporation into DNA and avoiding chromosomal lesions	yjjX
yjjX_yjjX_2	Phosphatase that hydrolyzes non-canonical purine nucleotides such as XTP and ITP to their respective diphosphate derivatives. Probably excludes non-canonical purines from DNA precursor pool, thus preventing their incorporation into DNA and avoiding chromosomal lesions	yjjX
yohK_1	effector of murein hydrolase	yohK_2
yohK_2	effector of murein hydrolase	yohK
group_3504		-
PHYLOGROUP D:		
group_1975	Transposase	-
group_4795	Transposase	-
group_5887	Transposase	-
htrE	Usher protein	htrE
ipaH3	regulation of response to stimulus	inlE

Cluster name	Description	Preferred name
upaG_ehaG	Autotransporter adhesin	xadA
yadM	Part of the yadCKLM-htrE-yadVN fimbrial operon. Could contribute to adhesion to various surfaces in specific environmental niches	yadM
yadN	Part of the yadCKLM-htrE-yadVN fimbrial operon. Could contribute to adhesion to various surfaces in specific environmental niches	yadN
yadV	chaperone	yadV
yadV_yadV_1	chaperone	yadV
group_5620		-
group_368		-
group_1801		-
group_1609		-
PHYLOGROUP E:		
acr1	Belongs to the short-chain dehydrogenases reductases (SDR) family	VY92_08700
cckA	Histidine kinase	virA
cdhR_2	Transcriptional regulator containing an amidase domain and an AraC-type DNA-binding HTH domain	ftrA
cdhR_3_pchR	Transcriptional regulator containing an amidase domain and an AraC-type DNA-binding HTH domain	adpA
crl__stxB_1 _stxB	nuclear chromosome segregation	-
cshA	dEAD DEAH box helicase	-
dmsA_2_dmsA_1	Belongs to the prokaryotic molybdopterin-containing oxidoreductase family	-
dmsB_3	anaerobic DMSO reductase chain C anchor subunit	-
elfC_1	Usher protein	-
elfG_2	Fimbrial protein	-
fbaA_3_fbaA_2	Aldolase	-

Cluster name	Description	Preferred name
fcl_1	Catalyzes the two-step NADP-dependent conversion of GDP- 4-dehydro-6-deoxy-D-mannose to GDP-fucose, involving an epimerase and a reductase reaction	fcl
fliC_fliC_1	Flagellin is the subunit protein which polymerizes to form the filaments of bacterial flagella	fliC
gbpR_gbpR_1	transcriptional regulator	-
gmd_1_gmd_3 _gmd_2	Catalyzes the conversion of GDP-D-mannose to GDP-4-dehydro-6-deoxy-D-mannose	gmd
gmm_1	Belongs to the Nudix hydrolase family	gmm
gnu_gnu_2	epimerase dehydratase	gne
group_1082	-	-
group_1085	Zn-dependent hydrolases of the beta-lactamase fold	-
group_1194	-	-
group_1195	OmpA family	-
group_1265	Phage derived protein Gp49-like (DUF891)	-
group_1525	Rhodanese	-
group_1922	Protein of unknown function (DUF3343)	-
group_1950	Type II restriction enzyme, methylase subunits	-
group_1956	domain, Protein	-
group_1964	-	-
group_2402	ImpA domain protein	-
group_245	COG4886 Leucine-rich repeat (LRR) protein	yopM
group_2458	Evidence 3 Function proposed based on presence of conserved amino acid motif, structural feature or limited homology	eaeH
group_2667	-	yjdP
group_2826	Belongs to the 'phage' integrase family	intB
group_3162	Stress-induced protein	yciG
group_3220	Type III secretion	invE

Cluster name	Description	Preferred name
group_330	4Fe-4S dicluster domain	-
group_349	sequence-specific DNA binding	-
group_549	SdiA-regulated	yjiK
group_653	nuclear chromosome segregation	-
group_654	HNH nucleases	-
group_667	D-lyxose isomerase	-
group_680	Part of the yfcOPQRSUV fimbrial operon. Could contribute to adhesion to various surfaces in specific environmental niches. Increases adhesion to eukaryotic T24 bladder epithelial cells in the absence of fim genes	yfcO
group_69	-	-
group_737	Protein of unknown function (DUF2773)	ydbD
group_780	conserved small protein	-
group_825	domain, Protein	-
group_841	Protein of unknown function (DUF1778)	-
group_911	Evidence 3 Function proposed based on presence of conserved amino acid motif, structural feature or limited homology	yeeJ
hcpA_1_hcpA_2	Type VI secretion system	-
htrC	-	htrC
htrE	Usher protein	htrE
intS_4_intS_3	Belongs to the 'phage' integrase family	intS
lpfA_1	Fimbrial protein	-
lpfC	Usher protein	-
lpfC_lpfC_1	Usher protein	-
lpfD	Fimbrial protein	lpfD
lpfE	Fimbrial protein	lpfE
manC1_1_manC1_3_manC1_2	Belongs to the mannose-6-phosphate isomerase type 2 family	cpsB
mhbM_mhbM_1	COG0654 2-polyprenyl-6-methoxyphenol hydroxylase and related FAD-dependent oxidoreductases	mhbM

Cluster name	Description	Preferred name
mhbT_mhbT_2 _mhbT_1	transporter	pcaK
nagI	Gentisate 1,2-dioxygenase	gtdA
nagK_1	COG0179 2-keto-4-pentenoate hydratase 2-oxohepta-3-ene-1, 7-dioic acid hydratase (catechol pathway)	mhbB
nagL	maleylacetoacetate isomerase	maiA
nanM_1	Converts alpha-N-acetylneuranimic acid (Neu5Ac) to the beta-anomer, accelerating the equilibrium between the alpha- and beta-anomers. Probably facilitates sialidase-negative bacteria to compete sucessfully for limited amounts of extracellular Neu5Ac, which is likely taken up in the beta-anomer. In addition, the rapid removal of sialic acid from solution might be advantageous to the bacterium to damp down host responses	-
ompN_1_ompC_2 _ompN_3_ompC_1	Belongs to the Gram-negative porin family	ompS1
ompR_2_phoP_1	response regulator	-
parD1	transcriptional regulators containing the CopG Arc MetJ DNA-binding domain	parD
parE4	Plasmid stabilization system	-
pcrA	DNA helicase	-
perB	Bacterial transferase hexapeptide (six repeats)	-

Cluster name	Description	Preferred name
rapA_1	Transcription regulator that activates transcription by stimulating RNA polymerase (RNAP) recycling in case of stress conditions such as supercoiled DNA or high salt concentrations. Probably acts by releasing the RNAP, when it is trapped or immobilized on tightly supercoiled DNA. Does not activate transcription on linear DNA. Probably not involved in DNA repair	-
rbsA_3_araG_2 _araG_3	import. Responsible for energy coupling to the transport system	rbsA2
rbsB_2_rbsB_1	PFAM Periplasmic binding protein LacI transcriptional regulator	-
rbsC_5_rbsC_1	Belongs to the binding-protein-dependent transport system permease family	-
rhlE_1	dEAD DEAH box helicase	-
sopA_3_sopA_1	ATP-dependent helicase activity	-
tfoX1	regulator of competence-specific genes	sxy
tir	Translocated intimin receptor (Tir) N-terminus	tir
uhpB_1__uhpB_3	Histidine kinase	uhpB
uhpC_2	transporter	uhpC
yadL	Fimbrial protein	-
yadM	Part of the yadCKLM-htrE-yadVN fimbrial operon. Could contribute to adhesion to various surfaces in specific environmental niches	yadM
yadN_yadN_2	Fimbrial protein	-
yadV_3	chaperone	-
yahA_yahA_1 _yahA_2	Hydrolyzes c-di-GMP (cyclic bis(3'-5') dimeric GMP) to 5'-pGpG, known as PDE-A activity. PDE-B activity, that is hydrolysis of 5'-pGpG to GMP, proceeds only very slowly	yahA
ybgD_ybgD_2	Fimbrial protein	ybgD
ydbD	Mn-containing catalase	-

Cluster name	Description	Preferred name
ydjH_2	PfkB domain protein	-
yraI_1	pilus organization	-
group_2064		-
group_543		-
group_3154		-
tlyC		-
group_2512		-
group_2003		-
group_2798		-
group_202		-
group_1827		-
group_1516		-
ysnE		-
PHYLOGROUP F:		
group_4803	Glycosyl Transferase	JD73_04250

Table 2: ST131-specific core genes assigned to functional COG category G (carbohydrate metabolism) by eggNOG-mapper (v2.0.8) bestOG assignment. Preferred name assigned by eggNOG-mapper.

Cluster name	Preferred name
nagA_2	*agaA*
bglB	*bglB*
uxaA_2	*uxaA*
manX_2	*agaF*
ycjU	*ycjU*
fbaA_2	*fba*
sgbE	*sgbE*
nagB_3	*agaI*
exuT_1	*exuT*
ygbM	*ygbM*
dhaM	*dhaM*
dhaK	*dhaK*
pgaB	*pgaB*
group_357	YPO0284
manR_2	-
rpiB	*rpiB*
kanE	*wcaI*
alsE	*alsE*
gntP	*gntP*
siaM_1	-
rhmT	*rhmT*
gatY_1	*gatY*
manZ_3	*agaE*
bcr_2	*tcaB*
kdgK_4	-
group_2529	IV02_08475
kdgK_3	*scrK*
rbsK_1	-
kbaY_2	-
mak_2	Z012_07775
gtfA	*ycjM*
ulaE_2	*sgbU*
xapB	*xapB*
alsC	*alsC*
dgoT	*dgoT*
ygbN	*ygbN*
nagB_1	*yieK*
xylB_1	-
scrK	*scrK*

Cluster name	Preferred name
group_2955	-
manC1	*cpsB*
group_2977	-
fucA_3_fucA_2	*ygbL*
treA_treA_1	*treA*
fruA_2	*fruA_2*
gatZ	*gatZ*
gatB_2_gatB_3	*gatB*
group_3501	*xylS*
group_3588	-
sacX_treP	*treP*
yiaO_yiaO_2	*yiaO*
malX_malX_2	-
ptsG_2_ptsG_1	-
group_3798	*ycjR*
alsK_alsK_1	*alsK*
tabA_2	*yjgK*
fcl_gmm	*fcl*
sorM	*sorM*
ulaA_1_ulaA_2	*ulaA*
yiaM_yiaM_1_yiaM_2	-
rbsC_1_xylH_1	-
rbsB_3__rbsB_4	-
group_4030	-
sorF_sorB_2	*sorB*
sorA_sorA_2_sorA_1	*sorA*
agaC_1_manZ_1_manZ_2	*agaD*
bglF_bglF_1_bglF_2	*bglF*
group_4092	*ygbK*
tabA_3	*yiaL*
crh_ptsH_2	*ptsH*
yihP_yicJ_3_yicJ_1	-
mngB_mngB_1_mngB_2	*mngB*
mngA_mngA_2	*hrsA*
siaP_2	-
lyx	*lyx*
group_4283	-
rhmA	*rhmA*
gatC_4	*gatC*
suyB_uxaA_3	-
ydjE_1__ydjE_3	*ydjK*
mhpT	*mhpT*
uidA_uidB_3	*uidA*

Cluster name	Preferred name
gatA_gatA_1	*gatA*
gatC_2_gatC_3	*gatC*
msmE_kojP__malP_3	*ycjT*
kgtP_1_proP_3_mppA	-
proP_2_proP_1_yhjE_1	*ydfJ*
fruA_3_fruA_1	-
yicJ_2_siaM_2	*yiaN*
gatY_2_gatY_1	*gatY*
dgoA_2_dgoK1_2	*dgoK*
uxaB	*uxaB*
yggF	*yggF*
group_721	-
chiA_chiA_1	*chiA*
group_759	*rbsK*
lacZ_lacZ_1	*lacZ*
yicL_yicL_2	*yicL*
rfbB_rmlB_rfbB_1_rfbD_1_rfbD_2_rfbB_2	-
xylH_xylH_1	*xylH*

Table 3: ST10-specific core genes assigned to functional COG category G (carbohydrate metabolism) by eggNOG-mapper (v2.0.8) bestOG assignment. Preferred name assigned by eggNOG-mapper.

Cluster name	Preferred name
bglH_1	*ascB*
rhaM	*rhaM*
ydjH	*ydjH*
yihT	*yihT*
dhaM	*dhaM*
ulaA	*ulaA*
fucA_3	*ygbL*
ygbN	*ygbN*
mngB	*mngB*
rpiB	*rpiB*
bglF_1	*ascF*
nepI	*nepI*
frlD_frlD_1	*frlD*
setA_setA_1_setA_2	*setA*
yidA_dgoK1_1	*dgoK*
gtfA	*ycjM*
dhaL_dhaK_1	*dhaK*
yggF_yggF_1	*yggF*
group_160	*frlC*
setC_setC_1_setC_2	*setB*
deoB_1_deoB_2	*yhfW*
fbaA_2_fbaA_1	*ydjI*
treA_treA_1	*treA*
rhmA_rhmT_1_rhmT	*rhmT*
mngA_mngA_1	*hrsA*
ribZ_ribZ_1	*yebQ*
mtlA_1_mtlA_3_mtlA_2_coaA_1_coaA_2	*cmtA*
uidB_uidB_1_uidB_2	*uidB*
ycjO_msmE	*ycjN*
malP_3_malP_1_malP_4	*aglA*
yihV	*yihV*
yicL_yicL_1	*yicL*
chbG_chbG_2	*chbG*
xapB	*xapB*
lsrB_lsrB_1	*lsrB*
tabA_2	*tabA*
ydjE_2	*ydjK*
melB__melB_1_melB_2	*melB*
ygbM_ygbM_2	*ygbM*

Cluster name	Preferred name
manR_2	*frvR*
yihN_yihN_2	*yihN*
por_2_por_1__yhjE_2	*ydfI*